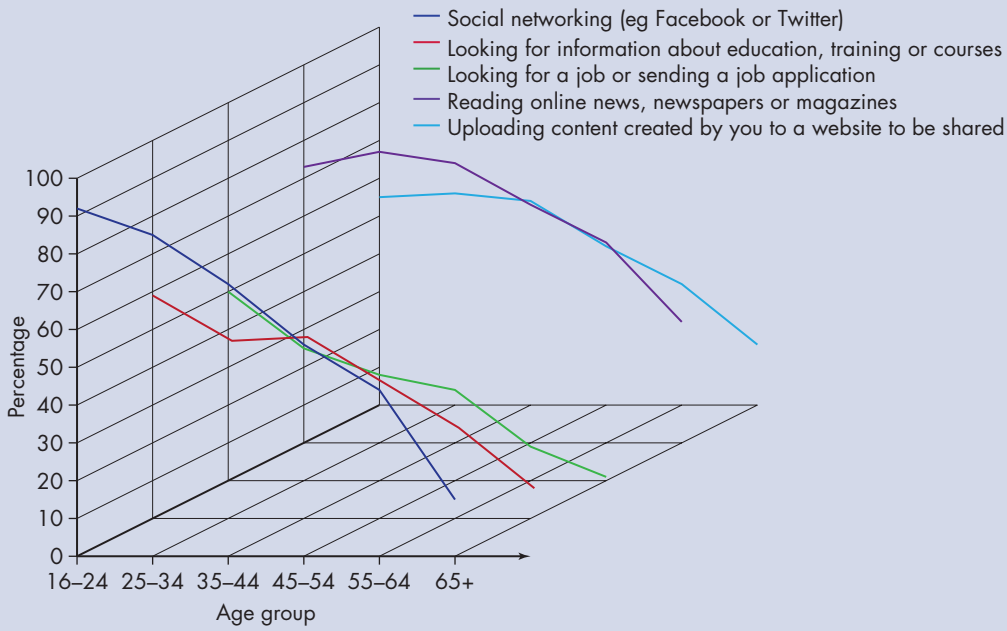
1 Analysis of data

Internet activities by age group, 2015



The chart shows the Internet activities of people in the UK during 2015.

Study the chart. What do you think of it?

Questions

Think about these questions.

1 How easy is it to interpret the chart? What other information would you like to see?

2 Why do the percentages for each age group, except the oldest age group, add up to more than 100%?

3 Do all the Internet activities decrease with age?

4 The Office for National Statistics reports that using the Internet for social networking has continued to grow, rising to 61% in 2015. This was an increase from 45% in 2011, and 54% in 2014. Social networking is widespread in all age groups, up to and including the 55–64 age group. In this group, 44% of adults reported using social networking. Of adults aged 65 and over, 15% use social networking. Social networking has become part of the daily lives of many adults. Of the 61% of adults who used social networking in the last three months, 79% did so more or less every day.

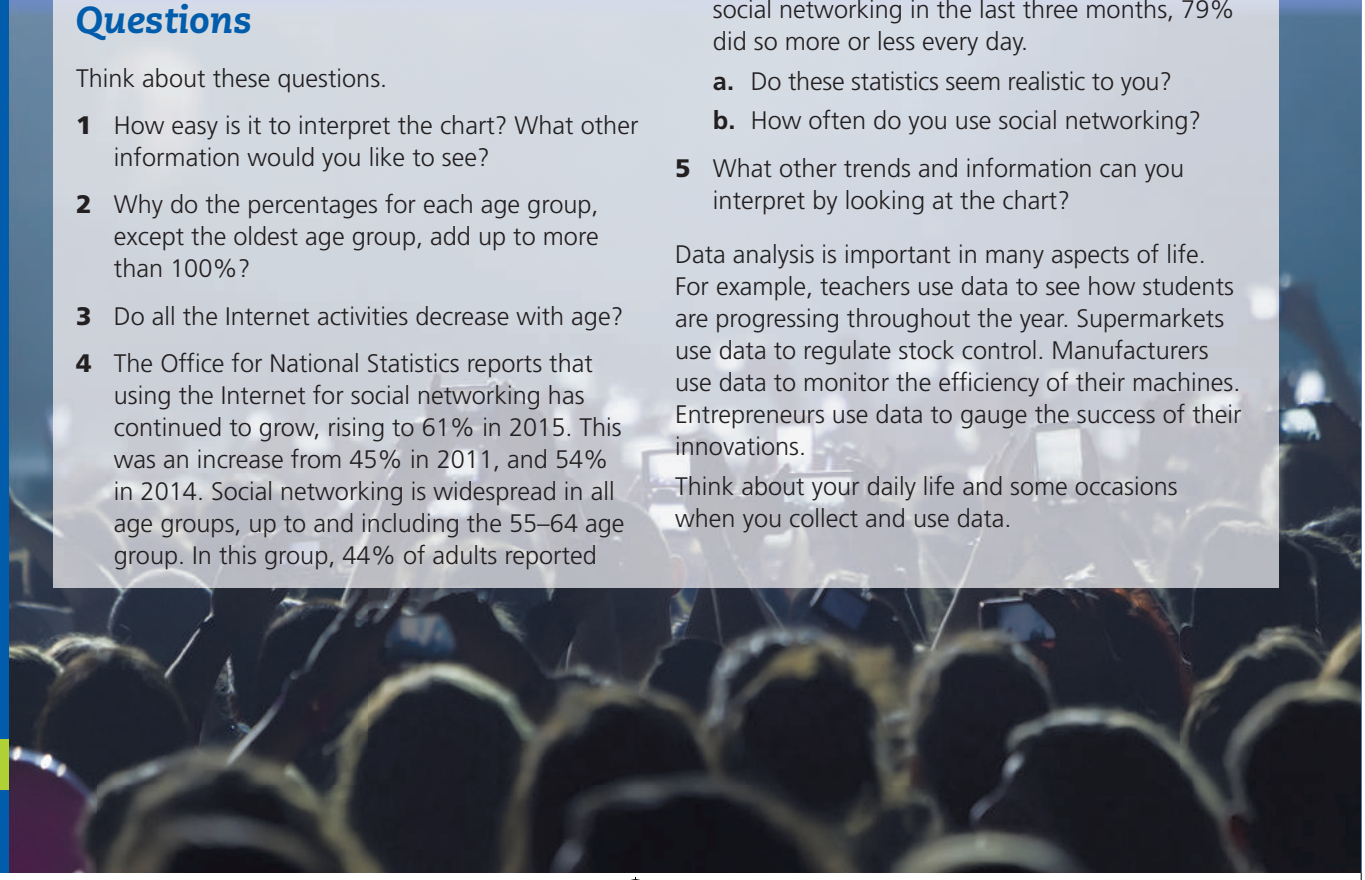
a. Do these statistics seem realistic to you?

b. How often do you use social networking?

5 What other trends and information can you interpret by looking at the chart?

Data analysis is important in many aspects of life. For example, teachers use data to see how students are progressing throughout the year. Supermarkets use data to regulate stock control. Manufacturers use data to monitor the efficiency of their machines. Entrepreneurs use data to gauge the success of their innovations.

Think about your daily life and some occasions when you collect and use data.



D1: Data

Learning objectives

You will learn how to:

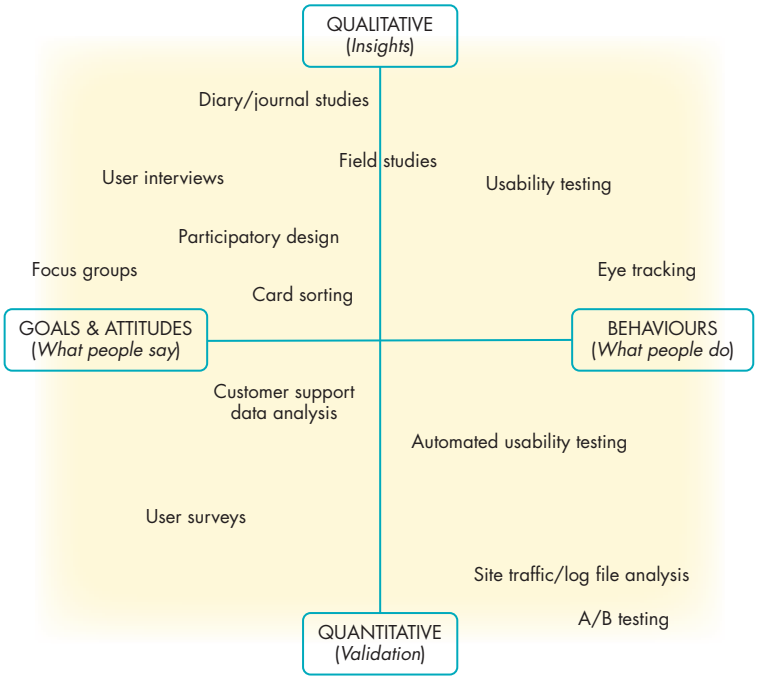
- Identify qualitative, quantitative, primary and secondary data.
- Identify discrete and continuous data, independent and dependent variables.

Introduction

Data is classified to make it easier to process. Data can be collected in the form of measurements or observations of variables. Different kinds of data are represented in different diagrams, according to the type of data. For example, the way an apple wholesaler for a supermarket chain would represent the mass of each apple grown on a particular tree is very different from the way the wholesaler would represent the taste of each variety of apple, or the amount of organic material used in the orchard to grow the apples. The wholesaler must identify the type of data in order to choose the best diagram to use.

Mathematics in the real world

Website designers use qualitative data and quantitative data when gathering research for the design, redesign or improvement of a website. Here is an example of the kind of diagram that a website designer might use.



The landscape of user research and testing techniques.

Qualitative and quantitative data

Qualitative data: non-numerical data that describes a quality.
Quantitative data: numerical data.

- **Qualitative data** is non-numerical and cannot be ranked in a meaningful way. Qualitative data describes a quality that cannot be counted or measured by an instrument. For example, a brick manufacturer might describe a brick by its colour or its roughness.
- **Quantitative data** is data that is numerical and can be counted or measured. For example, the brick manufacturer might describe bricks by their mass or the number of bricks needed to build a house.

Discrete and continuous data

Discrete data: data that is counted.
Continuous data: data that is measured.

Quantitative data can be divided into **discrete** or **continuous** data.

- **Discrete data** is data that is counted, for example, the number of days you walk to college in a week.
- **Continuous data** is data that is measured, for example, the time it takes you to walk to college.

So, if the brick manufacturer describes bricks by their mass, this is an example of continuous data, because the mass can come as any number on a continuous scale. The number needed to build a house, however, is discrete data because the bricks must be provided as a whole number.

Discussion

Choose one of the following scenarios and discuss it with your peers. In each scenario, identify the data that you would collect; then explain why and whether this data would be qualitative or quantitative, discrete or continuous.

