16 Summarising data

In this chapter you will learn:

- different ways to measure the centre of data
- different ways to measure the spread of data
- how to work with data that has been summarised
- some useful ways of representing data
- the effects of constant changes to data
- how to describe the strength of a relationship between two variables
- to find the equation of the line of best fit for two variables.

Introductory problem

The magnetic dipole of an electron is measured three times in a very sensitive experiment. The values obtained are 2.000 0015, 2.000 0012 and 2.000 0009. Does this data support the theory that the magnetic dipole is 2?

Huge amounts of data are collected in scientific experiments and social surveys. This data in its raw state can be very difficult to interpret, so we often summarise the important features using statistics – individual numbers or diagrams that display some aspect of the data, such as where its centre lies or how spread out it is.

16A Measures of the centre of data

In everyday language the word *average* is often used without specifying whether it refers to the mean, median or mode. This can lead to some confusing newspaper headlines. In mathematics it is important to be precise.

What are the pros and cons of representing a complex set of data by a single number?

One of the things we very often want to know about a set of data is a single number that could represent the whole set. This value is called an *average*. There are several different types of average. For the International Baccalaureate® you need to know about three kinds: the mean, the median and the mode.

The **mean** is the value you get when you add up all the numbers in a set and divide by the number of items in the set. It is what is usually meant when people refer to an 'average'. If the numbers in the set are $x_1, x_2, \ldots$, then the mean is usually denoted by $\bar{x}$; sometimes it is also given the symbol μ.

See Prior Learning section Y on the CD-ROM for how to find the mean, median and mode.

Technically, this is the *arithmetic* mean. There are several other types of 'mean' – the geometric mean, the harmonic mean and the quadratic mean, to name a few. Each has many uses, and perhaps most interestingly, if all the numbers in the data set are positive, these means always come out in the same order: harmonic ≤ geometric ≤ arithmetic ≤ quadratic.

The **median** is the number you get when you put all the data in order and find the middle value. If there is not just one middle value, you take the mean of the two middle values.

The **mode** is the value that occurs most frequently in the data set. There can be no mode or more than one mode for a given set of data.

Why are there so many different averages? They all have advantages and disadvantages, and depending on the situation, one type may be more suitable than the others for representing the data.

Type of average	Advantage	Disadvantage
Mean	Takes into account all of the data values	Can be skewed by outliers
Median	Is not skewed by outliers	Does not take into account all of the data
Mode	Can be used for non-numeric data	Can take more than one value

Worked example 16.1

Wages in a factory are £3000, £5000, £10 000 and £150 000. Find the mean, median and mode. Why does the mean give an unrepresentative average?

The mean is the sum of all the data values divided by the number of data items.

$$\text{mean} = \frac{£3000 + £5000 + £10000 + £150000}{4}$$

$$= £42000$$

With 4 items, the median is the average of the second and third in the ordered list.

$$\text{median} = \frac{£5000 + £10000}{2}$$

$$= £7500$$

→

continued . . .

The mode is the most frequent value.

There is no mode because all 4 values occur equally often.

The mean is unrepresentative because it is skewed by the extra-large value of £150 000 (an outlier).

EXAM HINT

You can use a calculator to find the mean and median of a set of data. See Calculator skills sheets 10 and 11 on the CD-ROM for how to do this. However, when the data set is small, you may find it faster to carry out the calculation by hand as above.

Exercise 16A

1. Find the mean, median and mode for the following sets of data.

(a) (i) 19.0, 23.4, 36.2, 18.7, 15.7

(ii) 0.4, −1.3, 7.9, 8.4, −9.4

(b) (i) 28, 31, 54, 28, 17, 30

(ii) 60, 18, 42, 113, 95, 23

2. Find the mean, median and mode for the following sets of data.

(a) 15, 15, 34, 15, 34, 4

(b) 3,−8,6,−8,14,22

3. A newspaper headline says ‘Half of children have below average intelligence’. Is such a statement always true?

4. For each of the following statements, decide whether it is always true, sometimes true or never true.

(a) The median is smaller than the mean.

(b) The median takes the value of one of the data items.

(c) The mean is a whole number.

(d) If all of the data items are whole numbers then so is the mean.

(e) If all of the data items are whole numbers then so is the mode.

(f) The mode is larger than the median.

(g) The mean is less than or equal to the maximum data value.

(h) There are the same number of data items with value below the median as above the median.

5. A sample of 14 measurements has a mean of 20.4, and another sample of 20 measurements has a mean of 16.8. Find the mean of all 34 measurements. *[5 marks]*

6. Jenny must sit four papers for an exam. The mean of the first three papers Jenny has sat is 72%.

(a) If she wants to get an overall mean of at least 75%, what is the lowest mark she can get in her fourth paper?

(b) What is the highest possible mean she can get over all four papers? *[6 marks]*

7. Five data items are as follows: $x, y, 1, 3, 10$

The mean is 5.4, and the median is 5. Find the values of x and y. *[5 marks]*

8. The table summarises the marks gained in a test by a group of students.

Mark	1	2	3	4	5
Number of students	5	10	p	6	2

The median mark is 3 and the mode is 2. Find the **two** possible values of p. *[6 marks]*

(*© IB Organization 2004*)

9. Amy and Bob are playing a computer game. Amy's average score on both level one and level two is higher than Bob's. Show that it is possible for Bob to still have a higher overall average across levels one and two.

The word *range* is used in several different contexts in mathematics, both technical and informal, and can have very different meanings. When talking about the range of a function, we mean the *set of values* which the function can output; the range of a data set, however, is a *single number* that represents the length of the interval in which the data values lie.

16B Measures of spread

Once we have a representative value for the centre of a data set, we may also be interested in how far away from this centre the other data values lie. This 'distance from the average' is called the spread (or dispersion) of the data, and, just as with the average, there are several different ways to measure the spread.

The simplest measure of spread is the **range**, which is the difference between the largest and smallest data values. The main disadvantage of the range is that it is extremely sensitive to outliers.

An improved measure that avoids the undue influence of outliers is the **interquartile range**, usually abbreviated IQR. Instead of taking the difference between the extreme values of the data, we arrange the data values in order and look at the difference between the data item one-quarter of the way up, called the **lower quartile** (abbreviated LQ or Q_1), and the data item three-quarters of the way up, called the **upper quartile** (UQ or Q_3). To find the quartiles Q_1 and Q_3 we split the ordered data into two halves, discarding the central item if there are an odd number of values, and then find the median of each half. In the same notation, the median of the whole data set can be called Q_2, the minimum value Q_0, and the maximum value Q_4.

KEY POINT 16.1

$$IQR = Q_3 - Q_1$$

As with the median, one disadvantage of the IQR is that it does not take into account all of the data. A commonly used measure that does use all the data values is the **standard deviation**, usually given the symbol σ. This measures the average distance of data items from the mean.

EXAM HINT

In the International Baccalaureate Standard Level Mathematics course, you will only be expected to find standard deviations using your calculator. See Calculator skills sheets 10 and 11 on the CD-ROM for instructions on how to do this.

EXAM HINT

Look carefully at whether a question is asking for standard deviation or variance.

The square of the standard deviation is called the **variance**. It is not a direct measure of spread, but in more advanced statistics it is a very convenient quantity to work with algebraically.

Worked example 16.2

Find the range, interquartile range and standard deviation of the following set of numbers:

1, 12, 9, 9, 15, 7, 5

Range = 15 − 1 = 14

Order the data and then split into halves. The number of values is odd, so omit the middle value.

Data in order:
1, 5, 7, $\boxed{9}$, 9, 12, 15

LQ = 5, UQ = 12, so IQR = 12 − 5 = 7

Use GDC to find the standard deviation.

$\sigma = 4.23$ (3 SF)

As a rough guide, about two-thirds of the data will be less than one standard deviation away from the mean. In a large data set, nearly all of the values will be within two standard deviations from the mean, and any value more than three standard deviations from the mean would be very unusual.

Exercise 16B

1. For each set of data calculate the standard deviation and interquartile range.

   (a) (i) 19.0, 23.4, 36.2, 18.7, 15.7

   (ii) 0.4, −1.3, 7.9, 8.4, −9.4

   (b) (i) 28, 31, 54, 28, 17, 30

   (ii) 60, 18, 42, 113, 95, 23

2. Six people live in a house. They record their ages in whole years. The oldest is 32 and the youngest is 20. State whether each of the following statements is true, false or impossible to be sure. For cases where it is impossible to be sure, give an example for which the statement is true.

