2 Representing and describing data: descriptive statistics

Statistics is concerned with the collection, analysis and interpretation of quantitative data. Statistical representations and measures allow us to represent data in many different forms to aid interpretation. Both statistics and probability provide important representations which enable us to make predictions, valid comparisons and informed decisions.

How can scientists determine whether a new drug is likely to be a successful cure?

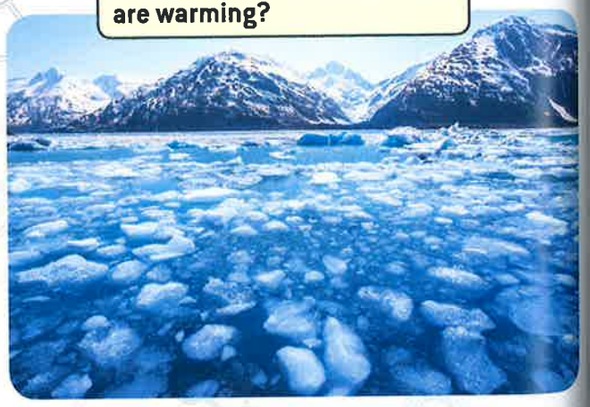
Concepts

- Representation
- Validity

Microconcepts

- Population
- Bias
- Samples, random samples, sampling methods
- Outliers
- Discrete and continuous data
- Histograms
- Box-and-whisker plots
- Cumulative frequency graph
- Measures of central tendency and dispersion
- Skewness
- Scatter graphs
- Correlation

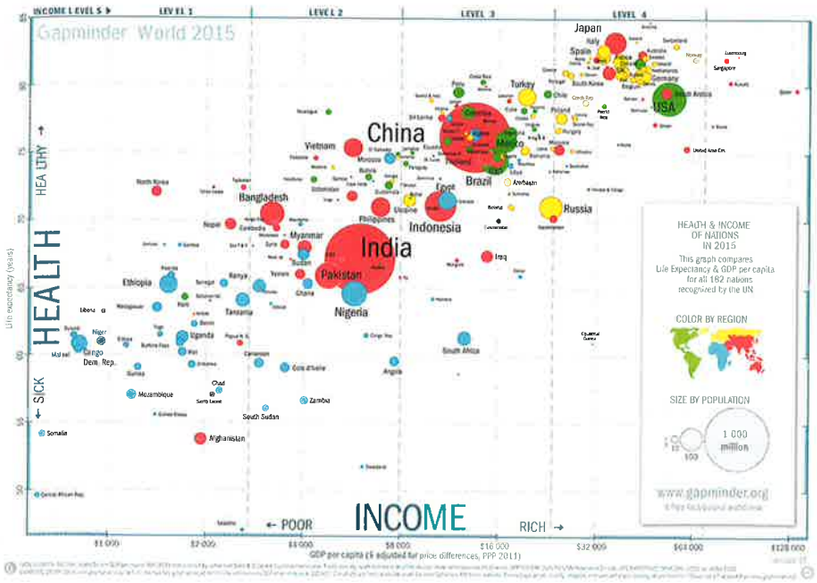
How can we tell if the oceans are warming?



How can a football coach determine whether a particular strategy is likely to be successful?

How can you persuade a potential customer that your product is better than the competition?

The picture shows a graph of GDP per capita (gross domestic product per person) and life expectancy taken from Gapminder (www.gapminder.org). Click the icon for a spreadsheet of the complete data for this graph.



- Name four pieces of information represented in this graph.
- How do you think this data could have been collected? How exact do you think it might be?
- Can you identify any relationships from the graph?
- Do you need to use all the data for analysis or can you just use a sample of the data?
- Do you find anything surprising in the graph?
- Describe the scale on the x-axis. Why do you think it has been done like that?

Developing inquiry skills

Write down any similar inquiry questions you might ask to investigate the relationship between two different quantities, for example, GDP per capita and infant mortality or life expectancy and population.

What questions might you need to ask in these scenarios which differ from the scenario where you are investigating. How are these questions different from those used to investigate the relationship between GDP and life expectancy?

Think about the questions in this opening problem and answer any you can. As you work through the chapter, you will gain mathematical knowledge and skills that will help you to answer them all.

Before you start

You should know how to:

1 Find the mean, median or mode of a set of numbers by hand, and the mean and median using statistical summaries on the GDC.

2 Find the mean, median or mode from a frequency table by hand or using statistical summaries on the GDC.

Skills check

Click here for help with this skills check

1 Find the mean, median and mode of these numbers by hand:
14, 15, 17, 22, 26, 22, 21, 16, 17, 22

2 Find the mean and median of the numbers above using your GDC.

2.1 Collecting and organizing data

Qualitative data is non-numerical, eg "it was fun", "blue".

Quantitative data is numerical. Quantitative data can be **discrete** or **continuous**.

Discrete data is data which takes specific (discrete) values, eg "number of accidents", "points in the IB diploma".

Continuous data is data which can take a full range of values, eg "height", "speed".

International-mindedness

Ronald Fisher (1890–1962) lived in the UK and Australia and has been described as "a genius who almost single-handedly created the foundations for modern statistical science". He used statistics to analyse problems in medicine, agriculture and social sciences.

Investigation 1

1 In a certain school, grades in the final IB HL Mathematics exam are given:

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

Factual Is the data discrete or continuous?

2 A frequency table can be created for the data in question 1.

Copy and complete the frequency table:

Grade, g	3	4	5	6	7
Frequency	1	5			

3 Heights, to the nearest centimetre, of primroses in the garden are measured and given:

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

Factual Is the data continuous or discrete?

4 A frequency table can be created for this data.

Copy and complete the grouped frequency table:

Height, h, in cm	$2.5 \le h < 3.5$	$3.5 \le h < 4.5$	$4.5 \le h < 5.5$	$5.5 \le h < 6.5$	$6.5 \le h < 7.5$
Frequency	1	5			

Explain why the two cases are different.

5 Conceptual When would discrete or continuous data occur? Think about how you would obtain the data and whether you would need any particular tools.

6 Ages, in years, of children in a nursery class are given:

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

Factual Is the data continuous or discrete?

7 Explain why the frequency table would again be different and create the table.

Note: Age is a special case of data as it can also be regarded as discrete when you are considering only completed years.

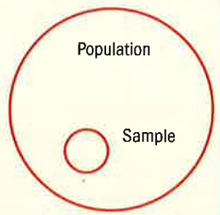
HINT

Think about when you become 3 and when you stop being 3.

A **population** includes all members of a defined group.

A **sample** is a subset of the population, a selection of individuals from the population.

Biased sampling is where the method may cause you to draw misleading conclusions about the population.



For example, drawing conclusions about people's use of public transport from a survey conducted in a train station.

The sampling methods below can be used if you can list every member of the population.

Simple random sampling: every member of the population is equally likely to be chosen. For example, allocate each member of the population a number. Then use random numbers to choose a sample.

Systematic sampling: find a sample of size n from a population of size N by selecting every kth member where $k = \frac{N}{n}$ to the nearest whole number.

For example, choosing every 15th student from the school register to find a sample of 100 from a population of 1500.

Stratified sampling: is selecting a random sample where numbers in certain categories are proportional to the numbers in the population.

For example, if 20% of students in a school were in Grade 7, then you would choose 20% of your sample from Grade 7.

Example 1

An educational psychologist recorded the IQs of 200 people. The 200 people are sorted in order of age and a sample of 40 is selected from the list and the mean of the sample is found, which will be used as an estimate for the mean of the whole population.

Identify the type of sample used in each case. Give any advantages and/or disadvantages in using this method to estimate the population mean.

a Take the first 20 numbers and the last 20 numbers.

b Take every fifth IQ.

c Generate 40 random numbers between 1 and 200 and use them to select the sample.

Continued on next page

a	This is a biased sample since it is not random. It is easy to use but it is unreliable. Only the youngest and oldest will be selected.	
b	This is systematic sampling since every fifth entry is selected. It is easy to use and is not time consuming.	You can start at any number and choose every fifth number until you have 40 numbers in total. It is possible to have different answer for this depending on the starting place. The fact you can start with any number means that each person has an equal chance of being selected.
c	This is a random sample. Each time a random sample is chosen, different numbers will be generated. The advantage is that it is truly random. The disadvantage is that it will produce different numbers each time and it can be time-consuming.	

If you are not able to list every member of the population then you have to generate a sample to represent the population in the best way you can.

Quota sampling: decide how many members of each group you want to sample and take samples from the population until you have a large enough sample for each group.

For example, the school canteen is considering introducing a new lunch menu and would like feedback from the students. The school has 250 boys and 300 girls and so the canteen manager decides to interview 25 boys and 30 girls to find out their opinion of the new menu. He stands at the entrance to the canteen and interviews the first 25 boys and 30 girls that come into the canteen.

Quota sampling is not random. It can be biased and unreliable. The advantage is that it is inexpensive, easy to perform and saves time.

However, it is more reliable than convenience sampling where people are selected based on availability and are not representative of the population. This type of sampling is also a non-probability sample and can also be biased and unreliable.

Convenience sampling: take samples from the members of the population that you have access to until you have a sample of the desired size.

For example, asking people at a shopping centre to fill out a survey until you have 50 responses.

Investigation 2

All of the students at Valley High School must join either the Robotics Club or the Astronomy Club. The Headteacher at Valley High School wants to know if the Robotics Club members are performing better in Mathematics than the Astronomy Club members. He finds the scores of all students in the school in their University entrance exam for the last five years.