1. Imagine that you are an animal breeder. You want to focus on breeding animals for people in your area. The data you collect will allow you to concentrate on the animal breeds that most people in your area want, and the care that these people would expect you to provide for the animals.
2. Imagine that you are a software games engineer. You want to focus on designing a game for 16–18 year olds that encourages them to engage in their studies. The data that you collect will allow you to design a game that works for people in your college, with the possibility of the game becoming national and perhaps even international.
3. Imagine that you are in charge of swimming at the local swimming pool. You want to focus on providing swimming lessons for all who use the baths. The data you collect will allow you to draw up a timetable of when the lessons should take place and the ages and stages to which these lessons would apply.

Exercise 1A

1 Look at the human variables that are listed in i–vi below.

a. List the variables that are qualitative and those that are quantitative.

b Identify which of the following quantitative data are discrete or continuous.

i Hair colour

ii Body temperature

iii Number of teeth

iv Religious faith

v Confidence that you will get this question correct

vi Time it takes to travel from home to college

2 **a** Write down two quantitative variables about your favourite band.

b Identify the variables as discrete data or continuous data.

3 Write down two qualitative variables about the government.

4 You are asked to complete an evaluation form to assess a school lesson, by rating it. Choose from: 1 (excellent), 2 (very good), 3 (average), 4 (poor) or 5 (very poor).

Is this data quantitative or qualitative? Why?

5 A botanist finds a skeleton of a newt, as shown.

a Write down three variables that the botanist might record.

b Say whether the variables are qualitative or quantitative.

c Identify the quantitative data as discrete or continuous.

Primary and secondary data

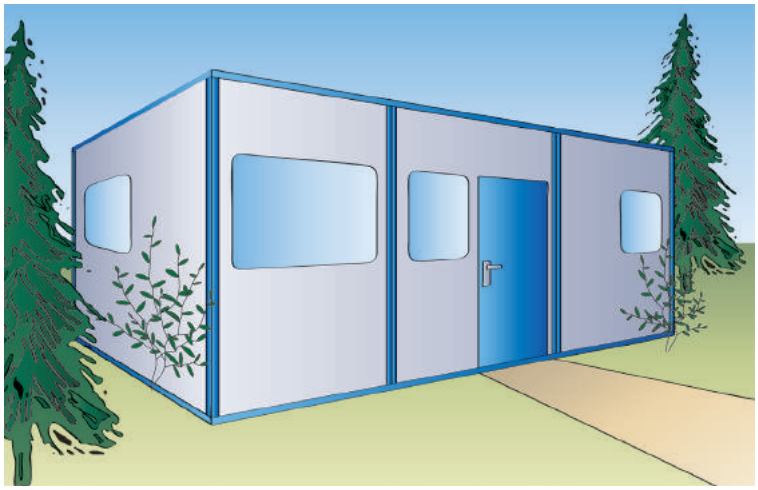
- **Primary data** is data that is collected by a researcher or research company through direct surveys, observations, or interviews with the subjects of the data. Primary data is expensive and time-consuming to obtain, but you know how it was collected – it has been observed or collected directly from first-hand experience.
- **Secondary data** is data that already exists, for example, data from the Office for National Statistics, or from other sources. It can also be data that has been processed in some way, for example, by grouping. Secondary data is cheap and easy to obtain but it might be outdated and from an unknown source. Therefore this type of data could be biased.

Primary data: data that comes directly from first-hand experience.

Secondary data: data that already exists or has been processed.

Mathematics in the real world

Portakabin is a UK-based company that was founded in 1961. It sells and hires out portable and modular buildings. The buildings are used, for example, as offices, nurseries, schools, hospitals, call centres and laboratories. *Portakabin* uses both primary and secondary data to keep up with the way customers' needs for accommodation may change.



Portakabin collects primary data from its customers and staff. The sales team collects data on a regular basis, by talking to customers. The team uses focus groups (groups of people selected to discuss a product before it is launched or to provide feedback on issues) in the working environment to find out what affects workers' performance and productivity. These results are used to develop new products and services. Surveys are used to extract quantitative data.

To improve working conditions, *Portakabin* uses secondary data. They found out from a Gallup survey that 66% of British workers consider that the quality of the working environment is important. It was also reported that noise disturbs 33% of workers, with the result that these employees are four times more likely to become disengaged from working.

Discussion

Choose one of the following scenarios and discuss it with your peers. In each scenario, identify the data you would collect, explaining why and whether this data would be primary, secondary, or both.

1. Imagine that you are an animal breeder and you wish to design and market a new food product for an animal of your choice. Decide if you will concentrate on what you think the person looking after the animal wants in terms of cost, essential food, or luxury food, or if you will concentrate on the needs of the animal and whether the animal will eat the food product.
2. Imagine that you are a software games engineer and you wish to design a new game based on grumpy guinea pigs. Decide if you will concentrate on a particular age range, the quality of the graphics, or the cost of the game to the consumer.
3. Imagine that you are in charge of the local swimming pool and you wish to design a new water feature. Decide if you will concentrate on a particular age range, a certain level of swimming ability, or the cost that will be charged to use the water feature.

Exercise 1B

1 The local maternity clinic is due for a revamp and you are involved.

Decide if you would focus on what patients want, or on what staff would like.

Assume that money is not an option, so you want the best available facilities and equipment.

a What secondary data might you use?

b What primary data might you collect?

2 A householder has a smart meter in her home, which she uses to record her use of gas and electricity.

The smart meter is linked to the energy company, which records and prepares the bills.

The householder records the amount of gas and electricity she uses each month.

The energy company sends a bill and a summary of the monthly usage during the year.

a Is the householder collecting primary or secondary data?

b Is the householder receiving primary or secondary data?

3 The police report that car crime has been reduced in a certain area.

Explain how the police could have used primary and secondary data to reach this conclusion.

4 A campaigning group wants to limit the bonuses paid to bankers.

They need to find out how much is being paid by various banks to the different categories of bankers.

Should they use secondary data or collect primary data, or both?

Think about the advantages, disadvantages, and possible bias of each.

5 Sam, a fruit wholesaler, wants to sell gooseberries.

a What primary data might Sam collect?

b What secondary data might Sam use?

Collecting data

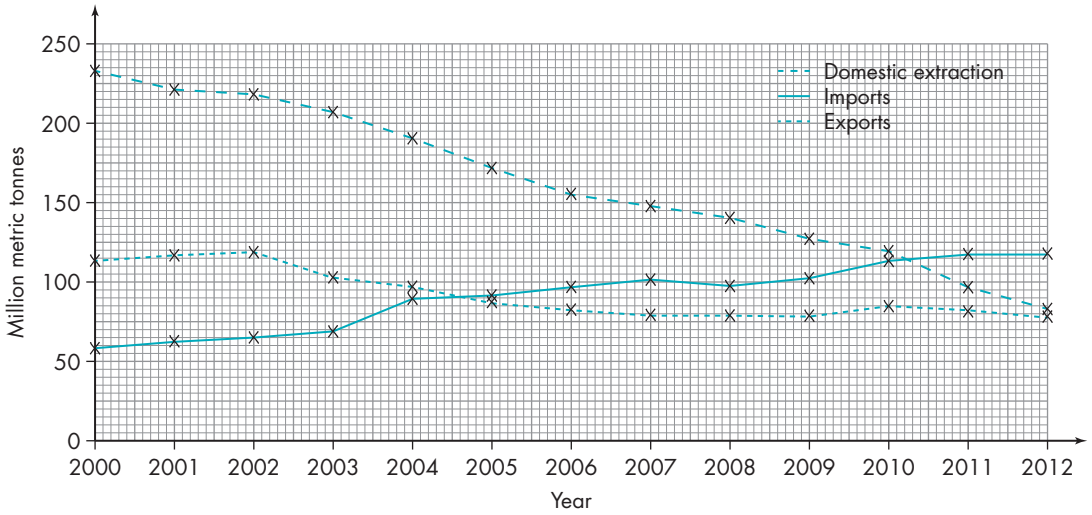
Data can be collected by direct observation, interviews, surveys, experiments and testing. Scientific data in Physics and Chemistry is often collected by doing experiments. In Psychology, data is often collected by testing, observation and interviews, though not exclusively. In experiments there are generally two variables: the variable that you can control, which is called the **independent** (or **explanatory**) variable, and the variable that results, which is called the **dependent** (or **response**) variable. In short, the dependent variable is dependent on the independent variable.

For example, if a farmer wanted to find out the effect of various amounts of fertiliser on his crops, then the amount of fertiliser the farmer puts on the field will be the independent variable, because the farmer can control that. The crops that result will be the dependent variable, as this would depend on the amount of fertiliser the farmer applies.

Mathematics in the real world

In 2014, the Office for National Statistics published a report to show that 'the extraction of oil and gas continues to decline'. The chart in the report shows that over the period 2000–2012, there has been a decline in the quantities of oil and gas extraction.

However, if you are interested in the whole picture of oil and gas supply in the UK over time, the chart does not show the whole picture. It does not show estimates that there are up to 1.3 billion tonnes of undiscovered oil resources, and up to 1010 billion cubic metres of undiscovered gas resources available in the UK.



Discussion

Choose one of the following scenarios and discuss it with your peers. In each scenario, talk about how and why you would collect the data you require and what would be the independent and dependent variables.

1. Imagine that you are an animal breeder and you want to design and market a new food product for an animal of your choice.
2. Imagine that you are a software games engineer and you wish to design a new game based on grumpy guinea pigs.
3. Imagine that you are in charge of swimming at the local swimming pool and you wish to design a new water feature.

Exercise 1C

1 A global software company decides to introduce a new operating system.

They need to research whether to sell the operating system for a one-off charge, charge a fixed amount each month, or charge an amount each month based on how much the system is used.

Advise the company on using primary or secondary data and *how* and *why* they should collect each type of data.

2 The intensive care ward at the local maternity clinic thinks that premature babies develop faster if they spend more time with their mothers.

a What data should the intensive care ward collect?

b Why should they collect it?

c How should they collect the data?

d What are the independent and dependent variables?

3 A high street coffee chain decides to move from using spoons to wooden stirrers.

Before they make the change they decide to do some research.

a What data should they collect?

b Why should they collect the data?

c How should they collect the data?

d What are the independent and dependent variables?

4 Think about your possible career (or interests) and how you will use data to examine something that interests you. Decide what information you would collect and whether you would use primary and secondary data (or both). Also, why should you collect the data? Use what you have learnt so far to write a short description of this (about 150 words). Justify the decisions you make.

D2: Collecting and sampling data

Learning objectives

You will learn how to

- Deduce properties of populations from a sample, whilst realising the limitations.
- Appreciate the advantages and disadvantages of various sampling methods.

Introduction

Sampling is used by researchers to collect data from some of the **population**, in order to make conclusions about all of the population.