   (a) The range is greater than the mean.

   (b) The median is 26.

   (c) The mean is greater than 10.

   (d) The mean is greater than the mode

   (e) The median is not a whole number.

3. Find the interquartile range of the following set of data:

$3, 4, 5, 5, 6, 8, 11, 13$ *[3 marks]*

4. The ordered set of data $5, 5, 7, 8, 9, x, 13$ has interquartile range equal to 7.

(a) Find the value of x.

(b) Find the standard deviation of the data set. *[5 marks]*

5. Three numbers a, b, c are such that $a < b < c$. The median is 12, the range is 12 and the mean is 14. Find the value of c.

[4 marks]

16C Frequency tables and grouped data

It is very common to summarise large amounts of data in a frequency table. This is a list of all the values that the data items take, along with how often each value occurs. Given a frequency table, we could always expand it into a list of all the data values and then calculate the statistics discussed in sections 16A and 16B, but usually it is enough to just imagine writing out the whole list.

For example, suppose we are given the following data:

x	Frequency
1	2
2	5
3	3

We could write this out as a list:

$$\underbrace{1,1,}_{\text{2 ones}}\ \overbrace{2,2,2,2,2,}^{\text{5 twos}}\underbrace{3,3,3}_{\text{3 threes}}$$

From this list we could calculate statistics such as the mean and standard deviation as before. With large data sets, however, it would be impractical to write out the full list. Instead, to find the mean we use the fact that we would add up each data value as often as it occurs. This leads to the following formula.

KEY POINT 16.2

From a frequency table, $\bar{x} = \dfrac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

where f_i is the frequency of the ith data value and $\sum_{i=1}^{n} f_i$, the sum of all the frequencies, is the total number of data items.

See section 6B for a reminder about $\sum$ notation.

Worked example 16.3

The numbers of passengers observed in cars passing a school are as follows:

Passengers	Frequency
0	32
1	16
2	2
3 or more	0

Find the median and mean number of passengers in each car.

There are 32 + 16 + 2 = 50 data items, so the median is the average of the 25th and 26th items. Both of these lie in the group corresponding to 0 passengers.

median = average of 25th and 26th numbers

so median = 0

For the mean, use the formula in Key point 16.2.

$$\bar{x} = \frac{(32\times 0)+(16\times 1)+(2\times 2)}{32+16+2}$$

$$= \frac{20}{50}$$

$$= 0.4$$

EXAM HINT

The mean does not have to be one of the data values, so do not round to the nearest whole number.

Data is often grouped because, even though some detail is lost, it is easier to get an overview of the data set. Would having the original data values necessarily be better?

So far we have been given the exact data values, but when dealing with **grouped data**, we no longer have this information. The simplest way of estimating the mean and standard deviation of grouped data is to assume that all the original data values in each group are located at the centre of the group, called the **mid-interval value**. To find the mid-interval value of a group, we take the mean of the largest and smallest possible values in the group, called the **upper and lower interval boundaries**. You can use your calculator to find the mean and standard deviation of grouped data by entering the mid-interval values for each group.

To find the median and interquartile range for grouped data you need to use cumulative frequency, which is covered in section 16D.

Worked example 16.4

Find the mean and standard deviation of the mass of eggs produced at a chicken farm. Explain why these answers are only estimates.

Mass of eggs (grams)	Frequency
[100,120[	26
[120,140[	52
[140,160[	84
[160,180[	60
[180,200[	12

Rewrite the table, replacing each group by the mid-interval value.

Midpoint	f_i
110	26
130	52
150	84
170	60
190	12

EXAM HINT

Any values that you have found which are not given in the question must be written down.

→

continued . . .

Use GDC to calculate the mean and standard deviation.

$\bar{x} = 148.3g$ (3 SF)
$\sigma = 21.2g$ (3 SF)

These answers are only estimates because we have assumed that all the data in each group is at the centre, rather than using the actual data values.

EXAM HINT

Whenever you find a mean or a standard deviation, it is always worth checking that the numbers make sense in the given context. For this data, an average of about 150 g seems reasonable.

Exercise 16C

1. Calculate the mean, standard deviation and median for the given data sets.

(a)

x	Frequency
10	7
12	19
14	2
16	0
18	2

(b)

x	Frequency
0.1	16
0.2	15
0.3	12
0.4	9
0.5	8

2 A group is described as '17–20'. State the upper and lower boundaries of this group if the data is measuring:

(a) age in completed years

(b) number of pencils

(c) length of a worm to the nearest centimetre

(d) hourly earnings, rounded *up* to whole dollars.

3. Find the mean and standard deviation of each of the following sets of data.

(a) (i) x is the time taken to complete a puzzle in seconds.

x	Frequency
[0,15[	19
[15,30[	15
[30,45[	7
[45,60[	5
[60,90[	4

(ii) x is the weight of plants in grams.

x	Frequency
[50,100[	17
[100,200[	23
[200,300[	42
[300,500[	21
[500,1000[	5

(b) (i) x is the length of fossils found at a geological dig, rounded to the nearest centimetre.

x	Frequency
0 to 4	71
5 to 10	43
11 to 15	22
16 to 30	6

(ii) x is the power consumption of light bulbs, rounded to the nearest watt.

x	Frequency
90 to 95	17
96 to 100	23
101 to 105	42
106 to 110	21
111 to 120	5

(c) (i) x is the age of children in a hospital ward, in completed years.

x	Frequency
0 to 2	12
3 to 5	15
6 to 10	7
11 to 16	6
17 to 18	3

(ii) x is the amount of tip customers paid in a restaurant, rounded *down* to the nearest dollar.

x	Frequency
0 to 5	17
6 to 10	29
11 to 20	44
21 to 30	16
31 to 50	8

4. In a sample of 50 boxes of eggs, the number of broken eggs per box is shown in the table.

Number of broken eggs per box	0	1	2	3	4	5	6
Number of boxes	17	8	7	7	6	5	0

(a) Calculate the median number of broken eggs per box.

(b) Calculate the mean number of broken eggs per box.

[4 marks]

5. The mean of the data in the table is 12.6 and the range is 15. Find the values of p and q.

x	Frequency
5	6
10	p
15	$2p$
q	2

[4 marks]

6. The mean of the data in the table is 30.4 and the range is 20. Find the values of p and q.

x	Frequency
20	12
40	p
60	q

[3 marks]

16D Cumulative frequency

In the previous section we saw how to estimate the mean and standard deviation of grouped data. Now we move on to finding the median and quartiles. This is easiest if we represent the data in a new way, using a **cumulative frequency** table or diagram. Cumulative frequency is a count of the total number of data items *up to a certain value*.

Worked example 16.5

Convert the 'masses of eggs' table from Worked example 16.4 into a cumulative frequency table.

Here is the original table of grouped data.

Mass of eggs (grams)	Frequency
[100,120[	26
[120,140[	52
[140,160[	84
[160,180[	60
[180,200[	12

→

continued . . .

The first column of the cumulative frequency table consists of the upper boundaries of the data groups. The second column counts how many items there are up to that point.

Mass of eggs (grams)	Cumulative frequency
120	26
140	78
160	162
180	222
200	234

Once we have organised the data in cumulative frequency form, we can draw a cumulative frequency diagram. This is a graph with data values along the *x*-axis and cumulative frequencies along the *y*-axis. For the example above, we know that 26 eggs are under 120 g, 78 eggs are under 140 g, and so on. Therefore we plot the points (120, 26), (140, 78), etc. In other words, for grouped data, we plot the cumulative frequency against the upper bound of each group.

Also, notice that besides the values in the cumulative frequency table of Worked example 16.5, there is another point we can plot: from the original data table we know that there are no eggs lighter than 100 g, so (100, 0) is the leftmost point in the diagram.

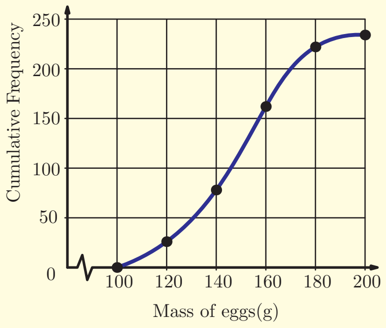
EXAM HINT

There is always an additional data point to be plotted, of the lowest lower bound and zero cumulative frequency.

Worked example 16.6

Plot a cumulative frequency curve for the 'mass of eggs' data from Worked examples 16.4 and 16.5.