1 Give reasons why the Headteacher would have removed the names of the students.

2 Give reasons why you may prefer to do further analysis of the data using a sample, rather than all the data available.

3 **Conceptual** Why do we need to take a sample from the population?

Five suggestions are made as to how a sample of size 30 could be taken to see if Robotics club members are performing better than Astronomy Club members in Mathematics.

- Randomly select 30 students from all the results.
- Randomly select 30 students from the 2017 cohort.
- Randomly select 15 Robotics Club members and 15 Astronomy Club members from all the results.
- Use stratified sampling to randomly select students from each year according to the numbers in that year.
- Randomly choose a student and then choose every 18th student from there (upon reaching the end of the data, return to the beginning).

4 For the first suggestion, a GDC can generate 30 random integers from 1 to 545. Use this method to create a sample of 30 students' Mathematics result, along with their club. Consider what you will do if a number is repeated, or if the number points to a missing data item.

5 **Factual** Does your sample have an equal number of Robotics Club and Astronomy Club members?

6 **Factual** Does your sample have representative numbers from the different years?

7 **Conceptual** Can you identify any bias created when you have used this method of sampling?

8 **Conceptual** Can you think of different occasions where this type of sampling could have created extreme bias?

Continued on next page

HINT
Make sure you can generate random numbers on your GDC.

9 Use similar methods to collect a sample of 30 students' results using suggestion 2 and suggestion 3.
For each suggestion consider the guiding questions above and state the advantages of that particular method of sampling in this example.
Conceptual Can you identify any additional bias that might have been created by the sampling method or a bias that has been eliminated?

10 For suggestion 4, explain why you need to select four students from 2013, six from 2014, five from 2015, seven from 2016 and eight from 2017. Collect the sample for suggestion 4.

11 For suggestion 5, explain why you would take every 18th student.
- **i** What if you took every 10th student?
- **ii** Collect the sample for suggestion 5.
- **iii** Explain why this method also selects students according to the number in a particular year.

12 Discuss the pros and cons of each suggestion.
Which suggestion do you think will be the most representative of the population – that is all Robotics Club and Astronomy Club members in Valley High School?

13 Conceptual What are the different sampling methods and how do they provide meaningful insight into the general population?

Keep your results – you will be using them in sections 2.2 and 2.3.

Exercise 2A

1 It is suggested that the English results at Valley High School have been improving in the last five years. Choose a suitable sampling method to collect data that might be used to give evidence as to whether this is true or not.
Collect the sample and save it for use in sections 2.2 and 2.3.

2 A dog kennel has 120 dogs. The ages of the dogs, in order, are shown below.
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3,
3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5,
5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7,
7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10,
10, 11, 11, 11, 11, 12, 12, 12, 13, 13, 13, 14, 14, 15, 16, 16.
- **a** Construct a table of values to represent the ages of the dogs and the frequency of their ages.
- **b** A sample of the population was taken to check the health of the dogs in the kennel. A systematic sample of every 5 dogs was used. How many dogs were taken for the sample?
- **c** In order to take a sample of 50 dogs, stratified by their age, state the number of dogs of each age you need to select.

3 In a poll to predict the results in an election, a sample from the population is taken with the intention of providing an accurate forecast. Explain why the sample would be stratified, and suggest different strata that might be important.

Developing inquiry skills

In the opening section you saw data from different countries regarding life expectancy, GDP, population and region. If you were to take a sample from all those countries to draw conclusions, how might you stratify your sample in order to reduce bias?

2.2 Statistical measures

Measures of central tendency (or averages)

- The most common measures of central tendency are the mean, median and mode.
- The **mode** of a data set is the value that occurs most frequently. There can be no mode, one mode, or several modes.
- The **median** of a data set is the value that lies in the middle when the data are arranged in size. When there are two middle values then the median is the midpoint between the two values.
- The **mean** of a data set is the sum of all the values divided by the number of values. For a discrete data set of n values the formula is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ where $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$. Σ means "the sum of". For example, the mean of the numbers 3, 4, 8, 12, 16 is the sum (Σ) of the numbers divided by 5.
- When there is a frequency table, you need to use the data values and the corresponding frequencies to calculate the mean.

TOK

Do different measures of central tendency express different properties of the data?

How reliable are mathematical measures?

HINT

Make sure you know how to find the statistical summaries on your technology – for both lists of data and for frequency tables.

Example 2

The number of days of sunshine in Helsinki in January was recorded for a period of 35 years and the data is given in the frequency table:

Number of days of sunshine	Number of years
3	1
4	2
5	1
6	2
7	7
8	5
9	9
10	8

- **a** State the modal number of days of sunshine in January in Helsinki for the period.
- **b** Calculate the mean number of days of sunshine in January in Helsinki for the period.
- **c** Determine the median number of days of sunshine in January in Helsinki for the period.
- **d** Comment on how a "day of sunshine" might be defined.

Continued on next page

a 9 days	There are 9 years where there were 9 days of sunshine.
b 7.94 days	Remember that the frequency table represents the data set {3, 4, 4, 5, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10}. $\frac{3+(2\times4)+5+(2\times6)+(7\times7)+(5\times8)+(9\times9)+(8\times10)}{35}$ $=\frac{278}{35}=7.94\,(3\text{ s.f.})$ Or the answer can be obtained directly from the statistical summaries on the GDC.
c 8 days	For 35 data items, the median will be the 18th. The 14th to 18th data items are all 8 so the median is 8 days. Or the answer can be obtained directly from the statistical summaries on the GDC.
d Days with more than 3 hours where you can see the sun.	Many interpretations possible.

Measures of dispersion

- Measures of dispersion measure how spread out a data set is.
- The most common measure of dispersion is the **range**, which is found by subtracting the smallest number from the largest number.
- The standard deviation, σ_n, gives an idea of how the data values are related to the mean. The standard deviation is also known as the root-mean-squared deviation; its formula is:

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}$$

EXAM HINT
In examinations you will use your technology to find the standard deviation.

From the first formula it can be seen that the standard deviation is found by considering the distance of each point from the sample mean. If the differences are larger then the value of σ will also be greater.

The **variance** is the standard deviation squared: σ_n^2.

TOK
Why are there different formulae for the same statistical measures like mean and standard deviation?

While the standard deviation is useful for interpreting the spread of data about the mean, other statistical processes such as the least squares regression, probability theory and investments use the variance.

If the data is arranged in ascending order the nth percentile is the piece of data that is n% along the list. Hence the median is also the 50th percentile. The **interquartile range** (IQR) is the **upper quartile**, Q_3, minus the **lower quartile**, Q_1.
When the data are arranged in order, the lower quartile is the median data point of the lower half of the data (at the 25th percentile) and the upper quartile is the median data point of the upper half of the data (at the 75th percentile).

- Outliers are extreme data values, or the result of errors in reading data, that can distort the results of statistical processes.
- Outliers can affect the mean by making it larger or smaller, but most likely will not affect the median or the mode.
- Outliers can affect the standard deviation by making it larger, but they most likely will not affect the interquartile range.

An **outlier** is defined as a data item that is more than 1.5 × IQR below Q_1 or above Q_2.

Example 3

The number of days of precipitation in January in London for 2008–2017 is given in the table:

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Days of precipitation	19	16	21	21	13	21	30	26	21	15

(data from weatheronline.co.uk)

a Write down the range of the number of days of precipitation in January in London for these years.
b Calculate the interquartile range of the number of days of precipitation in January in London for these years.
c Find the standard deviation of the number of days of precipitation in January in London for these years.
d Find whether the 30 cm precipitation in January 2014 is an outlier.

a 17 days	The maximum value is 30 and the minimum value is 13. 30 − 13 = 17.
b 5 days	Upper half of the data is 21, 21, 21, 26, 30. The median of those is the middle term which is 21.

Continued on next page

c 4.80 days

Lower half of the data is 13, 15, 16, 19, 21.

The median of those is 16.

Interquartile range is $Q_3 - Q_1 = 21 - 16 = 5$

Or the answer can be obtained directly from the statistical summaries on the GDC.

This value should be obtained directly from the GDC.

d $Q_3 = 21$ and IQR = 5 so it is true that $30 > 21 + 1.5 \times 5$ so the data item for 2014 is, as expected, an outlier.

Investigation 3

1 In section 2.1, you collected five samples from the Mathematics results from Valley High School. Use each of the samples obtained to find an estimate of the mean and standard deviation of the Mathematics results all Robotics and Astronomy Club members in Valley High School over the period 2013–2017.

The mean and standard deviation of the population is in fact:

	Mean	Standard deviation
Robotics Club	75.3	12.5
Astronomy Club	78.5	9.74

Compare your answers with the actual values. Which seemed to be the best sampling method?

2 **Conceptual** Explain why you would have used technology for all those calculations.

Investigation 4

Complete the table for the following sets of numbers:

A: Find the mean and standard deviation of the numbers 3, 4, 6, 8, 9, 10, 15, 17.

B: Add 3 to each of the numbers in A and then find the mean and standard deviation.

C: Subtract 2 from each of the numbers in A and then find the mean and standard deviation.

D: Add 5 to each of the numbers in A and then find the mean and standard deviation.

E: Multiply the numbers in A by 3 and then find the mean and standard deviation.

F: Multiply the numbers in A by −2 and then find the mean and standard deviation.

G: Multiply the numbers in A by 0.5 and then find the mean and standard deviation.

	Mean	Standard deviation
A		
B		
C		
D		
E		
F		
G		

1 **Factual** What happens to the mean when you add or subtract a number from each term?