A population is generally connected with a complete set of people. However, in statistics, a population can also be everything involved in the study, for example: it might be all the microprocessors made by a firm, or to a diamond merchant, all the diamonds from a particular mine. As a rough guide, if there are n items in the population then the sample size should be $\sqrt{n}$. There are various ways to sample data and this chapter will look at the advantages of **random**, **cluster**, **stratified** and **quota** sampling methods taking into account any **bias**.

Sampling: used to collect data from part of a population.

Population: a complete set of items that share a common property.

Mathematics in the real world

In the 2015 General Election a pre-election poll organised by YouGov surveyed 20 000 people to give an indication of the likely outcome of the election. On Election Day, an exit poll was taken at a sample of polling stations (about 100 out of 40 000), with a sample of about 100–200 voters at each polling station. The problem with exit polls is that the voters or polling stations selected might not be typical of the UK population as a whole (and the people asked might not tell the truth). However, exit polls are generally more accurate than a pre-election poll because the people involved have actually voted. In a pre-election poll, the people asked might not be the people who vote. The exit poll will thus produce a very different outcome to those of the pre-election polls.

The table shows the differences between the pre-election poll, the exit poll and the actual results. At the time, the result predicted by the exit poll was so different to all the pre-election polls that many people did not take it seriously, but in the end, the exit poll proved much closer to the actual result than the pre-election poll.

	Pre-election poll	Exit poll	Actual result
Conservatives	284	316	331
Labour	263	239	232
Liberal Democrats	31	10	8
SNP	48	58	56

Bias

In statistics, **bias** means that results are distorted. This can happen in many ways, for example, picking a non-representative sample such as asking all the students who study Mathematics in college where the College Prom should be held. Other ways in which bias could affect the results is by people being untruthful, errors in processing or recording results, poor survey design, and getting no response to a survey.

Bias: a distortion of results.

Discussion

Choose one of the following scenarios and discuss it with your peers. In each scenario, explain how you test the item mentioned – with a census or a sample – and why you chose that method.

1. Imagine that you are an animal breeder. You have been told about a new food product that can help to improve your animals' health.
2. Imagine that you are a software games engineer. You have designed a new game for older teenagers.
3. Imagine that you are in charge of swimming lessons at the local swimming pool. You want to know parents' reactions to changing the times of the lessons for 3-year-olds.

Exercise 1D

1 An estate agency wants to know the following information. Should the agency use a census or a sample?

- **a** The number of rooms in a bungalow in England
- **b** The number of rooms in a bungalow in an adjacent street
- **c** The number of people who offer less than the asking price of a house
- **d** How long an advert should appear on a website

2 The Government wants to close the bars in the House of Commons.
What population will be affected?

3 A community college wants to close all its facilities on a Sunday.
What population will be affected?

4 The group leader of the village toddler group (ages 2–3) wants to plan an end-of-year excursion.

- **a** Should the group leader use a census or a sample to decide where they should go?
- **b** What population will be affected?

Sampling methods

- **Random sampling** is used to give every item in the population (the sampling frame) the same chance of being chosen. Each item is given a number, then numbers are picked at random, using either random number tables or a computer to identify the items to be measured or surveyed.
- **Cluster sampling** is often used in market research. The population is divided into clusters (groups), then a sample of clusters is selected using random sampling, and all items in those clusters are surveyed.

Random sample: used to collect data from part of a population without bias.

Cluster sample: used to collect data from all members of a randomly selected cluster (or group).

- **Stratified sampling** is used to ensure that each stratum (or layer) is properly represented in the sample. The population is divided into strata (layers) then a random sample from each stratum is selected using random sampling and all selected items are surveyed.
  For example, in a college of 150 males and 250 females, a stratified sample of 80 students is to be selected to participate in a survey about transport. Since male students represent 150 of the college population of 400, then $\frac{150}{400}$ of the sample of 80 should be male. $\frac{150}{400} \times 80 = 30$ So, 30 males must be selected randomly. Therefore 50 female students will also be selected randomly.
- **Quota sampling** is also used in market research. The person doing the surveying is told the quota (amount) of items from each section to be surveyed, and is then free to select those items as she or he wishes.

Stratified sample: used to collect random data from a stratum (or layer) where the number of items selected is proportional to the size of the stratum in the population.

Quota sample: used to collect used to collect data chosen by the sampler from a stratum (or layer) where the number of items selected is proportional to the size of the stratum in the population.

Census or sample?

A **census** is used when every member of a population provides data. The UK National Census is taken every 10 years. It is used to collect data from every household so that plans can be made about many things such as the number of schools, colleges and teachers needed, or how much the NHS might have to spend on the care of elderly people. A large census like this is very expensive, so a **sample** is often used to save both time and money. Conclusions can be made from the sample about the whole population from which it came. The larger the sample the more accurate it is likely to be, but increasing the sample size costs extra time and money.

Census: used when every member of the population provides data.

Sample: part of a population.

Mathematics in the real world

- **Random sampling** is used in the National Lottery when six balls are selected at random from 49.
- **Cluster sampling** has become a popular method to perform immunisation coverage surveys. This method has proved to be very suitable in dispersed rural populations in Kenya.

  The names of children aged 5–14, who attended school were known, and 30 clusters were needed.

  The total school population was divided by 30 to determine the sampling interval, a. Then a random number, n, was chosen and the school of the nth child in alphabetical order was the first cluster to be surveyed. The school of the $(n + a)$th child in alphabetical order was the second cluster to be surveyed. The school of the $(n + 2a)$th child in alphabetical order was the next cluster to be surveyed, and so on, until 30 clusters (in this case, primary schools) were surveyed. If the same school is selected twice, then the school of the next child (the $(n + ra + 1)$ th child) on the list is substituted in its place.
- **Stratified sampling** is used for a political survey. If the survey needs to reflect the diversity of the population, the researcher would want to include participants of various groups such as race or religion, based on their proportionality to the total population. A stratified survey could therefore claim to be more representative of the population than a survey done using other sampling methods.
- **Quota sampling** is something you may have experienced when people who are doing a market survey in the street stop you and ask if you would answer some questions about a certain product. The marketing

people will have been told that they need to ask a certain number of people aged, for example, 16–20, 21–30, 31–40. Thus, they are stratifying the population, but rather than selecting the people in each group randomly, they are free to choose who they will question.

Advantages and disadvantages of various types of sampling

Type	Conditions of use	Advantages	Disadvantages
Random	Population members are similar to each other	Free from bias	Tedious and time-consuming
Cluster	Population consists of units rather than individuals (for example, the types of trees in parks in different areas of the UK)	Cheaper than random sampling, as it can reduce travel and admin costs; can show regional variation	Not a genuine random sample; can be subject to bias if only a few clusters are used
Stratified	Population members are similar to each other but contain several easily identifiable groups (such as gender and religion)	More accurate than simple random sampling, as a fair proportion of responses from each stratum is obtained; can show different tendencies in each stratum	Tedious and time- consuming
Quota	When there are several easily identifiable groups (such as gender and religion) and stratified sampling is not possible	Simple to take	Not genuinely random, since the surveyor picks the items and so is likely to be biased

Discussion

Choose one of the following scenarios and discuss it with your peers.

1. Imagine that you are an animal breeder and you wish to design and market a new food for an animal of your choice. What kind of sampling method would you use to find out what food the animal owners would buy? Why? Would you use the same sampling method to find what food the animals would eat? Why? In each case say what the sampling frame is.

2. Imagine that you are a software games engineer and you wish to design a new game based on grumpy guinea pigs. What kind of sampling method would you use to find out what features the gamers would buy? Why? What sampling frame would you use? Why?

3. Imagine that you are in charge of swimming at the local pool and you wish to design a new water feature. What kind of sampling method would you use to find out what features the bathers would use? Why? What sampling frame would you use? Why?

Time and money

There is always a fine balance to be found between removing bias and increasing sample size. If a sample is too small then it could easily be biased, as it may not be representative of the population. However, if the sample is large, this increases the cost of the survey, which may then become costly to conduct. Therefore, there is a need to balance the accuracy of the survey with the cost. At the start of this section it was mentioned that if there are n items in the population then the sample size should be $\sqrt{n}$.

Exercise 1E

In each of the following questions, estimate the size of the population and then suggest a suitable sample size. At this point in the course the estimate of the population size is not important. You will learn more about this kind of estimation in Chapter 3. What matters is that the sample size should be suitable for the population estimated.

1 The local maternity clinic is due for a revamp.

The registrar decides to give a survey to all the people who enter the clinic on Friday to find out their opinion on what catering facilities should be provided.

Explain why the sample is biased and suggest a better way of sampling.

2 A householder has a smart meter to record the use of gas and electricity.

This is linked to the energy company so that the bills can be prepared.

The householder wishes to estimate the use of his gas and electricity over a year to help with budgeting.

The householder records the amount of gas and electricity used in March and uses this to work out an estimate of the gas and electricity used throughout the year.

Explain why the sample is biased and suggest a better way of sampling.

3 The police want to know if car crime has been reduced in County Durham.

a What is the population they should use?

b What would be a suitable method of sampling?

c Why would that be suitable?

4 A campaigning group wants to limit the bonuses paid to bankers.

They decide that they need to know the amounts that bankers are paid, what non-bankers think of certain amounts of bonuses, and if it would be better to pay bankers a higher salary to avoid confrontation in future.

a What would be a suitable method of sampling to find out the information they want?

b Why would the method be suitable?

5 A biologist has an interest in oak galls (a species of small wasps).

Oak galls are in small spheres formed on oak trees; when the grubs inside turn into tiny wasps they burrow out by making small holes.

To investigate the number of oak galls that come from oak trees, the biologist collects all the galls that have fallen into his garden from a nearby oak tree.

a What is the population?

b Why might the sample be biased?

c Explain a better sampling method.

6 Look at your answer to Exercise 1C question **4**, about your possible career (or interests) and how you will be using data to examine something that interests you. Think about how you would collect the data you described. What possible sampling methods might you use? How could you avoid bias? Write a short description of about 150 words, justifying the decisions you make.

D3: Representing data numerically

Learning objectives

You will learn how to:

- Calculate measures of location (mean, median, mode, quartiles and percentiles) and spread (range, interquartile range and standard deviation).
- Interpret these measures and use them to make conclusions.

Introduction

Most of these measures (**mean**, **median**, **mode** and **quartiles**, **range** and **interquartile range**) you will have met before in your GCSE course. This section will focus on **percentiles** and **standard deviation,** although there will be some questions to remind you of what you have met before.

Here is a summary of the measures you should have met before.