Plot the five points from the cumulative frequency table of Worked example 16.5, plus the additional point (100, 0), and join them up with an increasing curve.



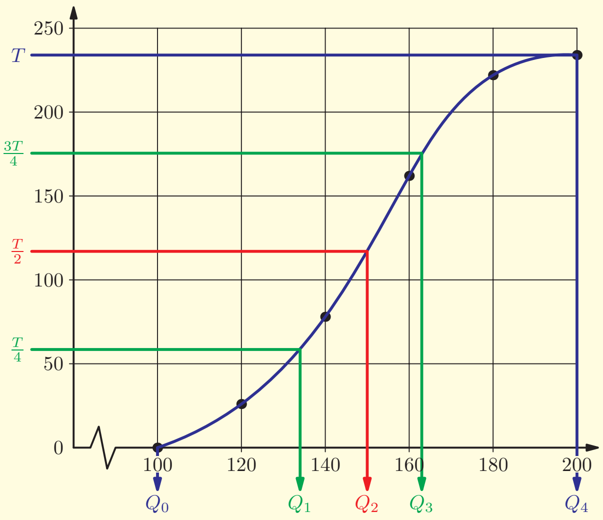
We can use the cumulative frequency graph to estimate the median and the quartiles. The median is the data value corresponding to the middle data item, so draw a horizontal line from the y-axis at *half* the total frequency (maximum y-value) until it meets the cumulative frequency curve, and then draw a vertical line down to the x-axis to obtain the median value.

To find the quartiles Q_1 and Q_3, the process is similar, except that the horizontal lines should be drawn at one-quarter and three-quarters of the total frequency.

Worked example 16.7

Find the median and interquartile range of the eggs data from Worked examples 16.4–16.6.

The total frequency is 234, so draw lines across from the y-axis at $0.5 \times 234 = 117$ for the median, $0.25 \times 234 = 58.5$ for the lower quartile, and $0.75 \times 234 = 175.5$ for the upper quartile. Where these horizontal lines meet the cumulative frequency curve, draw vertical lines downward to the x-axis to find the values of the median and quartiles.



Read off the values on the x-axis for Q_1, Q_2 and Q_3.

Median (Q_2) ≈ 150 g

Upper quartile (Q_3) ≈ 163 g

Lower quartile (Q_1) ≈ 134 g

$\text{IQR} = Q_3 - Q_1 \approx 29$ g

EXAM HINT

Drawing in appropriate vertical and horizontal lines on cumulative frequency graphs can get you method marks in the exam even if the cumulative frequency graph itself is wrong.

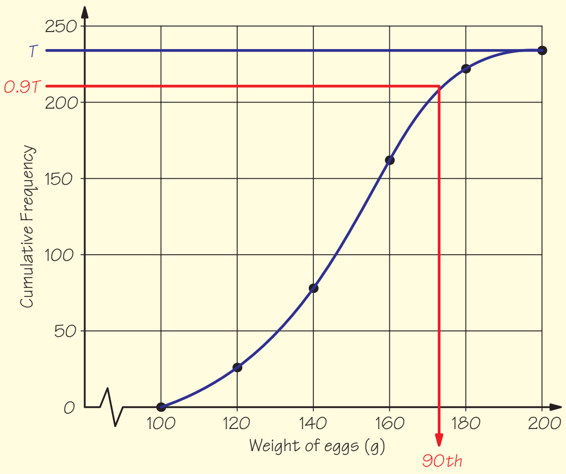
The median and quartiles are specific examples of *percentiles*, which tell you the data value at a given percentage of the way through the data when the items are arranged in ascending order. The median is the 50th percentile and the lower quartile is the 25th percentile.

Worked example 16.8

(a) Find the 90th percentile of the mass of eggs from Worked examples 16.4–16.7.

(b) The top 10% of eggs are classed as 'extra large'. What range of masses corresponds to extra large?

The total frequency is 234, so draw a line across from the y-axis at $0.9 \times 234 \approx 211$. Where this horizontal line meets the cumulative frequency curve, draw a vertical line downward to the x-axis to find the 90th percentile.



Read off the value on the x-axis.

(a) From the diagram the 90th percentile is approximately 173 g.

Top 10% means 90th percentile and above.

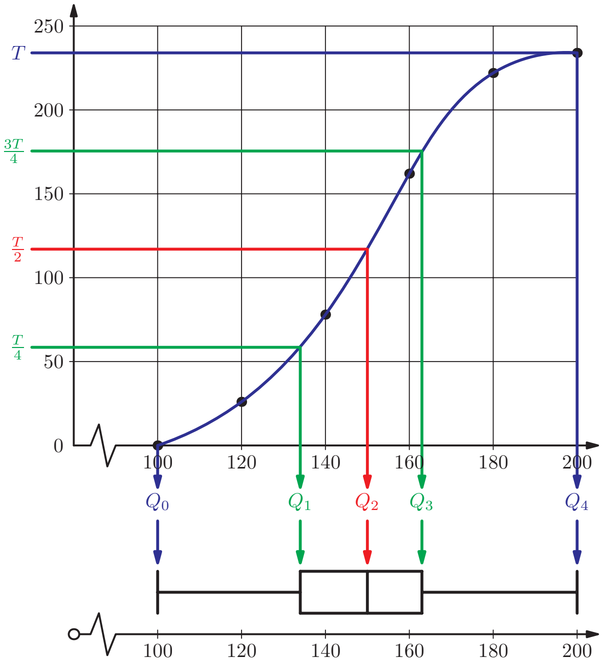
(b) Eggs with weights in the range [173, 200[are classified as extra large.

A useful way of visually representing the information in a cumulative frequency diagram is a **box and whisker plot**, which shows the median and lower and upper quartiles in a 'box', and the smallest and largest data values at the ends of 'whiskers'. The IQR is the length of the central box.

To obtain a box and whisker plot from a cumulative frequency graph, draw the same lines you would for finding the median and quartiles, but extend the vertical lines beyond the *x*-axis.

Worked example 16.9

Draw a box and whisker plot for the eggs data from Worked examples 16.4–16.8.



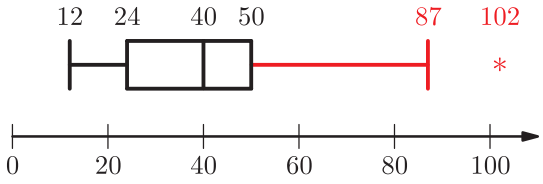
Since quartiles and the interquartile range are less influenced by extreme data values, they can be used to define such 'outliers'.

KEY POINT 16.3

An **outlier** is any data value that is more than 1.5 × IQR above the upper quartile or more than 1.5 × IQR below the lower quartile.

An outlier is usually marked with a cross on a box and whisker plot; the whiskers are ended at the most extreme data values that are not outliers.

So, for a data set with least value 12, lower quartile 24, median 40, upper quartile 50 and highest two values 87 and 102, the box and whisker, graph would be represented as shown below.



$$\text{IQR} = 26, \ \text{Q3} = 50$$

$$\text{Q3} + 1.5\ \text{IQR} = 89$$

End whisker at 87, label outlier 102 with $*$

Exercise 16D

1. Draw cumulative frequency diagrams for each of the data sets in Exercise 16C question 3. Hence estimate the median and the interquartile range.

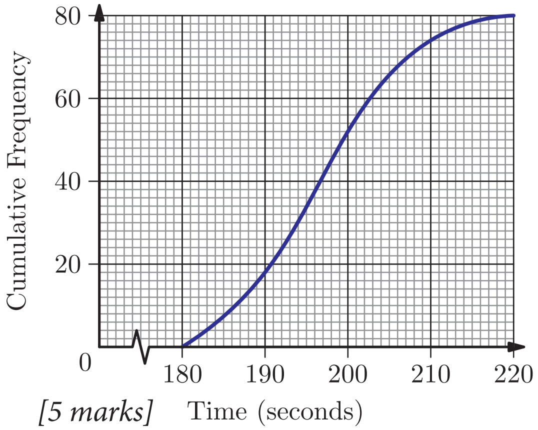
2. For each of the data sets in question 1 (Exercise 16C question 3) draw a box and whisker plot.

3. 80 students were asked to solve a simple word puzzle, and the times they took, in seconds, were recorded. The results are shown in a cumulative frequency graph.