2 **Factual** What happens to the standard deviation when you add or subtract a number from each term?

3 **Factual** What happens to the mean when you multiply each number by a constant?

4 **Factual** What happens to the standard deviation when you multiply each number by a constant?

5 **Conceptual** Why does adding a constant to every value in a data set result in no change in the standard deviation?

6 The mean of a set of numbers is 10 and the standard deviation is 1.5.

a If you add 3 to each term, write down the new mean and standard deviation.

b If you multiply each term by 4, write down the new mean and standard deviation.

International-mindedness

The 19th century German psychologist Gustav Fechner popularized the median as a measure of central tendency although French mathematician Pierre Laplace had used it earlier.

The mean of a set of numbers is $\bar{x}$ and the standard deviation is σ_x.

If you add k to or subtract k from each of the numbers then the mean is $\bar{x} \pm k$ and the standard deviation is σ_x.

If you multiply each number by k then the mean is $k \times \bar{x}$ and the standard deviation is $|k| \times \sigma_x$.

Exercise 2B

1 Find the mean, median and mode for the following data sets and comment on any pieces of data that you think may be outliers.

a The times of 25 telephone calls in minutes:

1.0, 1.5, 2.3, 2.6, 2.8, 3.0, 3.4, 3.8, 4.1, 4.5, 4.6, 4.8, 5.2, 5.3, 5.5, 5.8, 6.0, 6.3, 6.6, 7.3, 7.5, 7.5, 7.5, 17.8, 25.0

b The heights, in metres, of 15 sunflowers:

1.1, 2.2, 2.5, 2.5, 2.5, 3.1, 3.5, 3.6, 3.9, 4.0, 4.1, 4.4, 4.6, 4.9, 6.1

c The results of a Geography test:

22, 39, 45, 46, 46, 52, 54, 58, 62, 62, 62, 67, 70, 75, 78, 82, 89, 91, 95, 98

2 In a survey, 25 people were asked how many times they visited the cinema in the last two weeks. The results are given in the frequency table:

Number of visits	0	1	2	3	5	8
Number of people	5	5	6	2	1	1

a Find the mean, median and modal number of visits.

b Find the standard deviation and the variance of the number of visits.

3 The table shows the number of orthodontist visits per year made by the students in Grade 10.

Number of visits	0	4	6	8	10	12	14
Frequency	3	2	8	4	2	12	5

a Find the mode, the median and the mean, and comment on which is the most appropriate to use.

b Find the standard deviation and comment on the result.

c Find the range and interquartile range, and comment on the spread of the data.

4 The heights in centimetres of 15 basketball players are:

175, 183, 191, 196, 198, 201, 203, 203, 204, 206, 207, 209, 211, 212, 213

The heights of 15 randomly chosen males are:

154, 158, 158, 162, 165, 168, 171, 176, 178, 180, 181, 182, 182, 183, 186

a Find the mean and standard deviation for each group.

b Compare your answers and comment on any similarities or differences.

5 Mr Jones, a teacher in Valley High School, collects the following data from his class and suggests that it provides evidence that Robotics Club members are better at Maths than Astronomy Club members.

Name	Club	Result	Name	Club	Result
Abdullah L	R	72	Justine H	R	83
Angela W	R	96	Kara F	A	70
Arthur B	R	84	Kay H	A	60
Brad S	A	61	Lia S	A	70
Dalia V	A	83	Marcus C	A	64
Eddie R	A	77	Marina B	R	89
Elsie D	A	78	Mattias L	R	52
Emma M	A	60	Natalya A	A	60
Ernesto H	R	65	Preston H	A	68
Hadassah H	R	79	Thalia C	A	69
Jenny E	A	83	Vance T	R	83
Joanna S	R	81	Waylon D	R	57
Jonathan S	A	69			

Calculate the mean, median, standard deviation and quartiles for the Robotics and Astronomy Clubs' Maths results in Mr Jones' class. State whether or not your calculations support Mr Jones' claim.

6 For the sample collected in Exercise 2A question 1, find the mean and standard deviation for the English results of each year. Comment on any changes in the distribution of English results over the five years that are indicated by these values.

7 The number of sweets in 25 bags has a mean of 30 and standard deviation of 3. In a special promotion, the manufacturer doubles the number of sweets in each bag. Write down the new mean and the new standard deviation of the number of sweets in a bag.

8 Mrs Ginger's Grade 8 class sat an English test. The test was out of 40 marks. The mean score was 32 marks and the standard deviation was 8 marks.

In order to change this to a mark out of 100, Mrs Ginger thinks that it would be all right to multiply all the scores by 2 and then add 20 to each one.

Mr Ginger thinks that it would be fairer to multiply all the scores by 2.5.

Miss Ginger suggests multiplying by 3 and subtracting 20 from each score.

a Write down the new mean and the new standard deviation for each suggestion.

Matty had an original score of 12, Zoe had an original score of 25 and Ans had an original score of 36.

b Find their new scores under all three suggested changes.

c Comment on how each of these three suggestions would affect students with low, middling and high marks out of 40.

Statistical measures of continuous data

Until now we have been looking at finding summary measures for discrete data.

For continuous data we can only consider estimations as we never know the exact value of any data item.

If data is continuous you find **estimates** for the mean, variance or standard deviation by assuming that all of the data values are equally spread around the midpoint.

You can find a **modal class** if the data are arranged in intervals of equal width.

TOK

Is standard deviation a mathematical discovery or a creation of the human mind?

Example 4

Heights of 200 fir trees are measured and the results recorded:

Height, h (m)	$0 < h \leq 1$	$1 < h \leq 2$	$2 < h \leq 3$	$3 < h \leq 4$	$4 < h \leq 6$	$6 < h \leq 10$
Frequency	17	35	69	51	22	6

a Find estimates for the mean and standard deviation of the height of these fir trees.

b State why your calculations are estimates.

a Estimate of mean = 2.85 Estimate of SD = 4.36 (3 s.f.)	First enter the data into your GDC using mid-points as the data points then find values as normal. Having entered the data all the summary values, for example the median or quartiles can be found directly from the GDC.
b They are only estimates because mid-points have been entered as estimates of what the data points might actually have been.	

Exercise 2C

1 For the following sets of data, find

i the modal class

ii an estimate for the mean

iii an estimate for the median.

Comment on the meaning of these values and state which one is most appropriate to use in each case, giving a reason for your answer.

a

Number of cars, n	Frequency
$0 \leq n < 30$	12
$30 \leq n < 60$	28
$60 \leq n < 90$	39
$90 \leq n < 120$	42
$120 \leq n < 150$	54
$150 \leq n < 180$	65

b

Speed of cars, s (mph)	Frequency
$40 \leq s < 45$	4
$45 \leq s < 50$	8
$50 \leq s < 55$	23
$55 \leq s < 60$	15
$60 \leq s < 65$	6
$65 \leq s < 70$	4

c

Time to complete a puzzle, t (minutes)	Frequency
$2 \leq t < 3$	2
$3 \leq t < 4$	5
$4 \leq t < 5$	3
$5 \leq t < 6$	7
$6 \leq t < 7$	4
$7 \leq t < 8$	9
$8 \leq t < 9$	3

2 Gal asked 60 people how much money they had spent the last time they had eaten in a restaurant. The table shows his results.

Cost of dinner, c (GBP)	Frequency
$10 \leq c < 20$	6
$20 \leq c < 30$	12
$30 \leq c < 40$	28
$40 \leq c < 50$	10
$50 \leq c < 60$	4

a Write down the modal class.

b Find an estimate for the mean and the median.

c Find an estimate for the standard deviation and comment on the result.

d Find an estimate for the variance, the range and the interquartile range and explain why these are all estimates.

3 The table shows the heights of 50 wallabies.

Height, x cm	Frequency
$150 \leq x < 160$	3
$160 \leq x < 170$	5
$170 \leq x < 180$	13
$180 \leq x < 190$	23
$190 \leq x < 200$	4
$200 \leq x < 210$	2

a Write down the modal class.

b Find an estimate for the mean and standard deviation; comment on your answer.

4 The table shows the monthly salaries of all the staff at Mount High College.

Monthly salary, \$$x$	Number of males	Number of females
$1000 < x \leq 1500$	4	9
$1500 < x \leq 2000$	8	14
$2000 < x \leq 2500$	14	11
$2500 < x \leq 3000$	16	10
$3000 < x \leq 3500$	7	3
$3500 < x \leq 4000$	2	1
$4000 < x \leq 4500$	3	0

a Find the mean and standard deviation for each group.

b Compare your answers and comment on any similarities or differences.

Developing inquiry skills

Take a sample from the data in the opening section and use it to calculate statistical summaries, to determine whether different regions have different life expectancies.

2.3 Ways in which you can present data

Frequency histograms

A **histogram** is very similar to a **bar chart**. However, in a histogram there are no spaces between the bars.

Bar charts are useful for graphing **qualitative** data such as colour preference, whereas histograms are used to graph **quantitative** data.

Frequency histograms, like bar charts, have the vertical axis representing frequency.

To draw a frequency histogram, you need to find the lower and upper boundaries of the classes and draw the bars between these boundaries.

International-mindedness

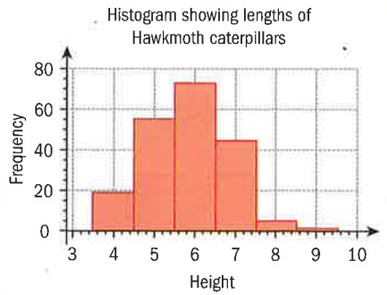
What are the benefits of sharing and analysing data from different countries?