Measure	Advantages	Disadvantages	Example
Mean	Probably the most used average; uses every item of data	Might not be representative if there is an extreme value	1, 50, 52, 56, 56, 56, 58 Mean = $\frac{1+50+52+56+56+56+58}{7} = 47$ This is a lower value than most of the data.
Median	Only looks at the middle value, so not affected by extreme values	Can be misleading because it does not consider all the values	1, 1, 1, 4, 56, 56, 58 Median is the 4th value; median = 4 Here the numbers below the median are close together, but the others and a long way above the median.
Mode	Easy to find: no calculation needed; the only average that can be used with qualitative data	Can be misleading; inappropriate if the data contains few duplicates	1, 1, 1, 4, 56, 56, 58 Mode is 1 This is not representative of the data as a whole.
Quartiles	They divide the data into four equal groups	Not easy to calculate if the data set is small	Often referred to as the **lower quartile**, **median** and **upper quartile** 0, 3, 7, 10, 12, 17 so the median lies between 7 and 10, that is, 8.5. The lower quartile (LQ) divides the data below the median (0, 3, 7) into two groups, and is therefore 3. The upper quartile (UQ) divides the data below the median (10, 12, 17) into two groups and is therefore 12.
Range	It measures the spread of the data	It only uses the two extreme values, which may not be representative of how the data is spread	1, 51, 51, 54, 56, 56, 98 The range is 98 – 1 = 97, but without the end values the range would be only 5.

Interquartile range (IQR)	Measures the range of the middle 50% of the data, so is not influenced by extreme values	Half the data plays no part in this measure of spread	0, 3, 7, 10, 12, 17, 20 LQ is the $\frac{7+1}{4}$ = 2nd value: LQ = 3 UQ is the $\frac{3(7+1)}{4}$ = 6th value: LQ = 17 IQR = UQ – LQ = 17 – 3 = 14

Measures of location are sometimes referred to as **measures of position** and **measures of spread** as **measures of dispersion**.

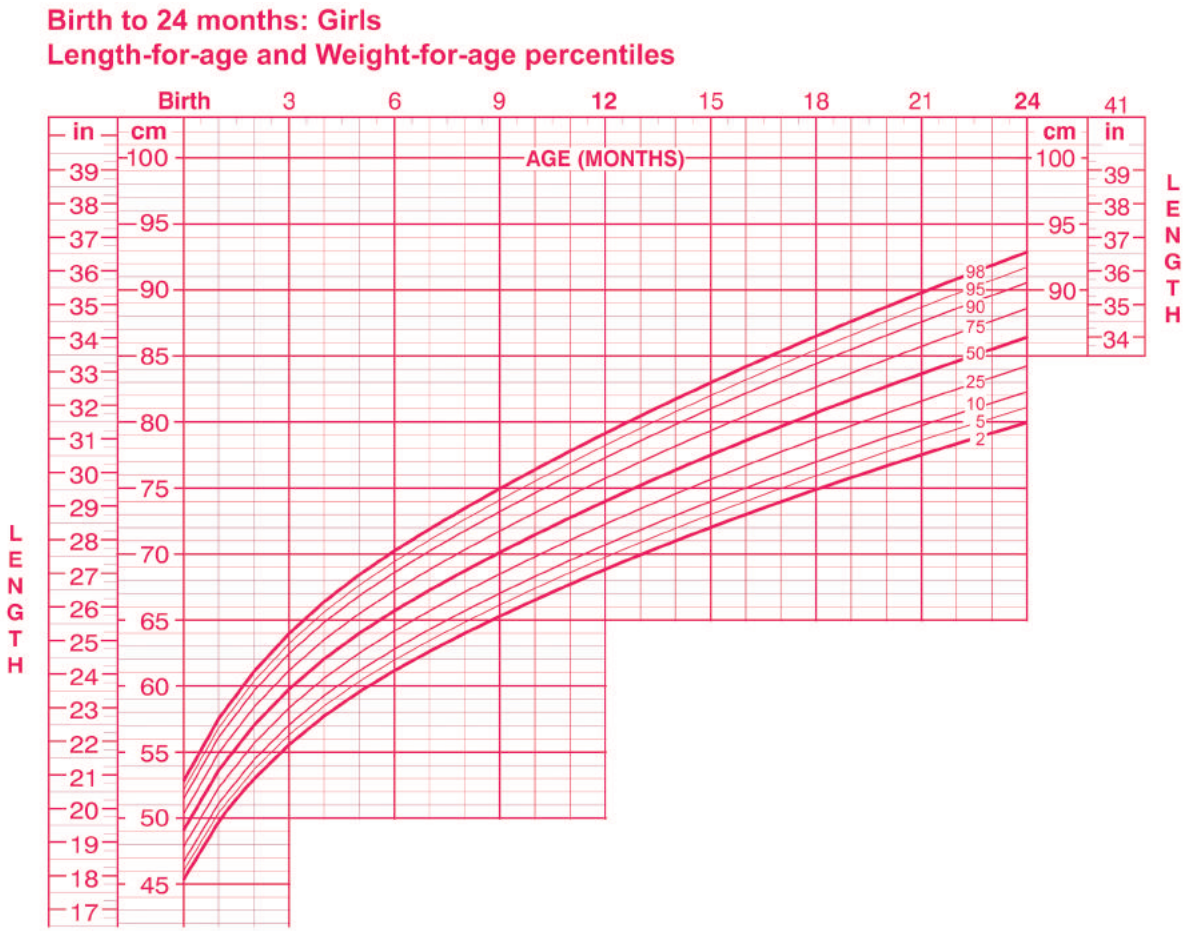
- **Percentiles** divide data into 100 equal parts, just as quartiles divide data into four equal parts (quarters). Therefore the lower quartile Q_1 is also the 25th percentile P_{25}, the median or Q_2 is the 50th percentile P_{50} and the upper quartile Q_3 is the 75th percentile, P_{75}. Sometimes the range between the 10th and 90th percentile is used, $P_{90} - P_{10}$, as the measure of spread because it uses more data and still avoids extreme values.

Percentile: a value below which a given percentage of the observations fall.

Mathematics in the real world

The World Health Organization publishes charts so that parents can see how their child is growing compared to other children. This also allows health workers to judge whether a child is progressing well or needs help.

Here is part of the chart for girls aged 0–24 months.



The numbers on the curves (98, 95, 90, 75, 50, 25, 10, 5, and 2) refer to P_{98}, P_{95}, P_{90}, P_{75}, P_{50}, P_{25}, P_{10}, P_5 and P_2. These are strategic percentiles, as they correspond to strategic points in a population that will become more apparent when dealing with standard deviation.

Discussion

Choose some of the following questions and discuss them with your peers.

1. Look at the chart. What information do you think it shows?
2. In which month is the length of the baby increasing most? How do you know without looking at the numbers on the scales?
3. Would you be worried if a 12-month-old girl was 71 cm long?

Exercise 1F

Use the chart above to help you to answer these questions.

1 Find the median length in centimetres of the following.

a A 6-month-old girl

b A 19-month-old girl

2 Work out the interquartile range in inches of a 15-month-old girl.

3 Work out P_{90} for a 17-month-old girl.

4 What percentage of 12-month-old girls would you expect to be longer than 79 cm?

5 What percentage of 17-month-old girls would you expect to be shorter than 77 cm?

6 The length of a baby girl is 70 cm. Between what ages might she be?

7 Work out $P_{90} - P_{10}$ for a two-year-old girl.

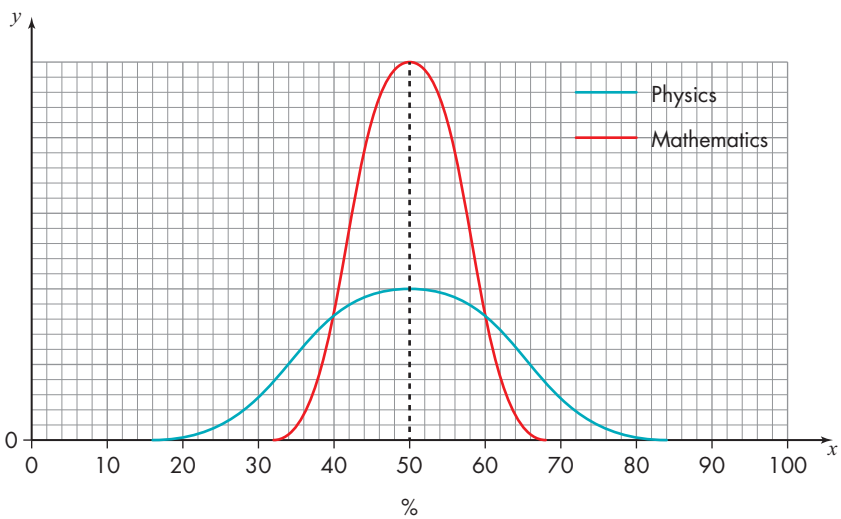
8 Seventy-five per cent of what age of baby girls is above 72 cm?

Standard deviation

Standard deviation: a measure of spread that uses all the data.

- **Standard deviation** is a measure of the spread of data using every item of data. It is especially useful when comparing values from different sets of data and when analysing the position of an item of data in a **normal** population. An advantage of standard deviation is that it uses all the data. A disadvantage is that it takes longer to calculate than the range or interquartile range if you do not use a calculator.

The following graph shows an example of standard deviation being used to compare two tests, one in Mathematics and one in Physics. The mean mark is the same in each case: 50%. The standard deviation in the Mathematics test is 5% and the standard deviation in the Physics test is 10%. This means that the marks in the Physics test are more spread out than the marks in the Mathematics test.



Calculating standard deviation

The standard deviation is sometimes called the 'root mean square deviation from the mean' because that is how it is calculated!

The reason why you square, then square root, is because if you added up the differences from the mean, that part would come to zero as some would be negative and some positive. Squaring a negative number makes it positive, so you square the differences from the mean before you add them up and then divide by the number of items. After you have done that you take the square root because squaring and square rooting are inverses of each other.

For a set of data which is expressed as $x_1, x_2, \ldots x_n$ with a mean $\bar{x}$ the standard deviation σ is

$$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

Beyond the spec

This looks like a very complicated formula, so here is a step by step method to show how it works with a set of data {2, 3, 6, 9, 15}.

Step	Algebraic notation	Example
1 Find the mean.	$\bar{x}$	$\frac{2+3+6+9+15}{5} = 7$
2 Find the deviations from the mean.	$x-\bar{x}$	2 – 7, 3 – 7, 6 – 7, 9 – 7, 15 – 7 that is –5, –4, –1, 2, 8
3 Square these deviations.	$(x-\bar{x})^2$	25, 16, 1, 4, 64
4 Find the mean of these squared deviations.	$\frac{\sum(x-\bar{x})^2}{n}$	$\frac{25+16+1+4+64}{5} = \frac{110}{5} = 22$
5 Find the square root of this result.	$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$	$\sqrt{22} = 4.69\ldots$

There is an easier formula to use which involve doing fewer subtractions.

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

Again, this looks like a very complicated formula, so here is a step-by-step method to show how it works with a set of data {2, 3, 6, 9, 15}.