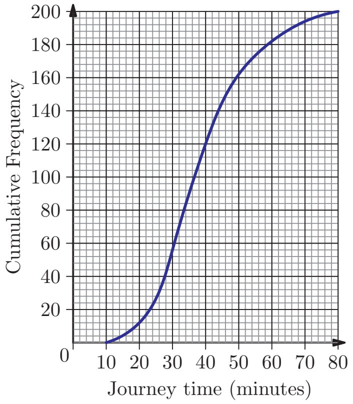
   (a) Estimate the median.

   (b) Estimate the interquartile range.

   (c) If the middle 50% of students took between c and d seconds to solve the puzzle, write down the values of c and d.



[5 marks]

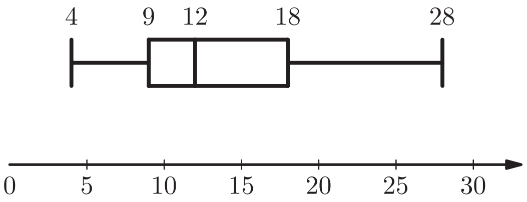


4. The cumulative frequency curve shows the amount of time 200 students spend travelling to school.

(a) Estimate the percentage of students who spend between 30 and 50 minutes travelling to school.

(b) If 80% of the students spend more than x minutes travelling to school, estimate the value of x. *[6 marks]*

5. From the box and whisker plot, state the median and interquartile range.



[3 marks]

6. The table summarises the number of pages in 200 books in a library.

No. pages	100–199	200–299	300–399	400–499	500–599	600–699	700–799
Frequency	12	36	42	53	33	20	4

(a) Estimate the mean number of pages.

(b) Fill in the following cumulative frequency table.

No. pages	199	299	399	499	599	699	799
Cumulative frequency	12	48					200

(c) On graph paper plot a cumulative frequency graph representing the data.

(d) Estimate the median of the data.

(e) Estimate the interquartile range of the data.

(f) Estimate the percentage of books with more than 450 pages. *[14 marks]*

16E Histograms

Another way of displaying data is to draw a histogram, which shows clearly how the data is distributed. At first sight a histogram may look just like a bar chart – in both types of diagram, the size of each bar represents the frequency. However, a bar chart is used for categorical or discrete data, and the width of the bars does not matter. A histogram, on the other hand, is used for continuous data (where data items may take any real value in a certain range); the axis along the bottom of the histogram represents a continuous interval of real numbers, and each bar is positioned over the entirety of the group to which it corresponds.

See Prior Learning section X on the CD-ROM for a review of bar charts.

Why are diagrams more useful than raw data? Statistical diagrams can be hugely informative, but they can also be very misleading. See if you can find some examples of misleading statistical diagrams. Are there any common features which lead you to a particular interpretation of the data?

Histograms can be drawn for grouped data where the groups are of unequal size. In this case, the area rather than the height of each bar represents the frequency associated with that group. You will not be expected to deal with this situation in the International Baccalaureate® course.

Worked example 16.10

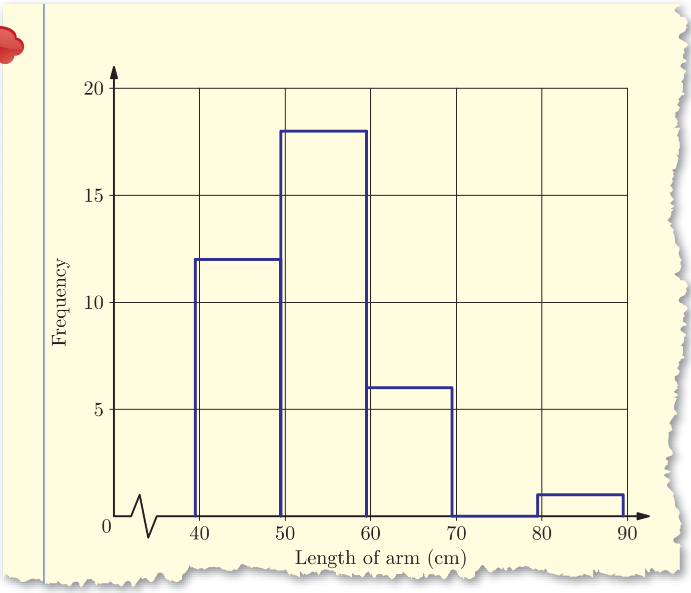
Draw a histogram to represent the following data.

Length of arm (to the nearest cm)	Frequency
40–49	12
50–59	18
60–69	6
70–79	0
80–89	1

→

continued . . .

The left and right ends of each bar are positioned at the lower and upper interval boundaries. Since the measurements have been rounded to the nearest cm, the group '40–49' actually represents all data values in the interval [39.5, 49.5 [,so 39.5 and 49.5 determine the boundaries of the first bar. Similarly, the second bar has its boundaries at 49.5 and 59.5, and so on. The height of each bar is the frequency of the group.



Exercise 16E

1. Draw histograms to represent the following data sets.

(a) (i) x is the time taken to complete a puzzle, in seconds.

x	Frequency
[0, 15[	19
[15, 30[	15
[30, 45[	7
[45, 60[	5
[60, 75[	3
[75, 90[	1

(ii) x is the birth weight of full-term babies in a delivery ward over the course of a week, in kilograms.

x	Frequency
[2.0, 2.5[	5
[2.5, 3.0[	49
[3.0, 3.5[	84
[3.5, 4.0[	63
[4.0, 4.5[	31
[4.5, 5.0[	9

(b) (i) x is the blood glucose level in a sample of 50 cats, to the nearest mmol l^{-1}.

x	Frequency
1 to 5	7
6 to 10	24
11 to 15	13
16 to 20	5
21 to 25	1

(ii) x is the power consumption of light bulbs, to the nearest watt.

x	Frequency
91 to 95	17
96 to 100	23
101 to 105	42
106 to 110	21
111 to 115	5

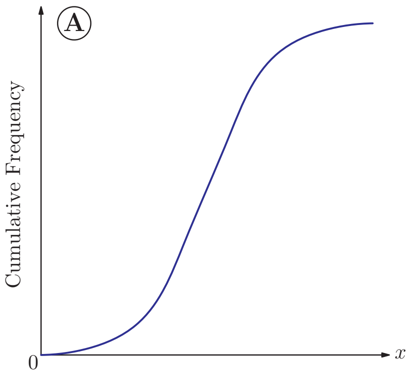
(c) (i) x is the age of teachers in a school.

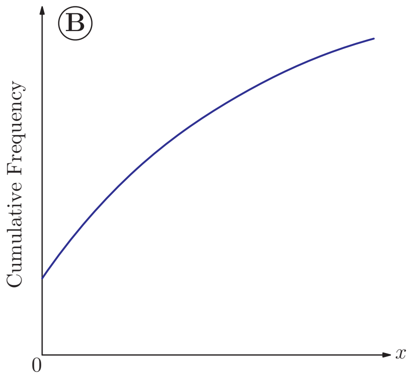
x	Frequency
21 to 25	7
26 to 30	16
31 to 35	23
36 to 40	18
41 to 45	3
46 to 50	8

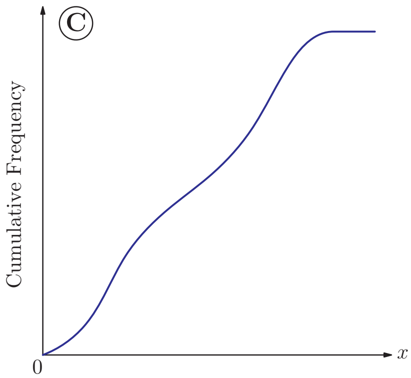
(ii) x is the total bill in a cafe, rounded up to the nearest dollar.

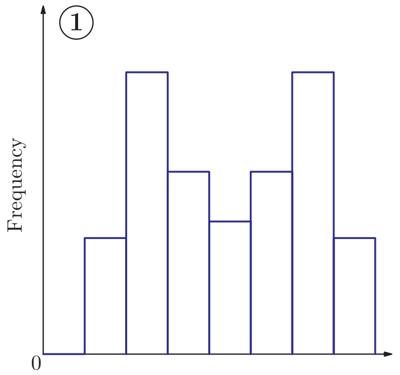
x	Frequency
1 to 10	16
11 to 20	21
21 to 30	43
31 to 40	26
41 to 50	8
51 to 60	2

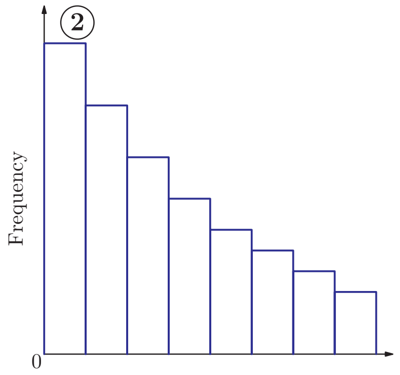
2. Match each histogram with the cumulative frequency diagram drawn from the same data.