Example 5

The data shown in the table was collected for Hawkmoth caterpillars, measured to the nearest cm.

Length, l (cm)	4	5	6	7	8	9
Frequency	19	56	74	45	5	1

Use the data to draw a frequency histogram.



As data goes from 3.5–9.5, our horizontal axes needs to include at least that. Choose 3–10.

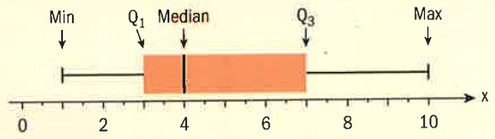
The maximum frequency is 74 so have the vertical axis up to say 80.

First bar has height of 19 and width from 3.5 to 4.5 etc.

Box-and-whisker diagrams

Box-and-whisker diagrams (often just called box diagrams) are another convenient way to present data to allow us to easily visualize characteristics of the data. They can be drawn for discrete or continuous data and are often very convenient for comparing sets of data.

To draw a box diagram, you first need to calculate the median, the quartiles, and the minimum and maximum values of the data. You then draw the box diagram as shown.



Example 6

The box-and-whisker diagrams below show the heights of all prominent peaks in different regions of central Asia.

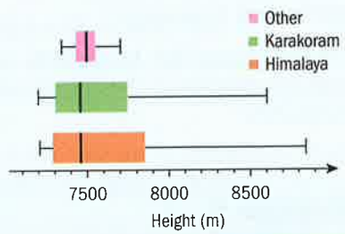
a State which region has the tallest mountain.

b State the region with the lowest standard deviation.

c Estimate the median height of peaks in the Karakoram.

d Estimate the interquartile range of heights of peaks in the Himalaya.

e Estimate the range of heights of peaks in the Karakoram.



a Himalaya	The maximum value is over 8800 which is higher than any other region (in fact Mount Everest which is 8848 m).
b Other	The data is less spread than the other regions.
c 7450 m	Somewhere between 7400 and 7500, about the same as for the Himalaya
d 550 m	About 7850 – 7300 = 550
e 1400 m	About 8600 – 7200 = 1400

The outliers are represented on a box-and-whisker diagram as separate crosses.

TOK

Can you justify using statistics to mislead others?

How easy is it to be misled by statistics?

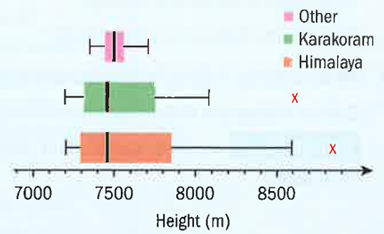
Example 7

The box-and-whisker diagrams from Example 6 are shown along with outliers.

a State in which region you would find the second highest mountain.

b Estimate the height of the second highest mountain in the Himalaya.

c Estimate the difference between the highest peak and the second highest peak in the Karakoram.



a The Karakoram	The outlier in the Karakoram is (slightly) higher than the maximum height in the Himalaya, apart from Everest (in fact it is K2).
b 8600 m	The maximum height inside the outlier is about 8600 m.
c 500 m	About 8600 – 8100 = 500 m

Investigation 5

Five investment companies, Altrucorp, Betterinvest, Cityshares, Dependshare and Eversafe each wish for you to invest your money through their company.

Eversafe promotes itself by saying that "over half its clients get a return of over 9.35% – more than 0.17% better than its nearest rival".

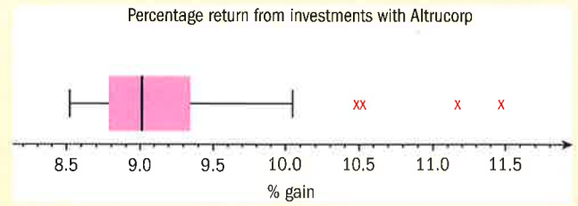
The summary data for the five companies is given here. Explain why Eversafe is justified in its claim.

Company	Number of investors	Mean % return	Median % return	Standard Deviation	Minimum % return	Maximum % return	Q_1	Q_3
A	106	9.12651	9.0063	0.522407	8.50703	11.5461	8.7799	9.35084
B	259	9.11148	9.17521	0.356837	7.46087	9.72693	8.91512	9.36484
C	978	9.09267	9.09045	0.195937	8.43358	9.63638	8.96163	9.22571
D	1222	9.09522	9.08718	0.491454	7.34769	11.0731	8.76	9.42048
E	312	9.08224	9.35156	0.794434	5.58167	9.59996	8.95071	9.53696

1 Draw a box-and-whisker diagram to illustrate the five companies. Explain how that would help to decide on which company you would use to invest your money.

2 The box-and-whisker diagram for Altrucorp is shown with outliers.

Why might this change your perception of the company?

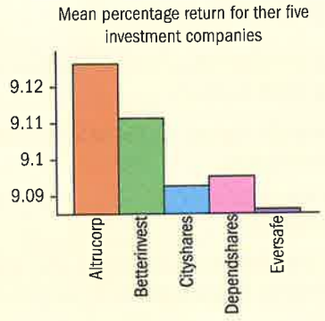


Continued on next page

3 Altrucorp uses this graph to promote itself.

Why might this be misleading?

Would it be dishonest advertising?



4 What statistic might Cityshares use to promote itself?

Draw a histogram to best promote the company.

Conceptual In what ways could you use graphics to mislead?

Exercise 2D

1 The number of days of precipitation in January in London for 2008–2017 is given in the table:

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Days of precipitation	19	16	21	21	13	21	30	26	21	15

(data from weatheronline.co.uk)

Draw a box-and-whisker diagram for the number of days of precipitation in London for the given years, marking the outlier.

2 The time, in minutes, to complete 200 games of chess is shown in the table.

Time, x minutes	Frequency
$20 \leq x < 30$	36
$30 \leq x < 40$	67
$40 \leq x < 50$	48
$50 \leq x < 60$	27
$60 \leq x < 70$	10
$70 \leq x < 80$	7
$80 \leq x < 90$	5

a Draw a histogram to represent this data.

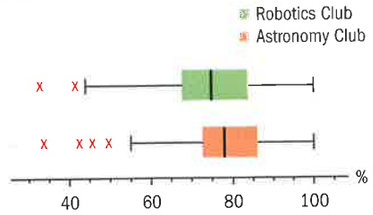
b Find the mean, median, Q_1, Q_3 and range, and determine if there are any outliers.

c Given that the quickest time was 26 minutes and the longest time was 84 minutes, draw a box-and-whisker plot to represent this data.

3 a Use the data from example 3 in section 2.2 to draw a histogram for the number of days of sunshine in Helsinki. You should regard the discrete data as continuous to draw a histogram (ie treat 1 day as an interval from 0.5 to 1.5).

b Comment on the symmetry of the data.

4 For each of the five samples you collected from the Mathematics results from Valley High School in section 2.1, draw separate box-and-whisker diagrams for the Robotics Club and Astronomy Club, clearly showing any outliers. The box-and-whisker diagram for the entire student population is shown.



State which of your samples best represents the school population over the period 2013–2017.

Cumulative frequency

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

- If data is continuous we find **estimates** for the median or interquartile range from a cumulative frequency curve or cumulative frequency polygon.
- To find any **percentile**, $p\%$, you read the value on the curve corresponding to $p\%$ of the total frequency.

Investigation 6

1 A sample of 200 Hawkmoth caterpillars were measured and their lengths to the nearest cm are given in the table:

Length, l (cm)	4	5	6	7	8	9
Frequency	19	56	74	45	5	1

Explain why 75 of the caterpillars are less than 5.5 cm.

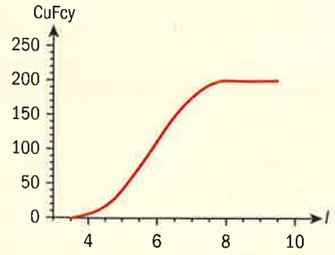
2 Use your answer from above to complete the cumulative frequency table:

Length, l (cm)	$l < 3.5$	$l < 4.5$	$l < 5.5$	$l < 6.5$	$l < 7.5$	$l < 8.5$	$l < 9.5$
Cumulative frequency			75				

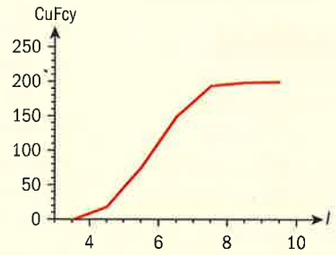
Factual How are all the values for the cumulative frequency calculated?

The points can be plotted and joined with a curve to create a cumulative frequency curve or joined with straight lines to create a cumulative frequency polygon. Unfortunately, that is not normally a task that can be completed on the GDC.

Cumulative frequency curve of caterpillar lengths



Cumulative frequency polygon of caterpillar lengths



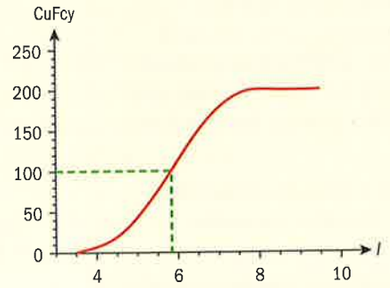
The curve shows how many caterpillars are less than a given length. Since half of the caterpillars are less than the median length, we can read off from the graph that 100 caterpillars are less than 5.8 cm. So the median is 5.8cm.

Continued on next page

EXAM HINT

If you are asked to draw a cumulative frequency curve, do **not** draw a cumulative frequency polygon.