Step	Algebraic notation	Example
1 Find the mean.	$\bar{x}$	$\frac{2+3+6+9+15}{5} = 7$
2 Square each item of data and sum them.	$\sum x^2$	4 + 9 + 36 + 81 + 225 = 355
3 Find the mean of the sum of the squares.	$\frac{\sum x^2}{n}$	$\frac{355}{5} = 71$
4 Subtract the square of the mean.	$\frac{\sum x^2}{n} - \bar{x}^2$	71 – 49 = 22
5 Find the square root of this result.	$\sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$	$\sqrt{22} = 4.69\ldots$

For discrete frequency distributions the formulae are:

$$\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}} \text{ or } \sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}.$$

$\sum f$, the sum of the frequencies, is used instead of n.

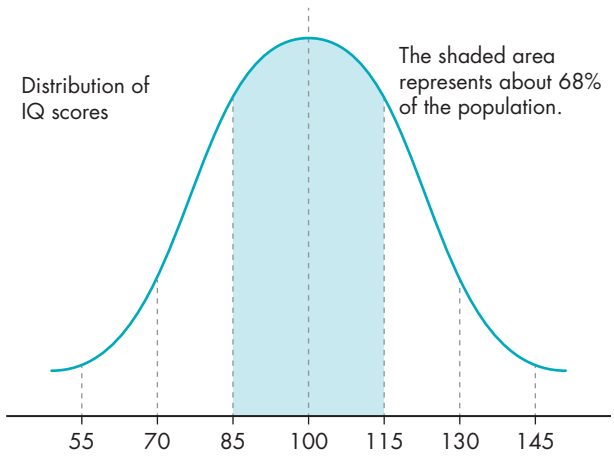
For grouped frequency distributions you use the mid-point of the group for x.

In the real world, no one works out standard deviation by hand. Everyone uses a calculator or computer. If you use a calculator, read the manual carefully to see how to enter the data efficiently. There is no difference in the marks awarded in the examination if you use the statistical buttons on a calculator or work it out step by step.

Mathematics in the real world

Intelligence quotient (IQ) scores are based around a normal distribution of intelligence where the mean score is 100 (this is what a person with average intelligence would expect to achieve - it is just a convenient number on a scale) and the standard deviation is 15 (again, just a convenient number with which to work).

About 68% of the population has an IQ within one standard deviation, either side of the mean, that is, between 85 and 115; around 95% have an IQ within two standard deviations (70 to 130) and about 99.7% within three standard deviations (55 to 145). If IQ had been set up with a mean score of 10 and a standard deviation of 3, then about 68% of the population would still have an IQ within one standard deviation, either side of the mean, that is, between 7 and 13; around 95% have an IQ within two standard deviations (4 to 16) and about 99.7% within three standard deviations (1 to 19).



Beyond the spec

Use of standard deviation in standardising scores

Standard scores (or **z-scores**) are often used to compare data from different samples where the means and standard deviations are known.

Here is some real data showing one year's results from the Biomedical Admissions Test given to applicants who want to study medicine and veterinary courses at Cambridge University.

Gender	Mean	Standard deviation
Female	5.03	0.60
Male	5.23	0.75

Imagine that as an admissions tutor you have only one place and two candidates: a male who scored 6.43 on the test and a female who scored 6.23 on the test. Which of these represents the better performance?

To compare the scores, calculate how many standard deviations each score is from the mean.

The woman's score is $\frac{6.23-5.03}{0.60}=\frac{1.2}{0.60}=2$ standard deviations above the mean.

The man's score is $\frac{6.43-5.23}{0.75}=\frac{1.2}{0.75}=1.6$ standard deviations above the mean.

Therefore the woman's performance is better than the man's performance because it is more standard deviations away from the mean.

This means that among the women, the score of 6.25 represented a better performance in that group, than a score of 6.43 by the man in his group.

This sort of standardising can be used in situations where it is known that one group is known to generally score higher than another group. Say, for example, that a university or college wants to have similar numbers of boys and girls on a particular course. They know that girls generally achieve higher scores in a particular test, therefore they can standardise the scores to achieve the required result.

Discussion

Choose some of the following questions and discuss them with your peers.

1. Do you think this is a fair way of comparing performances? If so, why? If not, why not?
2. What do you think the male score should be to be comparable with the female score, above?
3. What do you think the female score should be to be comparable with the male score, above?
4. What can you say about a standardised score if it is above average?

Exercise 1G

In all these questions you can use the statistical buttons on a calculator (no extra credit is given for using a step-by-step method). If your calculator uses buttons with σ_n and σ_{n-1}, the preferred use is the σ_{n-1} button, but there are no penalties for using the σ_n button.

1 Calculate the mean and standard deviation of the following.

Check your results using a calculator.

a 3, 5, 6, 9, 11, 12, 17

b 30, 50, 60, 90, 110, 120, 170

c 9, 11, 12, 15, 17, 18, 23

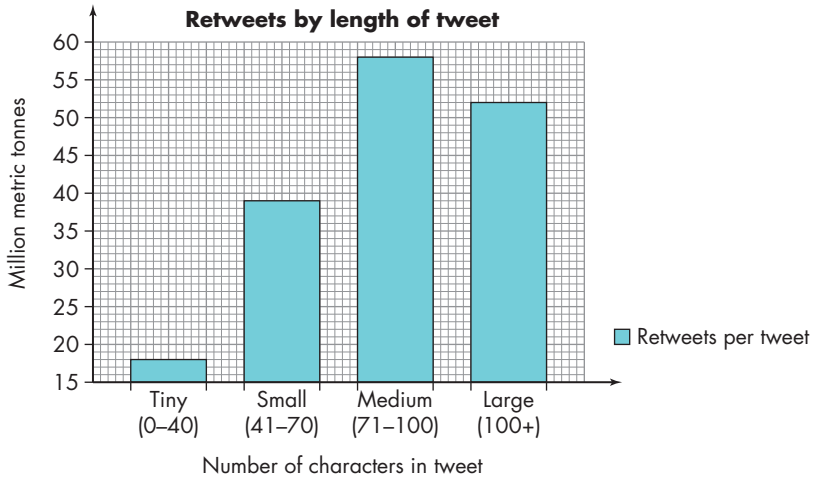
2 Here are the lengths of the stirrers from Exercise 1C, question 3.

Length in millimetres	11.0	14.0	17.8	19.0
Frequency	2	52	14	13

a Calculate the mean and standard deviation length of stirrer.

b What can you say about the mean length compared with the median and modal length?

3 Based on the information in this bar chart of retweet lengths, calculate the mean and standard deviation of a length of retweet.



Data source: Retweets by Length of Tweet, Track Social.

Beyond the spec

4 Use the formulae to calculate the mean and standard deviation for the following data.

a $\frac{\sum x^2}{n} = 69, \bar{x}^2 = 44$

b $\frac{\sum (x-\bar{x})^2}{n} = 16, n = 8, \sum x = 72$

c $n = 10, \sum x^2 = 612.5, \sum x = 35$

5 Find the better result in each of the following. Use standardised scores to decide.

a 56 where $\bar{x} = 50$ and $\sigma = 3$ or 45 where $\bar{x} = 50$ and $\sigma = 3$.

b 23 where $\bar{x} = 30$ and $\sigma = 4$ or 57 where $\bar{x} = 65$ and $\sigma = 3.7$.

c 4.5 where $\bar{x} = 3.9$ and $\sigma = 0.45$ or 3.9 where $\bar{x} = 2.1$ and $\sigma = 1.2$.

6 Look at your answers to Exercise 1C, question **4** and Exercise 1E, question **6** about your possible career (or interests) and how you will use and collect data to examine something that interests you. Think about how you would analyse the data you described. What possible statistical measures might you calculate? Why would you use them? Write a short description of about 150 words, justifying the decisions you make.

D4: Representing data diagrammatically

Learning objective

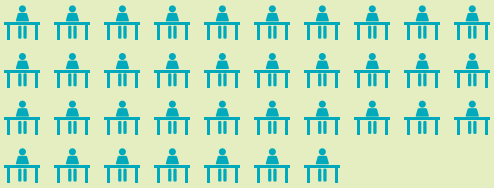
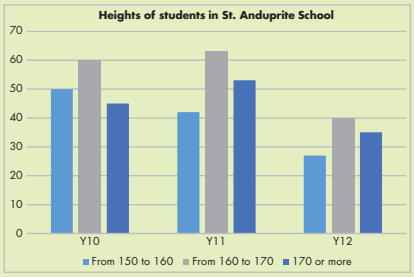
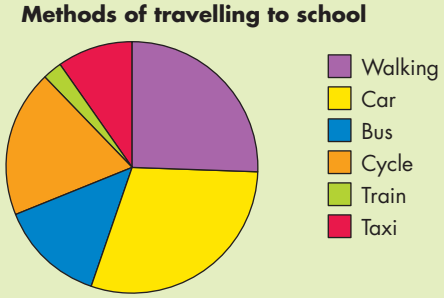
You will learn how to:

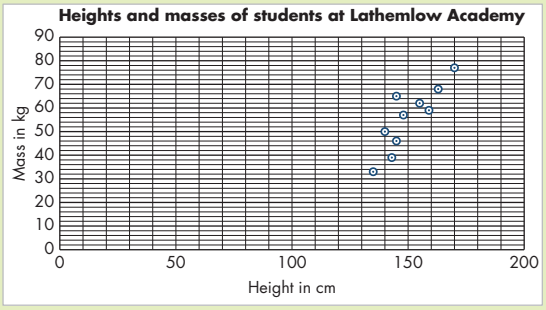
- Construct and interpret diagrams for grouped discrete data and continuous data, know their appropriate use and reach conclusions based on these diagrams.

Introduction

There are many ways to represent data in diagrams and you will have had considerable experience in drawing and interpreting them in your GCSE course. The diagrams you will concentrate on here (**histograms**, **cumulative frequency graphs**, **box-and-whisker plots** and **stem-and-leaf diagrams**) will develop those skills.

Here is a resume of the diagrams with which you will be familiar, and their distinguishing features.

Diagram	Advantages	Disadvantages	Example
Pictogram	Data is easily counted and represented using symbols; often used with qualitative data	Difficult to draw and hard to represent fractions of symbols	Number of countries participating in the International Mathematical Olympiads 1976 Austria  1986 Poland 
Bar chart (including multiple and composite charts)	Good for comparing data in different categories; easily understood	Shows the number of items, but not their actual values	
Pie charts (including proportional charts)	Good visual representation of the proportions; often used with qualitative data	Individual data is lost; not really suitable if there are many categories	

Scatter graphs	Used to determine correlation between two variables	Time-consuming to draw	

Stem-and-leaf diagrams

These are simple diagrams that are good for comparing small amounts of data. The data is not lost.

They are not really appropriate if there is a lot of data, as the appearance can be off-putting.

Example 1

Two classes sit the same test that is marked out of 50.

Class A students score 42, 36, 27, 27, 13, 41, 48, 17, 29, 45, 17, 10, 30, 22, 8, 38, 41, 34, 28, 26, 46, 18, 32, 17, 24, 38, 16, 21 and 12 marks.

Put the scores into a stem-and-leaf diagram by writing the stems (the digits in the tens column) in a vertical line, then adding the leaves (the units digits) in order alongside the appropriate stem.