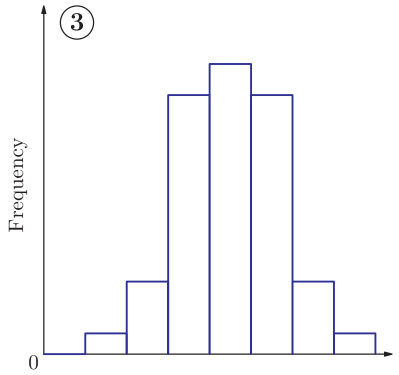




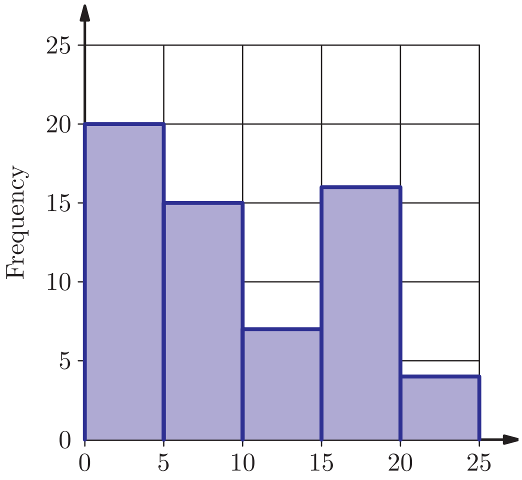








3. From the given histogram calculate the mean and standard deviation of the data.



[6 marks]

16F Constant changes to data

Suppose you need to calculate the mean of the following data:

135408, 135409, 135405

You do not actually have to add up these values; you can just look at the last digit, since this is the only thing that differs between the items. The mean of 8, 9 and 4 is 7, so the mean of the data is 135407.

Effectively, what we have done is subtract 135400 from each of the data items, calculate the mean of the much smaller numbers and then add the 135400 back. This is a common trick for simplifying statistical calculations. It relies on the following fact.

KEY POINT 16.4

If you increase (or decrease) every data item by the value x, all measures of the centre of the data will also increase (or decrease) by x.

Adding on the same quantity to every data item does not change how spread out the data is.

KEY POINT 16.5

If you increase (or decrease) every data item by the value x, all measures of the spread of the data will remain unchanged.

What about the effects of multiplying or dividing all of the data by the same value?

KEY POINT 16.6

If you multiply (or divide) every data item by the value x, all measures of the centre of the data will also be multiplied (or divided) by x.

KEY POINT 16.7

If you multiply (or divide) every data item by the value x, all direct measures of the spread of the data will also be multiplied (or divided) by x.

Since standard deviation is multiplied by x, it follows that the variance will be multiplied by x^2.

Worked example 16.11

A taxi driver records the distance travelled per journey, d, in kilometres. The cost of each journey, c, is given by \$4 plus \$3 per km travelled. On one particular day the average distance travelled was 2.9 km, with a standard deviation of 1.1 km. Find the mean and standard deviation of the cost per journey on that day.

Express c in terms of d.

$c = 3d + 4$

The mean is transformed in the same way as each individual value in the data set.

$$\begin{aligned}\bar{c} &= 3\bar{d} + 4\\ &= 3 \times 2.9 + 4\\ &= 12.7\end{aligned}$$

The standard deviation is only affected by the multiplication.

$$\begin{aligned}\sigma_c &= 3\sigma_d\\ &= 3 \times 1.1\\ &= 3.3\end{aligned}$$

Exercise 16F

1. In each of the following situations find $\bar{y}$ and σ_y.

(a) (i) $y = x - 25,\ \bar{x} = 28,\ \sigma_x = 14$

(ii) $y = x + 5,\ \bar{x} = 12,\ \sigma_x = 3$

(b) (i) $y = 2x,\ \bar{x} = 9.3,\ \sigma_x = 2.4$

(ii) $y = \frac{x}{7},\ \bar{x} = 49,\ \sigma_x = 14$

(c) (i) $y = 5x - 2,\ \bar{x} = 0,\ \sigma_x = 1$

(ii) $y = 3x + 7,\ \bar{x} = 9,\ \sigma_x = 10$

(d) (i) $y = 3(x - 2) + 1,\ \bar{x} = 4,\ \sigma_x = 4$

(ii) $y = 4(x + 1),\ \bar{x} = 10,\ \sigma_x = 3$

2. In each of the following situations find the median and interquartile range of b.

(a) (i) a has median 2.3 and IQR 2.4; $b = a + 4$

(ii) a has median 7.1 and IQR 2; $b = a - 3$

(b) (i) a has median 9 and IQR 3; $b = 2.5a$

(ii) a has median 8 and IQR 6; $b = 1.2a$

(c) (i) a has median 22 and IQR 7; $b = 2.4a + 1.8$

(ii) a has median 10 and IQR 8; $b = 9a - 4$

3. Consider the data set $\{x-2,\ x,\ x+1,\ x+5\}$.

(a) Find the mean of this data set in terms of x.

Each number in the above data set is now decreased by 6.

(b) Find the mean of the new data set in terms of x. *[4 marks]*

4. In some countries the fuel efficiency of a car is measured in miles per gallon (mpg); in other countries it is measured in kilometres per litre (kpl). 1 mile per gallon is equivalent to 0.354 kilometres per litre. A certain make of car has a median efficiency of 32 mpg with variance 60 mpg^2. Find the median and variance of the efficiency in kilometres per litre. *[4 marks]*

5. A website gives the following instructions for converting temperatures in Celsius to temperatures in Fahrenheit:

Take the temperature in Celsius and multiply by 1.8.

Add 32.

The result is in degrees Fahrenheit.

(a) The mean temperature in a fridge is 4°C. Find the mean temperature in Fahrenheit.

(b) The interquartile range of the temperatures in the fridge is 2°F. Find the interquartile range of the temperature in Celsius. *[5 marks]*

6. The variable x has median 20 and interquartile range 10. The variable y is related to x by $y = ax - b$. Find the relationship between a and b so that the median of y equals the interquartile range of y. *[5 marks]*

7. Key point 16.7 assumes that x is positive. Investigate what happens when x is negative. Find a rule that works for all values of x.

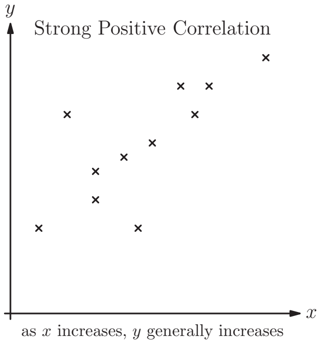
16G Correlation

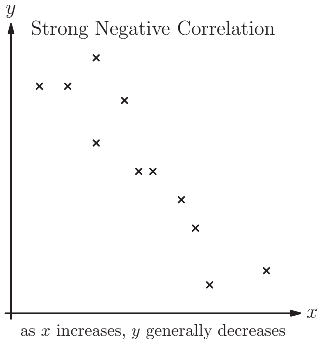
So far we have been focusing on one variable at a time, such as the mass of eggs or the time taken to complete a puzzle. Now we will look at two variables, represented by two data sets, and investigate whether there is a relationship between them. The two sets of data are collected such that from each individual source (a person, for example) we record two values, one for each variable (for example, age and mass). Data that comes in pairs in this fashion is said to be **bivariate**.

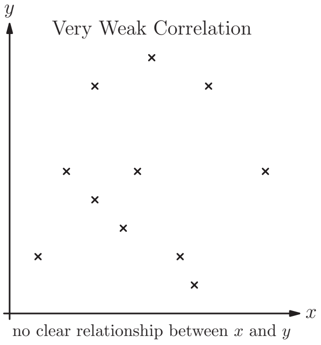
Two variables may be *independent*: knowing the value of one gives us no information about the other – for example, the IQ and house number of a randomly chosen person. Alternatively, there may be a fixed relationship between the variables: once we know the value of one, we can determine exactly the value of the other – for example, the length of a side of a cube and the volume of the cube. Usually, however, the situation is somewhere in between: if we know the value of one variable, we can make a better guess at the value of the other variable, but we cannot be absolutely certain – for example, the total mark achieved by a student in paper 1 of an examination and their mark in paper 2. Where the relationship between two variables lies on this spectrum from completely independent to totally deterministic is called the **correlation** of the two variables.