Cumulative frequency curve of caterpillar lengths



When using the GDC, there were 74 data points entered as 6 cm. The actual lengths of all these caterpillars was somewhere between 5.5 cm and 6.5 cm. When calculating the statistical summaries on the GDC, it gives the median as 6, whereas it appears as 5.8 from the graph.

3 Why should this be different?

25% of caterpillars will be less than the value of Q_1. Use the graph to read off the approximate value of Q_1.

4 **Factual** How would you use the cumulative frequency graph to find an estimate for the upper quartile of the data?

Similarly use the graph to find the value of Q_3 and of the interquartile range.

5 How could you use the graph to find how many caterpillars were more than 7 cm?

6 **Conceptual** What is the purpose of using cumulative frequency graphs?

7 Use the cumulative frequency curve to find the 95th and the 90th percentiles.

EXAM HINT

Always show lines on the graph to show how you have found the required value.

TOK

Why have mathematics and statistics sometimes been treated as separate subjects?

Exercise 2E

1 The table shows the average times, in minutes, that 100 people waited for a train.

Time, x minutes	Frequency
$0 \le x < 2$	5
$2 \le x < 4$	11
$4 \le x < 6$	23
$6 \le x < 8$	31
$8 \le x < 10$	19
$10 \le x < 12$	8
$12 \le x < 14$	3

a Draw a cumulative frequency table for this data.

b Sketch the cumulative frequency curve.

c Use your graph to find an estimate for the median and interquartile range.

d Find the 10th percentile.

The train company will refund the fare if their customers have to wait 11 minutes or more for a train.

e Determine the number of customers who can claim for a refund of their fare.

2 Nuria recorded the number of words in a sentence in one chapter of her favourite book. The results are shown in the table.

Number of words, x	Frequency
$0 \le x < 4$	5
$4 \le x < 8$	32
$8 \le x < 12$	41
$12 \le x < 16$	28
$16 \le x < 20$	22
$20 \le x < 24$	12
$24 \le x < 28$	7
$28 \le x < 32$	3

a Construct a cumulative frequency table for this data.

b Sketch the cumulative frequency curve.

c Use your graph to find an estimate for the median and interquartile range.

d Determine if there are any outliers.

e Find the 90th percentile.

f The smallest sentence had 1 word and the longest sentence had 31 words. Draw a box-and-whisker plot to represent this information.

g A children's book has, on average, 8 words in a sentence and an adult book has, on average, 15 words in a sentence. State the type of book you think Nuria is reading. Explain your answer.

3 In example 6, the heights of 200 fir trees were given as:

Height, h (m)	$0 < h \le 1$	$1 < h \le 2$	$2 < h \le 3$	$3 < h \le 4$	$4 < h \le 6$	$6 < h \le 10$
Frequency	17	35	69	51	22	6

a Construct a cumulative frequency table for the heights of fir trees.

b Draw a cumulative frequency curve for the heights of fir trees.

c Estimate the median height of the 200 fir trees.

d Estimate the interquartile range of the heights.

e Estimate the 10th percentile of the heights.

f If the tallest 12 trees are to be felled, estimate the height of the smallest tree felled.

4 Use the data in Section 2.3, investigation 6, to draw a box-and-whisker diagram to illustrate the lengths of the 200 Hawkmoth caterpillars.

5 A tourist attraction is open 350 days in the year. The number of visitors each day for the 350 days was recorded and the results are shown in the table.

Number of visitors, n	Frequency
$100 \le n < 200$	24
$200 \le n < 300$	36
$300 \le n < 400$	68
$400 \le n < 500$	95
$500 \le n < 600$	73
$600 \le n < 700$	38
$700 \le n < 800$	16

a Draw a suitable graph to represent this data.

b Use your graph or the data to find an estimate for the median and interquartile range.

c State whether or not there are any outliers.

d The smallest number of visitors was 185 and the largest number was 792. Draw a box-and-whisker plot to represent this data.

If the number of tourists is less than 350, then the attraction loses revenue.

e Determine the number of days that the attraction loses revenue.

Developing inquiry skills

Would you have a better understanding of a set of data after looking at either:

a the raw data

b summary statistics

c statistical charts

d a combination of two or more of the above.

At the end of section 2.2, you collected a sample from the data in the opening section. Use that data to draw statistical charts to illustrate your findings.

International-mindedness

Hans Rosling (1948–2017) was a professor of international health at Sweden's Karolinska Institute. He co-founded the Swedish chapter of Medécins Sans Frontières, and was able to clearly show the importance of collecting and understanding real data in order to understand situations and plan for the future.

2.4 Bivariate data

- Are students who are good at Mathematics also good at Physics?
- Is smoking linked to lung cancer?
- Is the re-election of a government influenced by the state of the economy?
- Is a good breakfast essential to success in school?

In the previous sections we have been looking at ways to analyse a set of data in one variable. Frequently it is necessary to look at ways in which one variable interacts with another.

Bivariate data has **two** variables; **univariate** data has only **one** variable.

With bivariate data you have **two sets of data** that you want to compare to see if there is any **correlation** between the two sets.

Mr Price was interested to find out if the number of past papers that his students completed had an effect on the grade they obtained in their final examination. The data he collected is shown below.

Number of past papers	2	6	5	1	4	8	3	12	7	4	2	8	10	9
Examination grade (%)	48	70	61	45	58	85	55	96	80	56	43	88	92	89

He plots all these points on a graph to see if there is any correlation between the two sets of data. The number of past papers is the **independent** variable and this is plotted on the x-axis. The examination grade is the **dependent** variable and this is plotted on the y-axis.

The pattern of dots or crosses will give him an indication of how closely the variables are related.

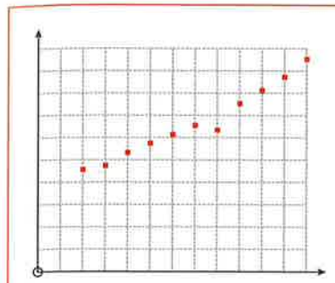
Do you think that the two sets of data are related?

How closely do you think they are related?

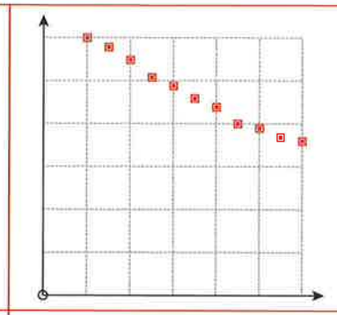
What advice would you give to students who have to take examinations?

Types of correlation

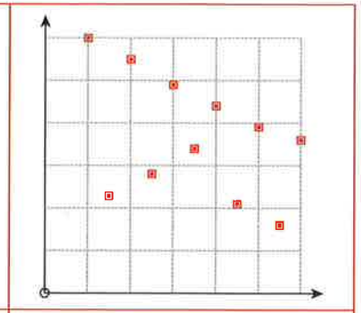
If you have date items $x_1, x_2, \ldots, x_n$ with associated data items $y_1, y_2, \ldots, y_n$ then we can draw a **scatter diagram** by plotting the data pairs (x_1, y_1), (x_2, y_2), …, (x_n, y_n).



Correlation can be **positive**. When the independent variable increases, so does the dependent variable.

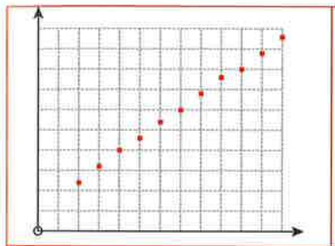


Correlation can be **negative**. When the independent variable increases, the dependent variable decreases.

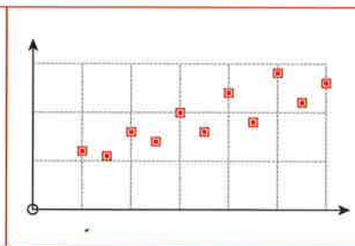


There can also be **no correlation**. This occurs when the points are scattered randomly.

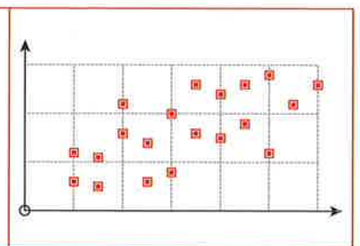
Correlation can also be described as strong, moderate or weak.



Strong



Moderate



Weak

Correlation does not imply **causation**. Two quantities may have very strong correlation but that may be due to an underlying cause in common, or simply coincidence.

TOK

To what extent can we rely on technology to produce our results?

Example 8

1 The table shows the number of schools and the number of restaurants in a town over a 40-year period.

Year	**1980**	**1985**	**1990**	**1995**	**2000**	**2005**	**2010**	**2015**	**2020**
Number of schools	12	13	15	15	16	17	17	18	19
Number of restaurants	28	30	33	34	36	36	38	39	40

Continued on next page

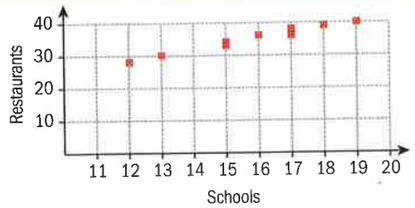
a Draw a scatter graph of the number of schools and the number of restaurants.

b Describe the correlation between the two sets of data.

c State whether or not you think that one set of data "causes" the other set.

d State another reason why the number of schools and the number of restaurants increased over the 40-year period.

1 a



b There is a strong, positive correlation.

c Not directly.

d The population in the town could be increasing every year, which could require more schools and more restaurants.