Notice that a key is always included, showing how the chart represents the data.

Then find the median: here there are 29 items of data, so the median is the 15th item. Counting from the largest item downwards (or the smallest item upwards) this is 27 (circled in red).

The median divides the data into two groups of 14 items each, so the quartiles will divide each of these groups equally and be found halfway between the seventh and eighth values, from the top, and halfway between the seventh and eighth values from the bottom, that is, the lower quartile is 17 (halfway between two 17s) and the upper quartile is 38 (halfway between the two 38s).

Class A

		Key 3 | 6 = 36 marks
4	1 1 2 5 6 8	
3	0 2 4 6 8 8	
2	1 2 4 6 7 (7) 8 9	
1	0 2 3 6 7 7 7 8	
0	8	

Class B students scored 21, 36, 45, 38, 15, 46, 29, 23, 35, 35, 19, 13, 40, 15, 24, 27, 41, 35, 14, 45, 37, 37, 28, 16, 42, 26, 33 and 46.

Add their marks to the other side of the stem-and-leaf diagram, so that the two classes can be compared.

Marks in a test

Class B		Class A
6 6 5 5 2 1 0	4	1 1 2 5 6 8
8 7 7 6 5 5 5 3	3	0 2 4 6 8 8
9 8 7 6 4 3 1	2	1 2 4 6 7 7 8 9
9 6 5 5 4 3	1	0 2 3 6 7 7 7 8
	0	8

Key 5 | 3 | 6 means 35 marks in Class B and 36 marks in Class A

Box-and-whisker plots

Box-and-whisker plots make it fairly easy to find quartiles and interquartile ranges. They are good for comparing two data sets and summarising data. However, some work must be done on the data before drawing the plot diagram, so drawing a stem-and-leaf diagram first will help a lot.

In box-and-whisker plot diagrams, the individual data is lost and the mean and mode are not identifiable. They are probably best drawn on squared paper.

Example 2

Use the same data as we used in the stem-and-leaf diagram above.

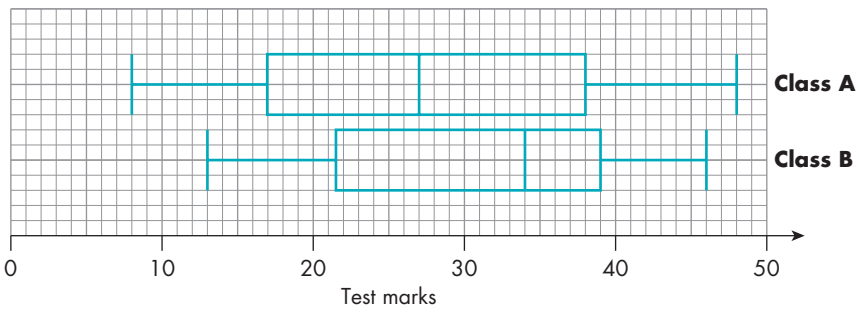
From the stem-and-leaf diagram, the data is summarised below:

	Minimum mark	Lower quartile	Median	Upper quartile	Maximum mark
Class A	8	17	27	38	48
Class B	13	21.5	34	39	46

The 'box' is a rectangle that is drawn from the lower to the upper quartile; the line inside it represents the median.

Draw the 'whiskers' at the minimum and maximum values, joined to the box, as shown.

The two box-and-whisker plot diagrams are shown above each other, so that you can compare the two classes easily. The plot diagrams can be drawn horizontally like this, or turned through 90°.



Discussion

Compare the two classes. Think about comparing the measures of spread and the measures of location. Which class do you think did better? Why?

Cumulative frequency graphs

These are used when grouped data is given, so the individual data have been lost.

It is easy to find percentiles and quartiles using cumulative frequency diagrams.

Example 3

Here is a grouped frequency table showing the time spent on mobile phone calls.

It shows, for example, that 28 calls lasted between 10 and 20 minutes.

Time, t, in minutes	$0 < t \leqslant 5$	$5 < t \leqslant 10$	$10 < t \leqslant 20$	$20 < t \leqslant 30$	$30 < t \leqslant 60$
Frequency	8	15	28	20	9

Now work out the cumulative frequencies by adding up the frequencies as you go along.

Time, t, in minutes	$0 < t \leqslant 5$	$0 < t \leqslant 10$	$0 < t \leqslant 20$	$0 < t \leqslant 30$	$0 < t \leqslant 60$
Cumulative frequency	8	23	51	71	80

Notice that the time intervals have also changed: 51 calls lasted less than or equal to 20 minutes.

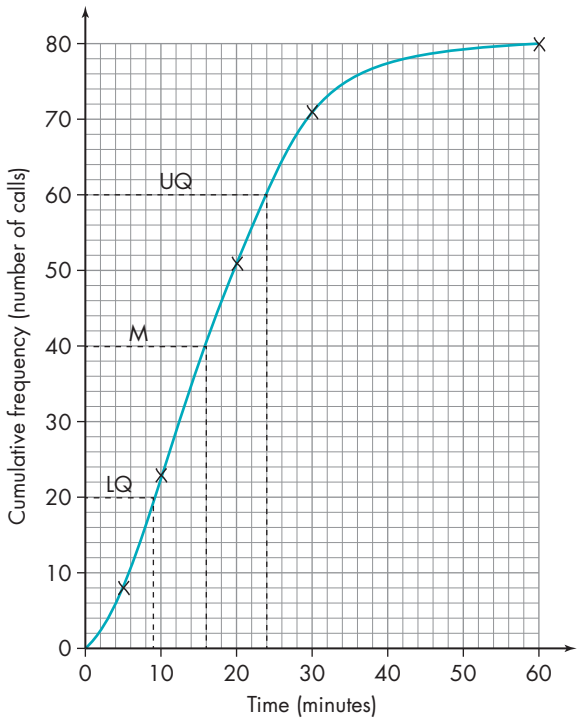
Now plot these cumulative frequencies on graph paper, making sure that you plot the cumulative frequencies at the end of the interval, that is, at (5, 8), (10, 23), (20, 51), (30, 71) and (60, 80).

You can also plot the point (0, 0), as no calls lasted 0 minutes or less.

Join the points with a smooth curve (not straight line segments) to create the graph shown here.

Find the median by drawing a line from 40 (half of 80) on the cumulative frequency axis to the graph, then down to the time axis: here you can see that the median is 16 minutes.

Read the upper quartile (24 minutes) and lower quartile (9 minutes) in a similar manner.



Histograms

Histograms are used especially when grouped continuous data is given, more so when the group intervals are not all the same. The difference between bar charts and histograms is that with bar charts, the height of the column is proportional to the frequency of that interval; with histograms, it is the area of the column that is proportional to the frequency. This means that the vertical axis should always be labelled with frequency density, not just frequency.

Frequency density: the number of items in a given unit.

Example 4

Here is a grouped frequency table showing the time spent on mobile phone calls that you used for the cumulative frequency graph.

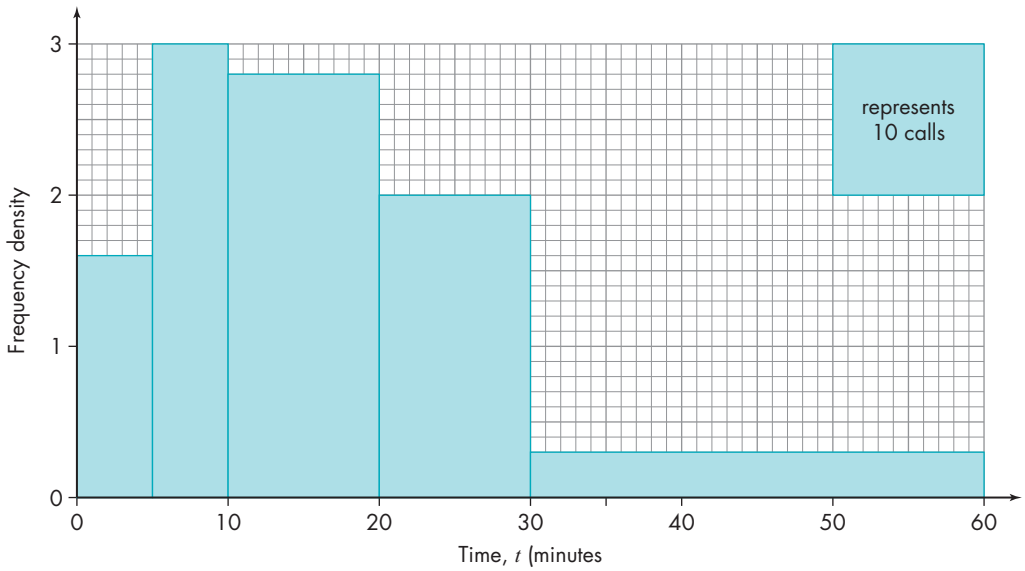
Time, t, in minutes	$0 < t \leqslant 5$	$5 < t \leqslant 10$	$10 < t \leqslant 20$	$20 < t \leqslant 30$	$30 < t \leqslant 60$
Frequency	8	15	28	20	9

Work out the frequency densities (in this case the number of calls per minute) by dividing each frequency by the width of its interval.

Time, t, in minutes	$0 < t \leqslant 5$	$5 < t \leqslant 10$	$10 < t \leqslant 20$	$20 < t \leqslant 30$	$30 < t \leqslant 60$
Frequency	8	15	28	20	9
Frequency density	$8 \div 5 = 1.6$	$15 \div 5 = 3$	$28 \div 10 = 2.8$	$20 \div 10 = 2$	$9 \div 30 = 0.3$

Finally, put this information into the histogram, as shown.

It is useful (but not essential) to shade a square stating the frequency it represents.



Notice that if you multiply the height of the column by its width, you obtain the number of calls (the frequency) of that interval.

Histograms are used because they do not distort the visual impression given by a bar chart when there are unequal class intervals.

Exercise 1H

1 Here are the scores of Sam and Ella after throwing darts at a dartboard.

Sam scored 27, 54, 16, 1, 39, 5, 60, 25, 8, 40 and 20.

Ella scored 26, 12, 51, 20, 50, 19, 48, 57, 30, 24, 21 and 15.

a Draw a back to back stem-and-leaf diagram showing Sam and Ella's scores.

b Draw box-and-whisker plot diagrams showing their scores.

c Write a short report to compare their scores, using your diagrams to help you.

2 Elsa and Christof find a snowdrift and decide to make snowballs for Olaf.

Here is a grouped frequency table showing the mass in grams of each snowball.