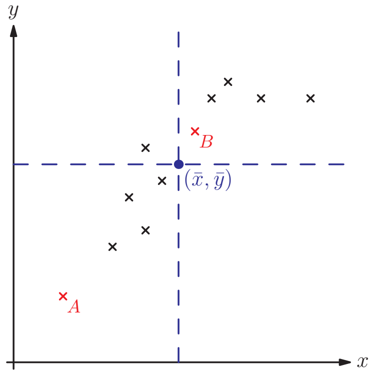
In this course we shall focus on *linear* correlation – the extent to which two variables X and Y are related by a relationship of the form $Y = mX + c$. If the gradient m of the linear relationship is positive, we say that the correlation is positive; if the gradient is negative, we describe the correlation as negative.

The relationship between two variables is best illustrated using a scatter diagram.









Rather than simply describing the relationship in words, we can try to find a numerical value to represent the linear correlation. The idea is to split a scatter diagram into quadrants around the mean point.

If there is a positive linear relationship, we would expect most of the data points to lie in quadrant 1 and quadrant 3. Points lying in those regions should increase our measure of correlation, while points lying in quadrants 2 and 4 should decrease the measure. We do not, however, want all points to be treated equally: point A seems to provide stronger evidence of a positive linear relationship than point *B*, so we would like it to count more.

KEY POINT 16.8

The **product–moment correlation coefficient**, usually denoted by r, is a measure of the strength of the relationship between two variables.

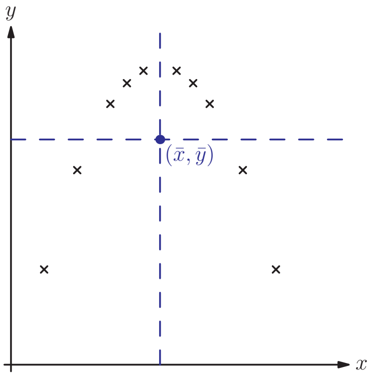
EXAMINERS' HINT

In the International Baccalaureate Standard Level Mathematics course, you will only be expected to find this coefficient using your calculator. See Calculator Skills sheet 14 on the CD-ROM for instructions.

r can take values between -1 and 1 inclusive. You need to know how to interpret the value of r that you find:

Value of r	Interpretation
$r \approx 1$	Strong positive linear correlation
$r \approx 0$	No linear correlation
$r \approx -1$	Strong negative linear correlation

Just because $r = 0$ does not mean that there is no relationship between the two variables – only that there is no *linear* relationship. The graph alongside shows bivariate data with a correlation coefficient of zero, but clearly there is a relationship between the two variables.



While the product–moment correlation coefficient can provide a measure of correlation between two variables, it is important to realise that even when r is close to ±1, a change in one variable does not necessarily cause a change in the other. The correlation might simply be due to coincidence or the influence of a third, hidden variable. For example, there may be a strong correlation between ice-cream sales and instances of drowning at a certain beach, but this does not imply that eating more ice cream leads to increased chances of drowning; rather, the hidden variable of temperature could cause both variables to rise.

There are actually many different measures of correlation. The r in Key point 16.8 is referred to as Pearson's product–moment correlation coefficient, or PPMCC.

Worked example 16.12

The following data were collected:

Score in a maths test (M)	Number of hours spent revising (R)
42	1.0
50	1.25
67	2.0
71	2.3
92	3.0

(a) Calculate the product–moment correlation coefficient.

(b) Interpret the value you found in part (a).

(a) From GDC, $r = 0.996$

The value is close to +1.

(b) This suggests a strong positive linear correlation between **M** and **R**.

Exercise 16G

1. Find the correlation coefficient for each of the following bivariate data sets.

(a) (i) $(2,-5),(0,3),(8,12),(5,19),(4,10),(10,24)$

(ii) $(1,0),(1,3),(2,6),(2,2),(4,4),(5,9)$

(b) (i) $(3,15),(17,9),(22,10),(33,7)$

(ii) $(22,50),(54,19),(100,0),(93,12)$

(c) (i) $(-2,3),(0,0),(2,1),(3,5),(4,2)$

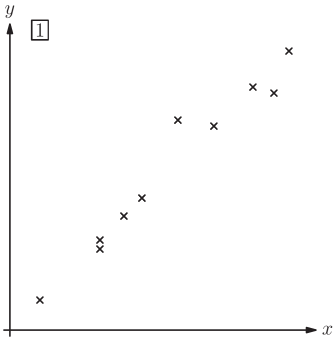
(ii) $(5,1),(9,3),(7,-2),(8,8)$

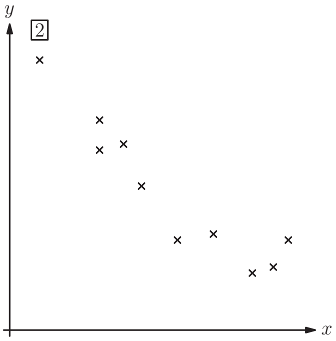
(d) (i) $(1,3),(2,5),(3,7),(5,11)$

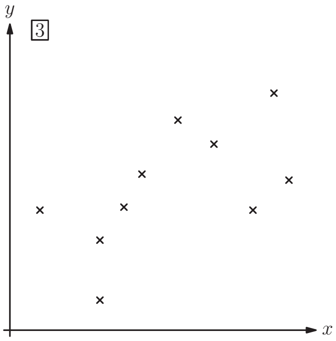
(ii) $(9,1),(4,6),(5,5),(11,-1)$

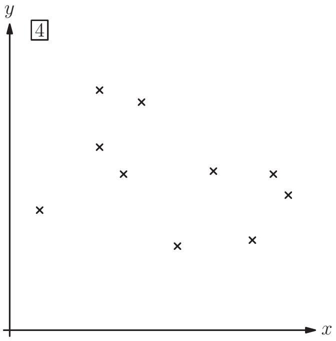
2. Match the scatter diagrams with the following values of r:

$A: r = 0.98$ $B: r = -0.34$ $C: r = -0.93$ $D: r = 0.58$









3. The following table gives data on life expectancy and average CPU speed of PCs over the past 20 years:

Year	Processor speed (Hz)	Life expectancy (months)
1990	6.6×10^7	74
1995	1.2×10^8	75
2000	1.0×10^9	77
2005	1.8×10^9	79
2010	2.4×10^9	80

(a) Calculate the correlation coefficient between processor speed and life expectancy.

(b) Interpret the value you found in (a).

(c) Does this result imply that CPU speed affects life expectancy? *[5 marks]*

4. A road safety group has tested the braking distance of cars of different ages.

Age in years	Braking distance in metres
3	31.3
6	38.6
7	40.1
7	35.1
9	48.4

(a) Find the correlation coefficient between a car's age and braking distance.

(b) Interpret this value.

(c) Farah says that this provides evidence that older cars tend to have longer stopping distances. State with a reason whether you agree with her. *[6 marks]*

5. The time spent in education and the income of six different 45-year-olds is recorded:

Years in education	Income ($)
12	64 000
14	31 000
14	36 000
18	54 000
18	62 000
20	48 000

(a) Find the correlation coefficient between these two variables.

(b) Gavin says that this provides evidence that spending more time in education means you will be paid more. State with two reasons whether you agree with him. *[5 marks]*

6. The following table shows the time taken for a chemical reaction to complete at different temperatures. The temperatures were recorded in both degrees Celsius and degrees Fahrenheit.

Temperature in °C (C)	Temperature in °F (f)	Time in seconds (t)
10	50	43
15	59	39
20	68	34
25	77	29
30	86	22
35	95	18
40	104	15

(a) Find the correlation coefficient between c and t.

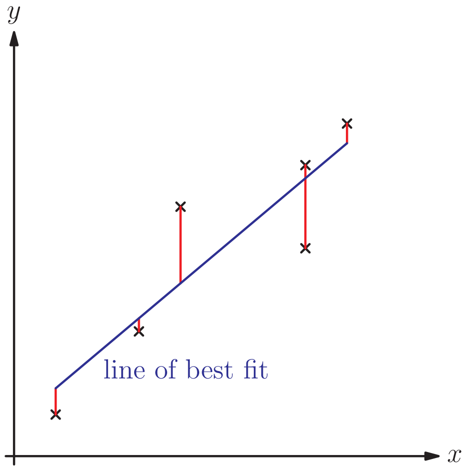
(b) Find the correlation coefficient between f and t.