- A **line of best fit** can be drawn on a scatter diagram, by plotting the point $(\bar{x}, \bar{y})$ and drawing a line through that point that best follows the trend of the data.
- If the gradient of that line is positive then we say that the data has **positive correlation**.
- If the gradient of the line is negative then it has **negative correlation**.
- The strength of correlation is determined by how close the data points are from the line.

Example 9

The results of ten students in their final Mathematics and Physics exams are given.

Student	1	2	3	4	5	6	7	8	9	10
Mathematics result (%)	78	56	88	93	44	76	33	59	82	99
Physics result (%)	84	62	84	100	51	90	42	74	80	89

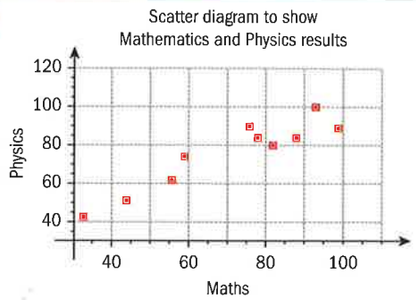
a Plot the information on a scatter diagram.

b Plot the point $(\bar{x}, \bar{y})$ and draw the line of best fit.

c Predict the Physics result for a student who scored 65% on their Mathematics exam.

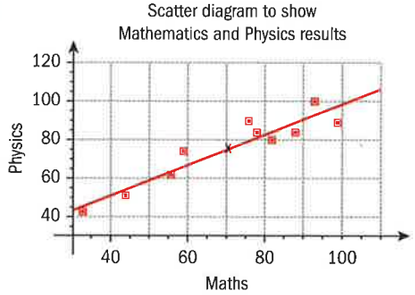
d State whether or not the results indicate that students who are good at Mathematics are also good at Physics.

a



Plot the points (78, 84), (56, 62) etc on a set of axes.

b



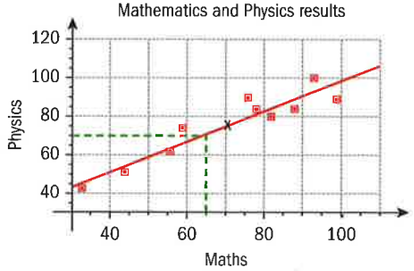
The mean of $x_1, x_2, \ldots, x_{10}$ is $\bar{x} = 70.8$.

The mean of $y_1, y_2, \ldots, y_{10}$ is $\bar{y} = 75.6$.

Mark (70.8, 75.6).

Draw line so that is best matches.

c



Use the line of best fit to predict the score.

Using the line of best fit to predict a data value within the range of the given data is called interpolation.

The predicted score is 70%

d There is some indication that high results in Mathematics correspond to high results in Physics

The line of best fit has a positive gradient and the data are closely clustered around the line of best fit.

The example above shows what appears to be quite a strong correlation.

We have a measure of the strength of linear correlation, called Pearson's product-moment correlation coefficient (PMCC), which is denoted by r.

PMCC can be calculated using technology and takes values between -1 and 1. $r = 1$ indicates perfect positive correlation, whereas $r = -1$ indicates perfect negative correlation. $r = 0$ indicates no correlation

TOK

Is there a difference between information and data?

Make sure you know how to calculate the value of r using your technology.

Example 10

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
x	227	456	509	497	596	573	661	741	809	717
y	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8

a For the data shown, calculate r, Pearson's product-moment correlation coefficient.

b Comment on the strength of correlation.

a 0.918	$r = 0.918$
b The data shows strong positive correlation.	r is close to 1 so there is strong positive correlation

In example 10, you found a strong positive correlation. This might be surprising if you consider that the data items for x were the number of people who died by becoming entangled in their bedsheets in the US and the data values for y were the amount of cheese per capita consumed in the US (in lb). (Data from http://www.tylervigen.com/spurious-correlations.)

Investigation 7

A drug is developed for treating a skin condition. A trial was undertaken to discover the effects of different daily dosages. The table gives effects on the decrease in the area of skin affected and the amount of fever the patient is showing.

Dosage (mg)	4	5	6	7	8	9	10	11	12
Percentage decrease in area of skin affected per year	5.9	8.2	12.7	20.6	18.7	8.4	24.9	34.2	35
Patient temperature (°C)	39.3	39.2	38.8	38.8	38.4	38.4	38	38.3	37.7

1 With a horizontal axis from 0 to 20 and a vertical axis from 0 to 40, plot the percentage decrease against the dosage.

2 Does your scatter diagram indicate that a higher dosage of the drug is likely to produce a greater percentage decrease in the skin affected?

3 Explain why we would consider dosage to be the independent variable and the percentage decrease in infected skin to be the dependent variable.

4 A researcher believes that one of the data points was recorded incorrectly during the trial.

- Which data point could be classified as an outlier?
- If the data was recorded correctly, give another explanation why the outlier does not fit the pattern of the other results.

5 Exclude the outlier to calculate the PMCC. Comment on the correlation.

6 Draw another scatter diagram with a horizontal axis from 0 to 20 and a vertical axis from 37 to 40. Plot the patient temperature against the dosage.

7 Continue to exclude the outlier to calculate the PMCC in this case. Comment on the correlation.

8 Do your results indicate that the dosage given influences the patient temperature?

9 Again in each case, excluding the outlier, calculate the coordinates of $(\overline{x}, \overline{y})$ and draw the line of best fit onto your graph.

10 The trial continued with increased dosages and the results are given.

Dosage (mg)	13	14	15	16	17	18	19
Percentage decrease in area of skin affected per year	36.3	35.2	36.6	33.1	36.6	25.8	21
Patient temperature (°C)	38.3	37.2	38	38.3	37.6	37.3	37.8

Plot the extra data points onto your scatter diagrams.

11 Do the data points maintain the same trend?

Give possible reasons why.

12 If you were to calculate the value of r for the whole data set explain what you would expect to find.

13 **Conceptual** Can we use extrapolation to predict beyond the data points? Why or why not?

Extrapolation means estimating a value at a point that is larger than (or smaller than) the data you have.

Trends in data are only valid for the range of study. We cannot **extrapolate** to draw conclusions outside of that range.

Exercise 2F

1 The table shows the size, in inches, of 10 laptop screens and the cost, in euros, of the laptop.

Size, inches	11.6	11.6	13.3	14	14	14	15	15.6	15.6	15.6
Cost, euros	145	170	700	450	370	175	320	500	420	615

a Plot the points on a scatter diagram.

b Describe and interpret the correlation.

c Comment on whether the size has an influence on the cost.

2 The table gives the heights, in cm, and weights, in kg, of 11 football players selected at random.

Height, h cm	161	173	154	181	172	184	176	169	165	180	173
Weight, w kg	74	76	61	80	76	88	79	76	75	83	75

a Plot the points on a scatter diagram.

b Calculate the coordinates of the point $(\overline{h}, \overline{w})$ and mark it on your graph.

c Draw a line of best fit.

d Predict the weight of a football player with height 170 cm.

e Calculate Pearson's Product-Moment Correlation Coefficient for the data.

f Describe the correlation. Interpret what this means in terms of the football players.

g Comment on whether the correlation might indicate a causation in this instance. Justify your answer.

3 A sample of 15 people were taken and given a vocabulary test. Their test results, v (%), were compared against their heights, h (cm), and the results given in the table:

h	1.18	1.26	1.32	1.50	1.63	1.70	1.69	1.45	1.56	1.23	1.44	1.53	1.60	1.38	1.30
v	51	64	59	67	80	82	77	75	67	54	66	81	75	69	54

a Draw a scatter diagram to show the data, with height as the independent variable.

b Calculate the coordinates of the point $(\overline{h}, \overline{v})$ and mark it on your graph.

c Draw a line of best fit.

d Klaus is 1.92 m tall, predict his vocabulary test result and comment on your prediction.

e Calculate Pearson's Product-Moment Correlation Coefficient for the data.

f Comment on the correlation of the data.

g Comment on whether this shows that taller people have better vocabularies.

4 Prices of unleaded fuel and diesel (in euros) in December 2017 are recorded across 28 EU countries.

Unleaded	1.168	1.403	1.064	1.259	1.203	1.156	1.597	1.229	1.416	1.421	1.369	1.558	1.151	1.389
Diesel	1.068	1.392	1.069	1.207	1.215	1.094	1.395	1.229	1.3	1.303	1.209	1.359	1.18	1.279
Unleaded	1.581	1.117	1.143	1.18	1.31	1.678	1.065	1.544	1.07	1.172	1.26	1.236	1.459	1.361
Diesel	1.444	1.037	1.039	1.043	1.18	1.37	1.041	1.354	1.083	1.035	1.217	1.157	1.467	1.394

http://www.fuel-prices-europe.info/

a Draw a scatter plot of the data, with the cost of unleaded fuel as the independent variable.

b Calculate the value of r and comment on the correlation between the cost of unleaded fuel and the cost of diesel.

Developing inquiry skills

For the country data given in the opening of this chapter, take a suitable sample to determine whether GDP and income have any correlation.

Chapter summary

- **Qualitative data** is non-numerical, eg "it was fun", "blue".
- **Quantitative data** is numerical. Quantitative data can be **discrete** or **continuous**.
- **Discrete data** is data which takes specific (discrete) values, eg "number of accidents", "points in the IB diploma".
- **Continuous data** is data which can take a full range of values, eg "height", "speed".
- A **population** includes all members of a defined group.
- A **sample** is a subset of the population, a selection of individuals from the population.
- **Biased sampling:** The sampling method is not random so not all members of the population are equally likely to be selected. Biased sampling may cause you to draw misleading conclusions about the population.