Mass, m, in grams	$40 < m \leqslant 50$	$50 < m \leqslant 60$	$60 < m \leqslant 80$	$80 < m \leqslant 100$	$100 < m \leqslant 105$
Frequency	9	30	38	32	11

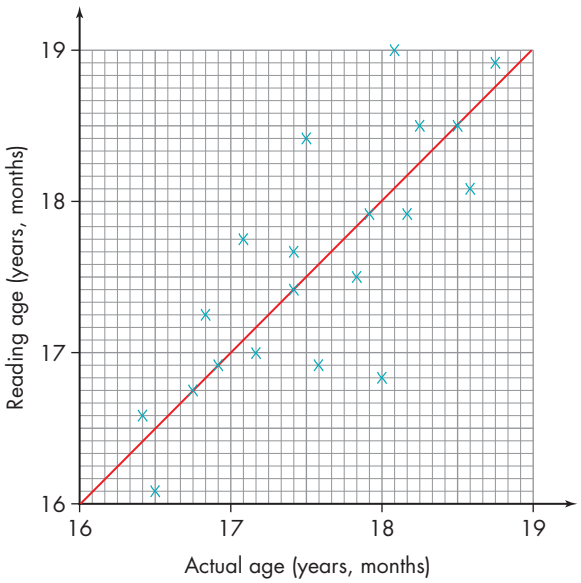
a Draw a cumulative frequency graph and use it to estimate:

i the median mass of a snowball

ii the upper and lower quartiles masses of the snowballs.

b Construct a frequency density table and use it to draw a histogram displaying the masses.

3 Here is a scatter graph of the actual age and reading age of a group of sixth formers. Describe the distribution, then use the scatter graph to answer the following questions.



a What is the reading age of the student who is 17 years and 10 months?

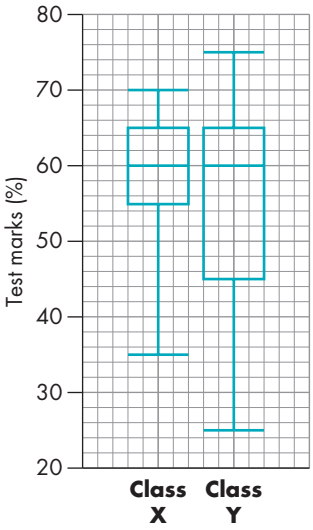
b What is the range of reading ages for the students?

c What is the greatest difference between the reading age and actual age of a student?

d What is the median reading age?

e What percentage of students has the same actual and reading age?

4 This diagram shows the percentage test marks of two student groups for the same test.



Use it to determine which of the following statements are true.

a The median is the same for both classes.

b Q_1 for Class Y is 20 marks less than Q_1 for Class X.

c The interquartile range for Class Y is twice the interquartile range of Class X.

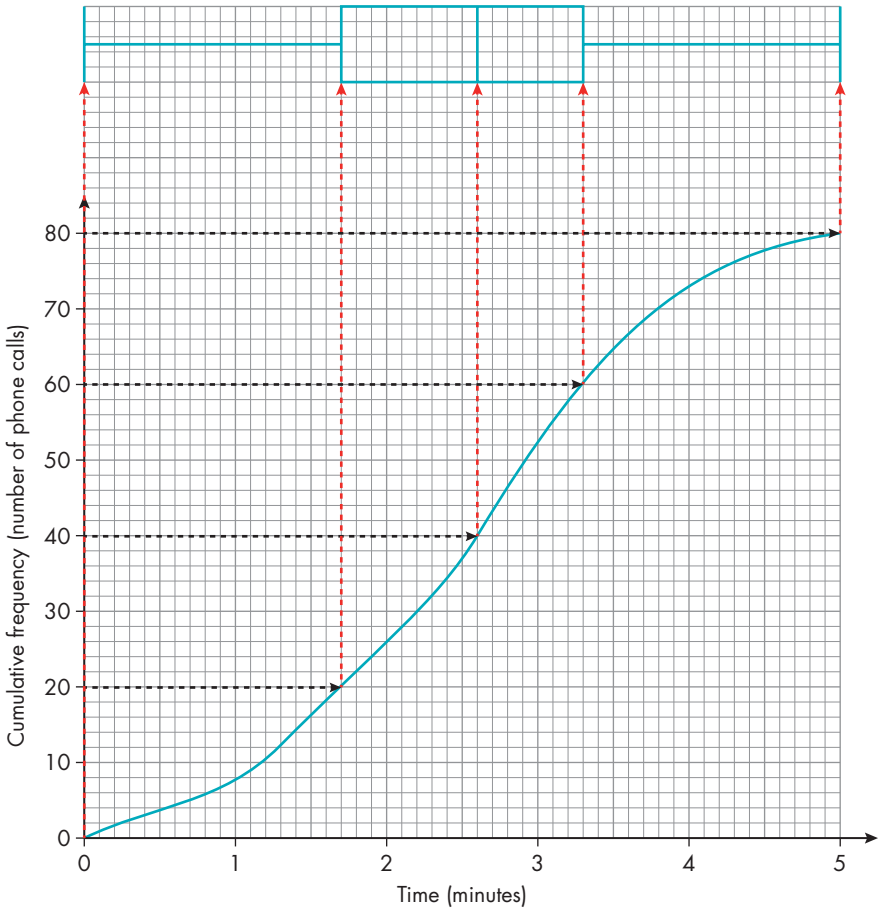
d $P_{75} - P_{50}$ is the same for both classes.

e $P_{25} - P_0$ is the same for both classes.

f The standard deviation for Class Y is less than the standard deviation for Class X.

g The mean mark is the same for both groups.

5 This diagram shows the time spent on 80 mobile phone calls.



Use the diagram to estimate the following statistics.

a The median time spent on mobile phone calls.

b The range of time spent on mobile phone calls.

c $Q_3 - Q_1$.

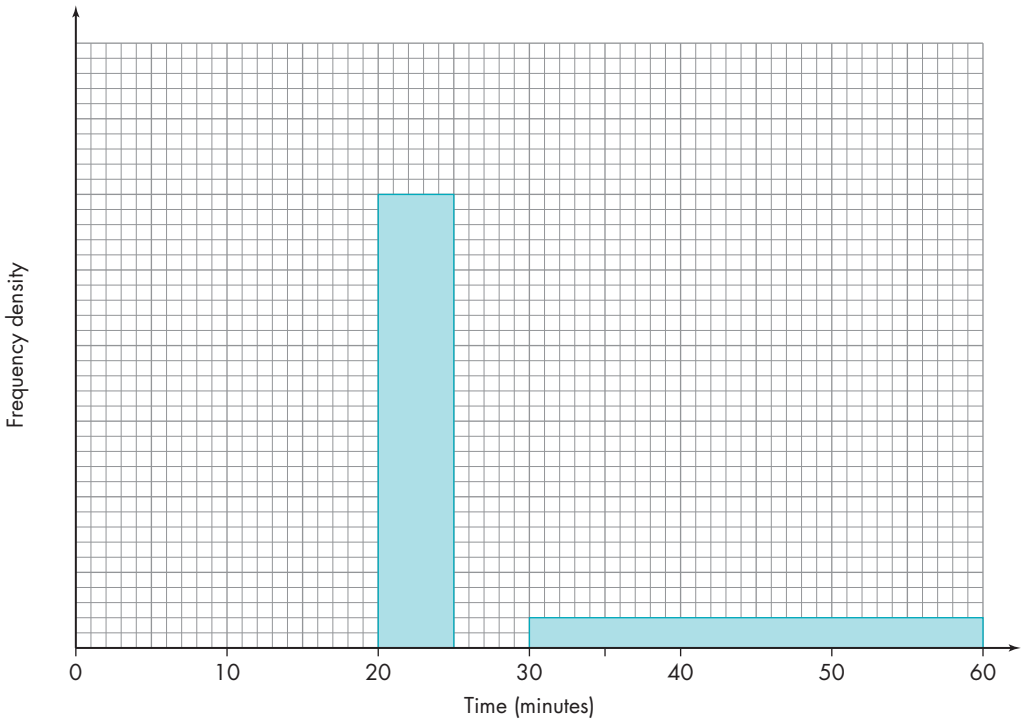
d P_{40}.

e How many calls lasted longer than 2 minutes 12 seconds?

6 Donald kept a record of the time he spent playing *Incandescent Ducks*.

He put the data into a table and a histogram, but Daisy distracted him before he finished.

Copy and complete the table and histogram.



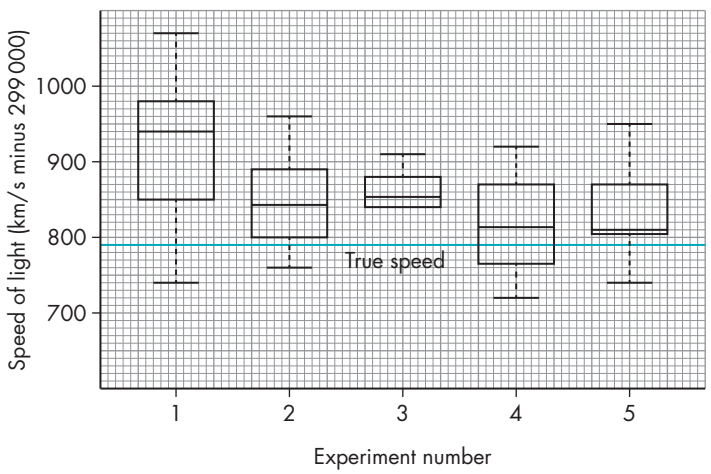
Time in minutes	$0 < t \leqslant 10$	$10 < t \leqslant 20$	$20 < t \leqslant 25$	$25 < t \leqslant 30$	$30 < t \leqslant 60$
Frequency	15	25	15	25	

7 Here are five box plots showing the results of five experiments to determine the speed of light, each experiment consisting of 20 runs.

a Which experiment would you consider to be the most reliable?

b Which would you consider to be the least reliable?

Support your decisions with statistical calculations derived from the box plots.



8 Look at your answers to Exercise 1C, question **4** and Exercise 1E, question **6** about your possible career (or interests) and how you will be using and collecting data to examine something that interests you. Now think about how you would display the data you described. What possible diagrams might you use? Would you decide to go for clarity or novelty in your diagrams? Write a short description of about 150 words, justifying the decisions you make.

9 A market researcher was asked to find what games were played by college students. The researcher conducted a survey of 100 students at two colleges and obtained the following results.

Game	Spy Mouse	Angry Birds	Candy Crush	Cut the Rope	Tiny Death Star
Simmonds	15	25	38	13	9
Thorntons	8	27	17	29	19

Construct an appropriate diagram to show these results, giving reasons why you chose that type of diagram and why you rejected at least one of the other types.

10 A college statistician wanted to compare the reaction times between students in Year 12 and Year 13. The students used Sheep Dash on the Internet, which gave them their average reaction time. The following table shows the results. The average reaction times have been rounded to the nearest hundredth of a second.

Year	12					13				
Times in seconds	0.36	0.87	0.94	0.91	0.15	0.98	0.28	0.28	0.72	0.42
	0.51	0.41	0.73	0.44	0.34	0.31	0.97	0.84	0.62	0.21
	0.99	0.20	0.80	0.95	0.22	0.73	0.30	0.25	0.95	0.70
	0.37	0.94	0.11	0.55	0.86	0.63	0.31	0.20	0.39	0.24

Construct an appropriate diagram to show these results, giving reasons why you chose that type of diagram and why you rejected at least one of the other types.