(c) Comment on your results in parts (a) and (b).

16H Linear regression

Suppose we have established that there is a linear relationship between two variables. We would then want to find the equation that best describes this relationship. To do this we use a method called *least-squares regression*.

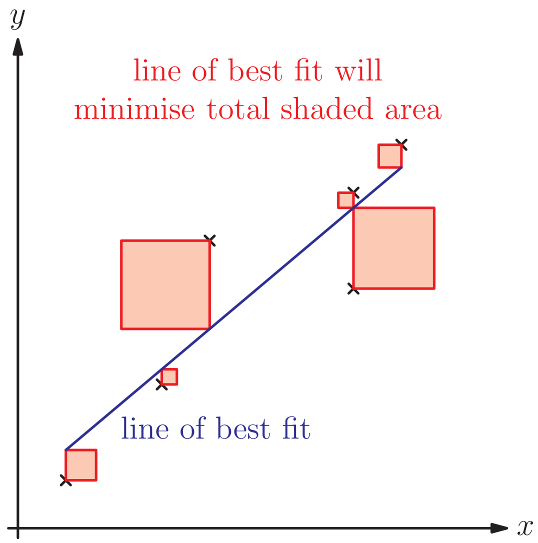
If we assume that two variables X and Y are related by $Y = mX + c$, then for each x-value (i.e. data value for the X variable), we can measure the distance of the corresponding y-value from the line $Y = mX + c$.



We could add up these distances and then try to minimise the total distance by varying the gradient and intercept of the line, but it turns out to be better to minimise the sum of the *squares* of these distances.

EXAM HINT

You may be asked whether a linear model is appropriate for a given set of bivariate data. In exam questions it will always be clear whether the correlation coefficient is strong enough to justify using a linear model (that is, r will be obviously close to or far away from ±1).



EXAM HINT

See Calculator skills sheet 14 on the CD-ROM for instructions.

To do this minimisation requires some fairly advanced calculus. Fortunately, as with the correlation coefficient r, you can use your calculator to obtain the equation of this **line of best fit**, also referred to as a **regression line**.

The line of best fit always passes through the mean point – the point with coordinates $(\bar{x}, \bar{y})$.

KEY POINT 16.9

The variable to be plotted on the x-axis is called the **independent variable**; it is the variable that can be controlled by the experimenter.

The variable to be plotted on the y-axis is known as the **dependent variable**.

Once you have found the line of best fit, you can use it to estimate values that are not among the data already collected. If we estimate values for the Y variable corresponding to values of X that are within the range of x-values already collected, we are **interpolating**. If we extend the regression line beyond the range of the x-values already collected and use it to predict a value of Y outside this range, we are **extrapolating**. It is important to treat the results of extrapolation with caution – there is no guarantee that the pattern will continue beyond the observed values.

Worked example 16.13

The following data were collected:

Score in a maths test (M)	Number of hours spent revising (R)
42	1.0
50	1.25
67	2.0
71	2.3
92	3.0

(a) Which is the independent and which is the dependent variable? Justify your answer.

(b) Find the equation of the regression line.

(a) Time spent revising is the independent variable, as it can be controlled. The score in the test is the dependent variable; we assume it can be affected by the time spent revising.

(b) From GDC: $M = 24.0R + 18.5$

Exercise 16H

1. Find the line of best fit for each of the following sets of bivariate data.

(a) (i) $(2,-5),(0,3),(8,12),(5,19),(4,10),(10,24)$

(ii) $(1,0),(1,3),(2,6),(2,2),(4,4),(5,9)$

(b) (i) $(3,15),(17,9),(22,10),(33,7)$

(ii) $(22,50),(54,19),(100,0),(93,12)$

(c) (i) $(-2,3),(0,0),(2,1),(3,5),(4,2)$

(ii) $(5,1),(9,3),(7,-2),(8,8)$

(d) (i) $(1,3),(2,5),(3,7),(5,11)$

(ii) $(9,1),(4,6),(5,5),(11,-1)$

2. In a restaurant, the regression line relating the number of calories, d, in a dessert and the quantity sold, n, is given by $n = 0.14d - 20$. The mean number of desserts sold is 36 and the median number of desserts sold is 32.

(a) Find the mean number of calories in a dessert, or state that it cannot be found.

(b) Find the median number of calories in a dessert, or state that it cannot be found. *[3 marks]*

3. The table shows the average speed v of cars passing positions d at 10 m intervals after a junction.

d **(m)**	***v*** **(km/hr)**
10	12.3
20	17.6
30	21.4
40	23.4
50	25.7
60	26.3

(a) Find the correlation coefficient.

(b) State, with a reason, which is the independent variable.

(c) Using an appropriate regression line, find the value of v when $d = 45$.

(d) Explain why you cannot use your regression line to accurately estimate v when $d = 80$. *[9 marks]*

4. The following data gives the height h of a cake baked at different temperatures T.

T **(°F)**	***h*** **(cm)**
300	16.4
320	17.3
340	18.1
360	16.2
380	15.1
400	14.8

(a) Find the correlation coefficient.

(b) State which variable is the independent variable.

(c) Find the regression line for this data.

(d) State two reasons why it would be inappropriate to use the regression line found in part (c) to estimate the temperature required to get a cake of height 20 cm. *[7 marks]*

5. State whether the following statements are true or false for bivariate data.

(a) If $r = 0$, there is no relationship between the two variables.

(b) If $Y = kX$, then $r = 1$.

(c) If $r < 0$, then the gradient of the line of best fit is negative.

(d) As r increases, so does the gradient of the line of best fit.

6. Data from an experiment is given in the following table:

x	y
–5	25
7	52
–6	35
–8	62
–4	13
–9	89
0	–3
–6	38

(a) Find the correlation coefficient for

(i) y and x

(ii) y and x^2

(b) Use least-squares regression to find a model for the data of the form $y = kx^2 + c$.

[7 marks]

Summary

- Measures of the *centre* of a data set include the **mean**, **median** and **mode**. The *spread* of the data values around the centre can be measured with the **range**, **interquartile range** (IQR = $Q_3 - Q_1$) or **standard deviation** (σ). The square of the standard deviation is called the **variance**. These values can all be worked out using your calculator for large data sets, or by hand for small data sets.
- To calculate the mean from a frequency table, see Key point 16.2.
- To estimate the mean and standard deviation of grouped data, we assume that every data item of a group is at the **mid-interval value** of the group, which is the mean of the upper and lower interval boundaries. The mid-interval value of each group is multiplied by the frequency in that group, and the mean and standard deviation can be worked out as normal using your calculator.
- **Histograms,** which are used for continuous data, provide a useful visual summary of data, giving an immediate impression of centre and spread.
- **Cumulative frequency** is a count of the total number of data items up to a certain value.
- Cumulative frequency diagrams involve plotting the cumulative frequency against the upper bound of each group as well as the leftmost point that represents a frequency of zero. These graphs are useful for finding the median, interquartile range and various percentiles, especially for grouped data. They also facilitate the construction of **box and whisker plots**, which are another good way of summarising data visually.

- If you increase (or decrease) every item in a data set by a value x, all measures of the centre of the data will also increase (or decrease) by x, while all measures of the spread of the data will remain unchanged.
- If you multiply (or divide) every item in a data set by a positive value x, all measures of the centre and all measures of the spread of the data will also be multiplied (or divided) by x.
- **Bivariate data** represents the paired measurements of two variables. The relationship, or correlation, between such variables can be visualised on a scatter diagram.
- The linear correlation is the extent to which two variables, X and Y, are related by a relationship of the form $Y = mX + c$. If m is positive, the correlation is also positive; if m is negative, so is the correlation. A numerical value to represent the linear correlation is the product-moment correlation coefficient.
- The **product–moment correlation coefficient** is a value between -1 and 1 which measures the strength of the linear relationship between two variables. 1 indicates a strong positive correlation.
- We can use least-squares regression to find the line equation of the **line of best fit** (or **regression line**) for the two variables. The line of best fit always passes through the mean point: $(\bar{x}, \bar{y})$
  - the x-variable is the **independent variable**; it is controlled
  - the y-variable is the **dependent variable**; its value depends on x.
- The line of best fit can be used to estimate data that is not among the collected data set:
  - **interpolating** is estimating Y from values of X within the range of the collected x-values
  - **extrapolating** is the process of extending the regression line beyond the range of collected x-values and using this to estimate the y-values; this is not very reliable because there is no guarantee that the pattern will continue beyond the observed values.