- **Simple random sampling:** every member of the population is equally likely to be chosen. For example, allocate each member of the population a number. Then use random numbers to choose a sample.
- **Systematic sampling:** find a sample of size n from a population of size N by selecting every kth member where $k = \frac{N}{n}$ rounded to the nearest whole number.
- **Stratified sampling:** is selecting a random sample where numbers in certain categories are proportional to the numbers in the population.
- **Quota sampling:** decide how many members of each group you want to sample and take samples from the population until you have a large enough sample for each group.
- **Convenience sampling:** take samples from the members of the population that you have access to until you have a sample of the desired size.
- The most common measures of central tendency are the mean, median and mode.
- The **mode** of a data set is the value that occurs most frequently. There can be no mode, one mode, or several modes.
- The **median** of a data set is the value that lies in the middle when the data are arranged in size. When there are two middle values then the median is the midpoint between the two values.
- The **mean** of a data set is the sum of all the values divided by the number of values. For a discrete data set of n values the formula is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, where $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$. Σ means "the sum of".
- When there is a frequency table, you need to use the data values and the corresponding frequencies to calculate the mean.
- Measures of dispersion measure how spread out a data set is.
- The most common measure of dispersion is the **range**, which is found by subtracting the smallest number from the largest number.
- The standard deviation, σ_n, gives an idea of how the data values are related to the mean. The greater the standard deviation, the more spread out the data.
- In examinations you will use your GDC to find the standard deviation.
- The **variance** is the standard deviation squared: $(\sigma_n)^2$.
- The **interquartile range** (IQR) is the **upper quartile**, Q_3, minus the **lower quartile**, Q_1.
- When the data are arranged in order, the lower quartile is the data point at the 25th percentile and the upper quartile is the data point at the 75th percentile.
- An **outlier** is defined as a data item that is more than 1.5 × IQR below Q_1 or above Q_3.
- Outliers are extreme data values, or the result of errors in reading data, that can distort the results of statistical processes.
- Outliers can affect the mean by making it larger or smaller, but most likely will not affect the median or the mode.
- Outliers can affect the standard deviation by making it larger or smaller, but they most likely will not affect the interquartile range.
- Given the mean of a set of numbers is $\overline{x}$ and the standard deviation is σ_x.
  If you add k to or subtract k from each of the numbers then the mean is $\overline{x} \pm k$ and the standard deviation is σ_x.
  If you multiply each number by k then the mean is $k \times \overline{x}$ and the standard deviation is $|k| \times \sigma_x$.

Continued on next page

- If data is continuous we find **estimates** for the mean, variance or standard deviation by assuming that all of the data values are equally spread around the midpoint.
- We can find a **modal class** if the data are arranged in intervals of equal width.
- Frequency histograms, like bar charts, have the vertical axis representing frequency.
- To draw a frequency histogram, you need to find the lower and upper boundaries of the classes and draw the bars between these boundaries.
- To draw a box-and-whisker plot you need five pieces of information: the smallest value, the lower quartile, the median, the upper quartile and the largest value.
- The outliers are represented on the box-and-whisker diagram as separate crosses.
- The **cumulative frequency** is the sum of all the frequencies up to and including the new value. To draw a cumulative frequency curve, you need to construct a cumulative frequency table, with the upper boundary of each class interval in one column and the corresponding cumulative frequency in another. Then plot the upper class boundary on the x-axis and the cumulative frequency on the y-axis.
- If data is continuous we find **estimates** for the median or interquartile range from a cumulative frequency curve or cumulative frequency polygon.
- To find any **percentile**, $p\%$, you read the value on the curve corresponding to $p\%$ of the total frequency.
- **Bivariate** data has **two** variables; **univariate** data has only **one** variable.
- With bivariate data you have **two sets of data** that you want to compare to see if there is any **correlation** between the two sets.
- A **line of best fit** can be drawn on a scatter diagram, by plotting the point $(\bar{x}, \bar{y})$ and drawing a line through that point that best follows the trend of the data.
- If the gradient of that line is positive then we say that the data has **positive correlation**.
- If the gradient of the line is negative then it has **negative correlation**.
- The strength of correlation is determined by how close the data points are from the line.
- PMCC can be calculated using the GDC and takes values between -1 and 1. $r = 1$ indicates perfect positive correlation, whereas $r = -1$ indicates perfect negative correlation. $r = 0$ indicates no correlation.
- Extrapolation means estimating a value at a point that is larger than (or smaller than) the data you have.
- Trends in data are only valid for the range of study. We cannot **extrapolate** to draw conclusions outside of that range.

Developing inquiry skills

Thinking about the opening problem:

- Has what you have learned in this chapter helped you to answer the questions?
- What information did you manage to find?
- What assumptions did you make?
- How will you be able to construct a model?
- What other things did you wonder about?

Thinking about the inquiry questions from the beginning of this chapter:

- Has what you have learned in this chapter helped you to think about an answer to most of these questions?
- Are there any that you are interested in and would like to explore further, perhaps for your internal assessment topic?

Chapter review

Click here for a mixed review exercise

1 The times, in minutes, it takes for 60 males and 40 females to swim 500 metres are:

Males:

16	14	17	8	12	11	13	15	12
10	10	9	13	16	15	8	9	10
11	10	9	13	18	15	15	13	14
14	10	20	9	7	20	18	13	14
15	15	12	11	15	10	10	11	11
17	18	19	21	9	18	16	15	15
18	12	12	13	15	10			

Females:

10	9	9	16	18	22	10	21	14
15	18	19	19	15	15	12	10	22
24	16	18	19	21	10	17	18	14
12	12	11	10	10	16	18	19	21
22	19	15	18					

a Find the mean and standard deviation for the males and the females and compare them.

b Find the mean and standard deviation for the 100 swimmers.

c Using a random sampling method to select 40 swimmers from the 100 swimmers and find the mean and standard deviation of the sample.

d Using a systematic sampling method, find the mean and standard deviation of 40 swimmers.

e Using a stratified sample, find the mean and standard deviation of 40 swimmers.

f In each case, compare your answers to the mean and standard deviation of the population.

2 Find the mean, median and mode for the following data sets. State which measure of central tendency is best to use in each case.

a The heights of 15 dogs, in cm:

7	23	32	41	32	56	64	67
88	91	110	78	56	45	32	

b The price of a pair of shoes in dollars:

46	54	58	62	62	79
96	120	135	185	270	300

c The hours Grade 12 students sleep:

4	7	6	6	8	6	9	8	6	5
4	5	5	6	8	8	8	6	7	

3 The data in the table show the lengths of 120 pike fish.

Length of pike, l cm	Frequency
$20 \le l < 30$	2
$30 \le l < 40$	12
$40 \le l < 50$	23
$50 \le l < 60$	46
$60 \le l < 70$	28
$70 \le l < 80$	9

a Write down the modal class.

b Find estimates for the median, mean and standard deviation.

c Draw a histogram to represent the data.

4 A company records its profits for the years 2000 to 2005. The results (to the nearest \$500) are shown in the table.

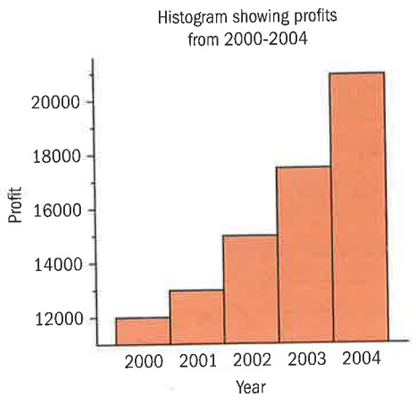
Year	2000	2001	2002	2003	2004
Profit (\$)	12 000	13 000	15 000	17 500	21 000

a Calculate the mean profit for the five years.

b Calculate the standard deviation of the profits over these five years.

c Calculate the percentage increase in profits from 2000 to 2001.

d The company illustrates its profits in its brochure using the bar chart shown:



i Explain why this diagram may give a misleading picture.

ii State reasons why the bar chart might be drawn in this way.

5 Ursula measures the heights of 35 tulips in her garden. The data she gathered is:

20	20	21	22	22	22	24
25	27	28	28	29	30	31
32	33	33	34	34	34	35
35	36	37	39	39	39	40
41	41	42	43	43	44	45

a Find the mean and standard deviation and comment on your answer.

b Find the range and interquartile range.

c Find the median and check whether there are any outliers.

d Draw a box-and-whisker plot to represent the data.

6 The number of push-ups that the Grade 11 students can perform is:

Girls:

2	4	5	5	7	8	8
10	10	13	15	15	15	18
20	21	22	23	24	25	25
26	27	28	30			

Boys:

5	8	10	12	12	12	15
18	18	20	21	22	22	25
25	28	30	31	31	35	35
38	45	46	48			

a Find the mean, median, Q_1, Q_3 and range for the girls and for the boys, and check whether there are any outliers.

b Draw box-and-whisker plots to represent the data.

c Compare the two plots.

7 The grouped frequency table shows the number of hours of voluntary service completed by the 200 students at a community high school.

Number of hours, x	Frequency
$0 \le x < 10$	8
$10 \le x < 20$	16
$20 \le x < 30$	41
$30 \le x < 40$	54
$40 \le x < 50$	36
$50 \le x < 60$	22
$60 \le x < 70$	17
$70 \le x < 80$	6

a Construct a cumulative frequency table for this data.

b Plot the points and draw the cumulative frequency curve.

c Use your curve to find an approximate value for the median and the interquartile range.