Case study

Claire has type 1 diabetes. She checks her blood sugar levels before meals so that she can inject the right amount of insulin that her body needs to convert the carbohydrate she eats into energy that her body can use. She also checks her blood sugar levels when she feels it is getting low (hypoglycaemia). Her target is to try and keep her blood sugar level between 3.9 and 7.8 mmol/L (millimoles per litre). If it is below 4 she drinks something that has a high sugar content to bring her level up. If it is above 10 she increases the amount of insulin she takes the next time, or she exercises to help lower the blood sugar level.

Here is the raw data over three days in May. Claire treats this as primary data and recognises it as continuous quantitative data.

Logbook

May 06, 2015 - Jun 02, 2015 (28 days)

12a	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12p	1p	2p	3p	4p	5p	6p	7p	8p	9p	10p	11p	12a
Wednesday, May 6, 2015																								Daily Totals
Glucose (mmol/L)							14.0					3.8			8.1	10.7		6.9						4.3	Average (6) 8.0
Notes																								
Thursday, May 7, 2015																								Daily Totals
Glucose (mmol/L)								7.0				12.3				4.2								4.9	Average (4) 7.1
Notes																								
Friday, May 8, 2015																								Daily Totals
Glucose (mmol/L)								6.3				7.2		4.3				8.8			9.3				Average (5) 7.2
Notes																								

Claire attends a diabetic clinic once or twice a year to talk to a nurse who specialises in diabetes.

Claire takes along a summary of her blood test results. The nurse receives Claire's logbook and treats that as secondary data. The nurse sees that Claire is doing regular checks. Results that are above Claire's target are lightly-shaded; those below target are shaded dark.

Claire and her nurse discuss the data. This time they decide that no major change is needed in Claire's dosage because she runs to college on a Wednesday which explains why her level was low before lunch. She has also had a blood sample taken, which shows that her control over the last six months has been good. Since the blood sample was taken and subjected to a full analysis, the nurse can compare it with Claire's previous results. This use of data over both short and long term means that Claire and her nurse are both confident that she is in very good health and can take care of herself with minimum fuss.

Both the nurse and Claire have become skilled at using statistics in this way to control Claire's diabetes. In the long run it helps Claire to lead a full life with few or no complications, and so cuts down on the cost to the NHS.

Here are Claire's summary results.

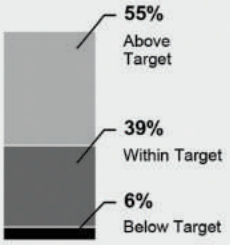
Snapshot

Jan 01, 2015 - Jun 02, 2015 (153 days)

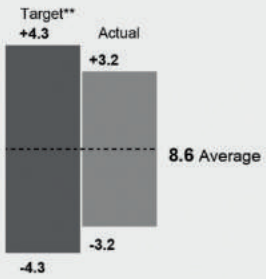
Glucose (mmol/L)

SELECTED REPORTING PERIOD

Tests



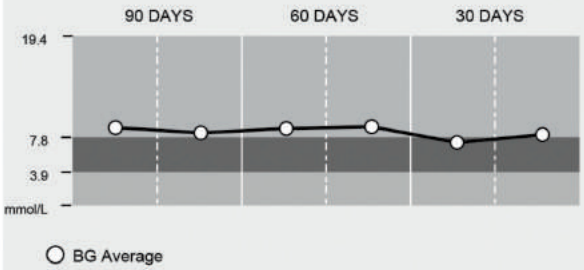
Standard Deviation



Average	8.6	mmol/L
High	18.4	mmol/L
Low	2.3	mmol/L
Standard Deviation	3.2	
Total Tests	725	
Avg. Tests/Day	4.7	
Days without tests	0	
Hypo Events	19	

Target Range: 3.9-7.8 mmol/L ** Target Standard Deviation < (Average/2) Hypo Threshold: 3.3 mmol/L

TRENDS(Last 90 days)



	30 DAYS	90 DAYS
Average	7.6	8.4
High	18.4	18.4
Low	2.5	2.3
Standard Deviation	3.3	3.3
Total Tests	144	418
Avg. Tests/Day	4.8	4.6
Days without tests	0	0
Hypo Events	10	14

By presenting the data in this format, the results are shown numerically as well as in graphs and charts. Claire is pleased that the trend in her blood sugar results shows that she is getting very close to the target zone. She is also very pleased that her standard deviation is well within the target zone for that measure: a guide for this is that it should be less than half the average result.

Since Claire has a reasonable knowledge of statistics she can interpret these results intelligently, helping her to understand her condition. As a result, Claire is able to lead a life that is relatively free of complications.

Project work

To help you bring together all the techniques you have developed from this chapter, choose one of the following projects. They are all concerned with the battle of Trafalgar (21 October 1805).

Project work – 1

The following information comes from a broadsheet published after the Battle of Trafalgar. It shows statistical data of the two fleets that took part in the battle.

THE ENGLISH FLEET consisted of
27 SHIPS OF THE LINE

	GUNS	MEN
VICTORY	110	837
ROYAL SOVEREIGN	110	837
BRITANNIA	100	800
TEMERAIRE	98	738
PRINCE	98	738
TONNANT	84	650
BELLEISLE	74	590
REVENGE	74	590
MARS	74	590
NEPTUNE	98	738
SPARTIATE	74	590
DEFIANCE	74	590
CONQUERER	74	590
DEFENCE	74	590
COLLOSUS	74	590
LEVIATHAN	74	590
ACHILLE	74	590
BELLEROPHON	74	590
MINOTAUR	74	590
ORION	74	590
SWIFTSURE	74	590
POLYPHEMUS	64	500
AFRICA	64	500
AGAMEMNON	64	500
DREADNOUGHT	98	738
AJAX	80	850
THUNDERER	74	590
Total	2,178	17,076

The Combined Enemy Fleet superior in Guns 474 in Men 8,124

THE Combined FLEETS OF FRANCE & SPAIN
33 SHIPS OF THE LINE

		GUNS	MEN	
SANTISUMA TRINIDAD	(S)	140	1200	Taken and Destroyed
BUCENTAURE	(F)	84	800	Taken and Destroyed
RAYO	(S)	100	1000	Taken and Destroyed
PRINCIPE de ASTURIAS	(S)	112	1100	Escaped
INDOMTABLE	(F)	84	800	Destroyed
FOUGOUEX	(F)	74	700	Taken and Destroyed
ACHILLE	(F)	74	700	Blown-up
SANTA ANA	(S)	112	1100	Taken, but got away
MONTANES	(S)	74	700	Escaped
HEROS	(F)	74	700	Escaped
SAN LEANDRO	(S)	64	600	Dismasted, but Escaped
SAN JUSTO	(S)	74	700	Dismasted, but Escaped
SAN ILDEFONSO	(S)	74	700	Taken
LE SWIFTSURE	(F)	74	700	Taken, formerly English
ALGESIRAS	(F)	74	700	Taken, but got away
PLUTON	(F)	74	700	Escaped, much damaged
NEPTUNE	(F)	84	800	Escaped
BAHAMA	(S)	74	700	Taken
SAN NEPOMUCENO	(S)	74	700	Taken
MONARCA	(S)	74	700	Destroyed
SAN FRANCISCO de ASIS	(S)	74	700	Destroyed
ARGONAUTE	(F)	74	700	On shore at Cadiz
LE BERWICK	(F)	74	700	Taken and Destroyed, formerly English
L'AIGLE	(F)	74	700	Taken and Destroyed
INTREPIDE	(F)	74	700	Taken and Burnt
SAN AGUSTIN	(S)	74	700	Taken and Burnt
REDOUBTABLE	(F)	74	700	Taken and Destroyed
ARGONAUTA	(S)	80	800	Taken and Destroyed
FORMIDABLE	(F)	80	800	Escaped } but later captured
MONT-BLANC	(F)	74	700	Escaped } but later captured
SCIPION	(F)	74	700	Escaped } but later captured
DUGUAY-TROUIN	(F)	74	790	Escaped } but later captured
NEPTUNO	(S)	80	800	Destroyed
Total		2,652	25,200	

(F) Denotes French ships (S) Denotes Spanish ships

Four ADMIRALS taken
One killed, and Three wounded.

Imagine that you are a statistician in 1805. Write a report for King George III, comparing the fleets. Your report should contain both numerical and graphical representations.

Project work – 2

What happened to the sailors on board the British ships at Trafalgar?

The following table lists the details of the numbers killed and wounded in each ship.

The fleet was split into two columns, the Weather column, which was north of the Lee column, both sailing eastwards in a line. The ships are listed with the top ship at the front of the column.

Weather column	Killed	Wounded
Victory	57	102
Temeraire	47	76
Neptune	10	34
Leviathan	4	22
Britannia	10	42
Conqueror	3	9
Agamemnon	2	8
Ajax	2	9
Orion	1	23
Minotaur	3	22
Spartiate	3	20
On its own		
Africa	8*	44

Lee column	Killed	Wounded
Royal Sovereign	47	94
Belleisle	33	94*
Mars	29	71*
Tonnant	26	50
Bellerophon	27	127*
Colossus	40	160
Achille	13	59
Dreadnought	7	26
Polyphemus	2	4
Revenge	28	51
Swiftsure	9	8
Defiance	17	53
Thunderer	4	12
Defence	7	29
Prince	0	0

* Indicates that not all sources agree on this figure

Write a report for King George III (who was on the throne in 1805). Use charts and diagrams to represent the casualties on board each ship.

You might like to rank the ships in order of safety and use appropriate averages.

What were the probabilities of being wounded or killed on each ship?

Project work – 3

What happened to the sailors on board the Combined Fleet of France and Spain?

In 1906, Edward Fraser estimated the numbers shown in this table.

Nationality	Killed	Wounded	Prisoners
French	3 373	1155	>4 000
Spanish	1022	1383	3 000–4 000

Write a report, comparing the number killed and wounded with those on the British fleet.

Check your progress

How confident are you feeling in your level of knowledge? What do you need to practise more?

Spec reference	Learning objective			
D1.1	Appreciate the difference between qualitative and quantitative data			
D1.2	Appreciate the difference between primary and secondary data			
D1.3	Collect quantitative and qualitative primary and secondary data			
D2.1	Infer properties of populations or distributions from a sample, whilst knowing the limitations of sampling			
D2.2	Appreciate the strengths and limitations of random, cluster, stratified and quota sampling methods and applying this understanding when designing sampling strategies			
D3.1	Calculate/identify mean, median, mode, quartiles, percentiles, range, interquartile range and standard deviation			
D3.2	Interpret these numerical measures and reach conclusions based on these measures.			
D4.1	Construct and interpret diagrams for grouped discrete data and continuous data, know their appropriate use and reach conclusions based on these diagrams			