Introductory problem revisited

> The magnetic dipole of an electron is measured three times in a very sensitive experiment. The values obtained are 2.000 0015, 2.000 0012 and 2.000 0009. Does this data support the theory that the magnetic dipole is 2?

From these three measurements, the mean magnetic dipole is 2.000 0012, which seems pretty close to 2. However, the standard deviation of the data is 0.000 000 245, so the mean is approximately 5 sample standard deviations away from 2. Therefore, within the natural variation observed, we cannot say with confidence that the magnetic dipole is 2.

The difference between 2.000 0012 and 2 may seem trivial, but it was what inspired Richard Feynman to create a new theory of physics called Quantum Electrodynamics, which did indeed predict this tiny difference from 2! This is an example of theory driving experiment, which in turn creates new theory – the interplay between theoretical mathematics and reality.

Mixed examination practice 16

Short questions

1. For the set of data

115, 108, 135, 122, 127, 140, 139, 111, 124

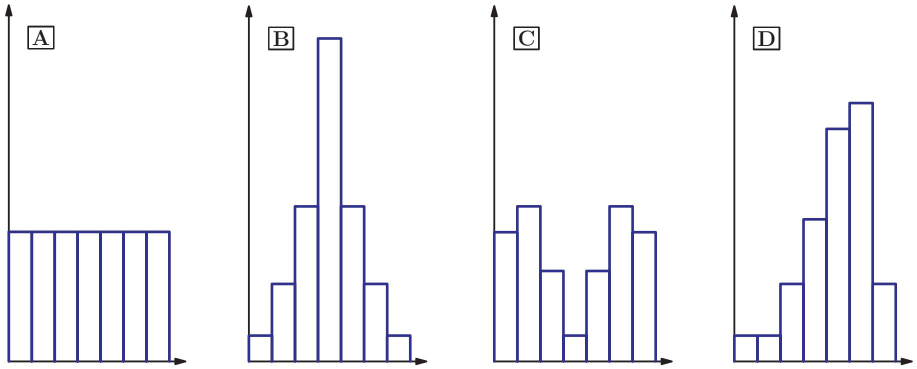
find:

(a) the interquartile range

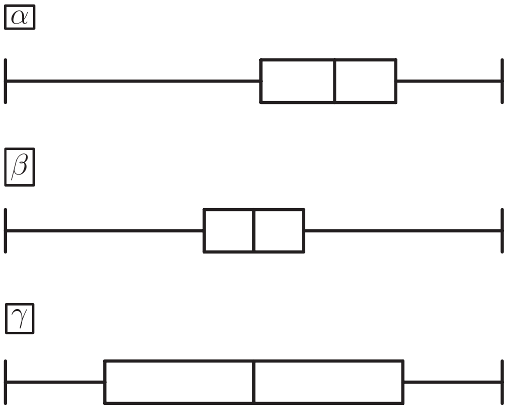
(b) the mean

(c) the variance. *[7 marks]*

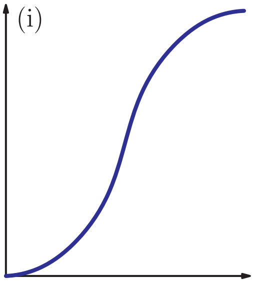
2. Four populations A, B, C and D are the same size and have the same range. Frequency histograms for the four populations are given below.

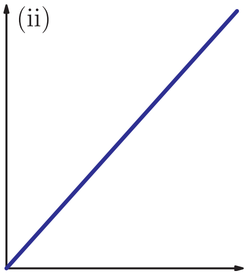


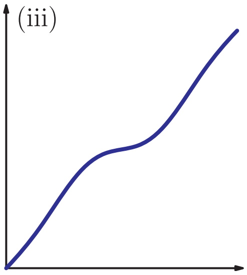
(a) Each of the three box and whisker plots shown corresponds to one of the four populations. Write the letter of the correct population for each of α, β and γ.



(b) Each of the three cumulative frequency diagrams below corresponds to one of the four populations. Write the letter of the correct population for each of (i), (ii) and (iii).



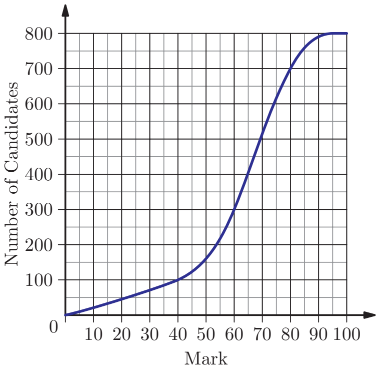




[6 marks]

(*© IB Organization 2006*)

3. A test marked out of 100 is taken by 800 students. The cumulative frequency graph for the marks is shown.



(a) Write down the number of students who scored 40 marks or less on the test.

(b) The middle 50% of test results lie between marks a and b, where $a<b$. Find a and b. *[6 marks]*

(*© IB Organization 2005*)

4. The number of bunches of flowers that a florist sells each day has a mean of 83 and a variance of 60. Each bunch sells for £10. The florist has fixed costs of £100 per day.

(a) Find the relationship between the number of flowers sold (f) and the profit made (p).

(b) Find the mean of p.

(c) Find the variance of p. *[6 marks]*

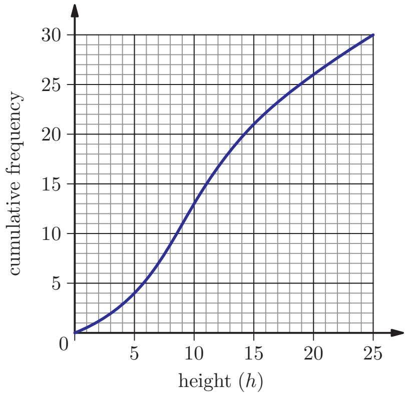
5. Three positive integers a, b and c, where $a < b < c$, are such that their median is 15, their mean is 13 and their range is 10. Find the value of c. *[6 marks]*

Long questions

1. The following is the cumulative frequency diagram for the heights of 30 plants given in centimetres.

(a) Use the diagram to estimate the median height.

(b) Complete the following frequency table.



Height (h)	Frequency
$0 < h \leq 5$	4
$5 < h \leq 10$	9
$10 < h \leq 15$	
$15 < h \leq 20$	
$20 < h \leq 25$	

(c) Hence estimate the mean height. *[8 marks]*

(*© IB Organization 2006*)

2. One thousand candidates sit an examination. The distribution of marks is shown in the following grouped frequency table.

Marks	1–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
Number of candidates	15	50	100	170	260	220	90	45	30	20

(a) Copy and complete the following table, which presents the above data as a cumulative frequency distribution.

Mark	≤10	≤20	≤30	≤40	≤50	≤60	≤70	≤80	≤90	≤100
Number of candidates	15	65					905			

(b) Draw a cumulative frequency graph of the distribution, using a scale of 1 cm for 100 candidates on the vertical axis and 1 cm for 10 marks on the horizontal axis.

(c) Use your graph to answer parts (i)–(iii) below.

(i) Find an estimate for the median score.

(ii) Candidates who scored less than 35 were required to retake the examination. How many candidates had to retake?

(iii) The highest-scoring 15% of candidates were awarded a distinction. Find the mark above which a distinction was awarded.

[16 marks]

(*© IB Organization 1999*)

3. The owner of a shop selling hats and gloves thinks that his sales are higher on colder days. He records the temperature and the value of goods sold on a random sample of 8 days:

Temperature (°C)	13	5	10	–2	10	7	–5	5
Sales (£)	345	450	370	812	683	380	662	412

(a) Calculate the correlation coefficient for the two sets of data.

(b) Suggest one other factor that might cause the sales to vary from day to day.

(c) Explain why temperature could be considered the independent variable.

(d) Find the equation of the regression line.

(e) Use this line to estimate the sales when the temperature is 0°C.

(f) Explain why it would not be appropriate to use the regression line to estimate sales when the temperature is –20°C. *[9 marks]*

4. A shopkeeper records the amount of ice cream sold on a summer's day along with the temperature at noon. He does this for several days, and gets the following results:

Temperature (°C)	**Ice creams sold**
26	41
29	51
30	72
24	23
23	29
19	12

(a) Find the correlation coefficient.

(b) By finding the equation of the appropriate regression line, estimate the number of ice creams that would be sold if the temperature were 25°C.

(c) Give two reasons why it would not be appropriate to use the regression line from part (b) to estimate the temperature on a day when no ice creams are sold. *[8 marks]*