The lowest number of hours completed was 8 and the greatest number was 76.

d Draw a box-and-whisker plot to represent the data.

8 Waiting times in a busy post office are recorded over the course of a day. The times are recorded in the frequency table:

Time (m)	3.5–4	4–4.5	4.5–5	5–5.5	5.5–6	6–6.5	6.5–7	7–7.5	7.5–8	8–8.5	8.5–9
Frequency	6	14	48	89	121	129	103	70	30	10	2

a Calculate estimates of the mean and standard deviation of the waiting time.

b Construct a cumulative frequency table for the data, and use it to draw a cumulative frequency curve.

c Use your graph to estimate:

i the median waiting time

ii the lower and upper quartile of the waiting times

iii the interquartile range

iv the 85th percentile of waiting times.

d Draw a box-and-whisker plot of the data.

e Determine, with reasons, whether any customers could be considered outliers.

9 Mr Farmer has 50 chickens. He collects data on the temperature and the average number of eggs that the chickens lay.

Temperature, °C	Number of eggs
14	43
15	44
16	48
17	46
18	50
19	48
20	50
21	52
22	53
23	55

a Draw a scatter graph to represent this information.

b Describe the correlation.

c State whether you think the temperature has an effect on the number of eggs laid. Give a reason for your answer.

10 Ten people were included in a survey to measure reaction time against age, and the data is shown in the table.

Age (years)	13	18	22	26	28	42	55	66	78	84
Reaction time (s)	0.45	0.44	0.54	0.55	0.61	1.02	0.77	0.93	0.88	1.11

a Draw a scatter diagram to illustrate the data.

b Identify an outlier in the data set.

c Remove the outlier and calculate the mean point for the rest of the data.

d Draw the line of best fit onto your scatter diagram.

e Calculate r, the PMCC for the data (excluding the outlier).

f Comment on the correlation of reaction time according to age.

Exam-style questions

11 **P1:** Eight primary school children were given a spelling test which was marked out of 20.

Their results were: 15, 20, 18, 4, 12, 17, 12, 9

Find:

a the range of the data (1 mark)

b the mean mark (2 marks)

c the median mark (1 mark)

d the modal mark (1 mark)

e the variance of the data. (2 marks)

12 **P2:** The following tables show the mean daily temperatures, by month, in both Tenerife and Malta.

Tenerife

Month	Mean daily temperature (°C)
January	19
February	20
March	21
April	21
May	23
June	25
July	28
August	29
September	28
October	26
November	23
December	20

Malta

Month	Mean daily temperature (°C)
January	16
February	16
March	17
April	20
May	24
June	28
July	31
August	31
September	28
October	25
November	21
December	17

a Find the mean temperature over the course of the year for Tenerife. (2 marks)

b Find the standard deviation of temperatures in Tenerife. (2 marks)

c Find the mean temperature over the course of the year for Malta. (2 marks)

d Find the standard deviation of temperatures in Malta. (2 marks)

e By referring directly to your answers from parts **a–d**), make contextual comparisons about the temperatures in Tenerife and Malta throughout the year. (4 marks)

13 **P2:** Ben practises playing the Oboe daily. The time (in minutes) he spends on daily practice over 28 days is as follows.

10, 15, 30, 35, 40, 40, 45, 55, 60, 62, 64, 64, 66, 68, 70, 70, 72, 75, 75, 80, 82, 84, 90, 90, 105, 110, 120, 180

a Find the median time. (2 marks)

b Find the lower quartile. (2 marks)

c Find the upper quartile. (2 marks)

d Find the range. (2 marks)

e Determine whether there are any outliers in the data. (4 marks)

f Draw a box-and-whisker diagram for the above data, marking any outliers as required. (3 marks)

14 **P2:** The following raw data is a list of the height of flowers (in cm) in Eve's garden.

26.5, 53.2, 27.5, 33.6, 44.6, 39.5, 24.9, 45.1, 47.8, 39.3, 33.1, 38.7, 44.1, 22.3, 44.1, 30.5, 25.5, 35.9, 37.1, 40.2, 23.3, 36.2, 34.8, 37.3

a Copy and complete the following grouped frequency table.

Height, (x cm)	Frequency
$20 \le x < 25$	
$25 \le x < 30$	
$30 \le x < 35$	
$35 \le x < 40$	
$40 \le x < 45$	
$45 \le x < 50$	
$50 \le x < 55$	

(3 marks)

b Find an estimate for the mean height, using the frequency table. (2 marks)

c Find an estimate for the variance, using the frequency table. (2 marks)

d Find an estimate for the standard deviation, using the frequency table. (2 marks)

e Eve's neighbour's garden was also surveyed. It was found that the flowers in the neighbour's garden had a mean height of 32.1 cm and standard deviation 7.83 cm. Compare the heights of the flowers in both gardens, drawing specific conclusions. (3 marks)

15 **P1:** Icicles creamery decided to analyse their ice cream sales to see if there was any correlation between sales and the average outdoor temperature for that particular month.

The following data was collected:

Month	Mean temperature (T°C)	Sales ($)
January	3	350
February	4	650
March	9	900
April	11	920
May	17	1080
June	22	1200
July	25	1260
August	29	1390
September	19	1220
October	11	880
November	8	770
December	6	500

a Plot the given points on a scatter diagram. (2 marks)

b Calculate the coordinates of the point $(\overline{T}, \overline{P})$ and hence draw a line of best fit. (3 marks)

c Calculate Pearson's Product Moment Correlation Coefficient for this data, and interpret your result. (2 marks)

d Comment on whether you can conclude, from this data, that outdoor temperature affects ice cream sales. (2 marks)

16 **P1:** An analysis was undertaken of the weight of new cars sold during one particular month. To aid in the calculation, 5000 kg was subtracted from every data value, and each result divided by 200. The mean of these new values was 9.6 and the standard deviation 2.15.

a Find the actual mean weight of the new cars sold. (2 marks)

b Find the standard deviation of the new cars sold. (2 marks)

17 **P1:** An analysis was undertaken in Sydney to determine if there was any correlation between an employee's salary ($\$s$ Australian dollars, AUD) and the distance they lived from the centre of Sydney (d km). The data from ten employees was collected, and the results were as follows.

Salary ($s AUD)	Distance from centre ($d km)
155000	0.3
92000	3.0
66000	2.5
72000	4.8
116000	1.2
153000	1.9
48000	4.0
118000	2.2
106000	4.1
140000	0.9

a Calculate Pearson's product-moment correction coefficient for this data, and interpret your result. (3 marks)

b Plot a scatter diagram to represent this data. Calculate values for $\overline{s}$ and $\overline{d}$, and hence draw a line of best fit on your scatter diagram. (5 marks)

c It is suggested that the scatter diagram could be used to determine the average salary of an employee living 7 km from the city centre.

Suggest two reasons why this may not be accurate. (3 marks)

Modelling and investigation activity

What's the difference?

Approaches to learning: Thinking skills, Communicating, Collaborating, Research

Exploration criteria: Presentation (A), Mathematical communication (B), Personal engagement (C), Reflection (D), Use of mathematics (E)

IB topic: Statistics, Mean, Median, Mode, Range, Standard deviation, Box plots, Histograms

Example experiment

Raghu does an experiment with a group of 25 students.

Each member of the group does a reaction test and Raghu records their times.

Raghu wants to repeat the experiment, but with some change.

He then wants to compare the reaction times in the two experiments.

Discuss:

How could Raghu change his experiment when he does it again?

With each change, is the performance in the group likely to improve/stay the same/get worse?

Alternatively, Raghu could use a different group when he repeats the experiment.

What different group could he use?

With each different group, is the performance likely to improve/stay the same/get worse?

Modelling and investigation activity

Your experiment

Your task is to devise an experiment to test your own hypothesis.

You will need to do your experiment two times and compare your results.

Step 1: What are you going to test? State your aim and hypothesis

Write down the aim of your experiment and your hypothesis about the result.

Why do you think this is important?

What are the implications of the results that you may find?

Make sure it is clear what you are testing for.

Step 2: How are you going to collect the data? Write a plan

- What resources/sites will you need to use?
- How many people/students will you be able to/need to collect data from to give statistically valid results?
- Exactly what data do you need to collect? How are you going to organize your data? Have you done a trial experiment?
- Are there any biases in the way you present the experiment? How can you ensure that everyone gets the same instructions?
- Is your experiment a justifiable way of testing your hypothesis? Justify this. What are the possible criticisms? Can you do anything about these?
- Is the experiment reliable? Is it likely that someone else would reach a similar conclusion to you if they used the same method?

Step 3: Do the experiment and collect the data.

Construct a results sheet to collect the data.

Give clear, consistent instructions.

Step 4: Present the data for comparison and analysis.

How are you going to present the data so that the two sets can be easily compared?

How are you going to organize the summary statistics of the two data sets so that you can compare them?

Do you need to find all of the summary statistics covered in this chapter?

Step 5: Compare and analyse.

Describe the differences between your two sets of data.

Make sure that your conclusion is relevant to your aim and hypothesis stated at the beginning.

Step 6: Conclusions and implications.

What are the conclusions from the experiment?

Are they different to or the same as your hypothesis? To what extent? Why?

How confident are you in your results? How could you be more certain?

What is the scope of your conclusions?

How have your ideas changed since your original hypothesis?

Extension

- How could you test whether the spread of the data has changed significantly, rather than the average?
- In what way could you incorporate the work you have done so far on bivariate data?
- It is possible to test data using a more "mathematical" test. Investigate, for example, the "difference in means test".