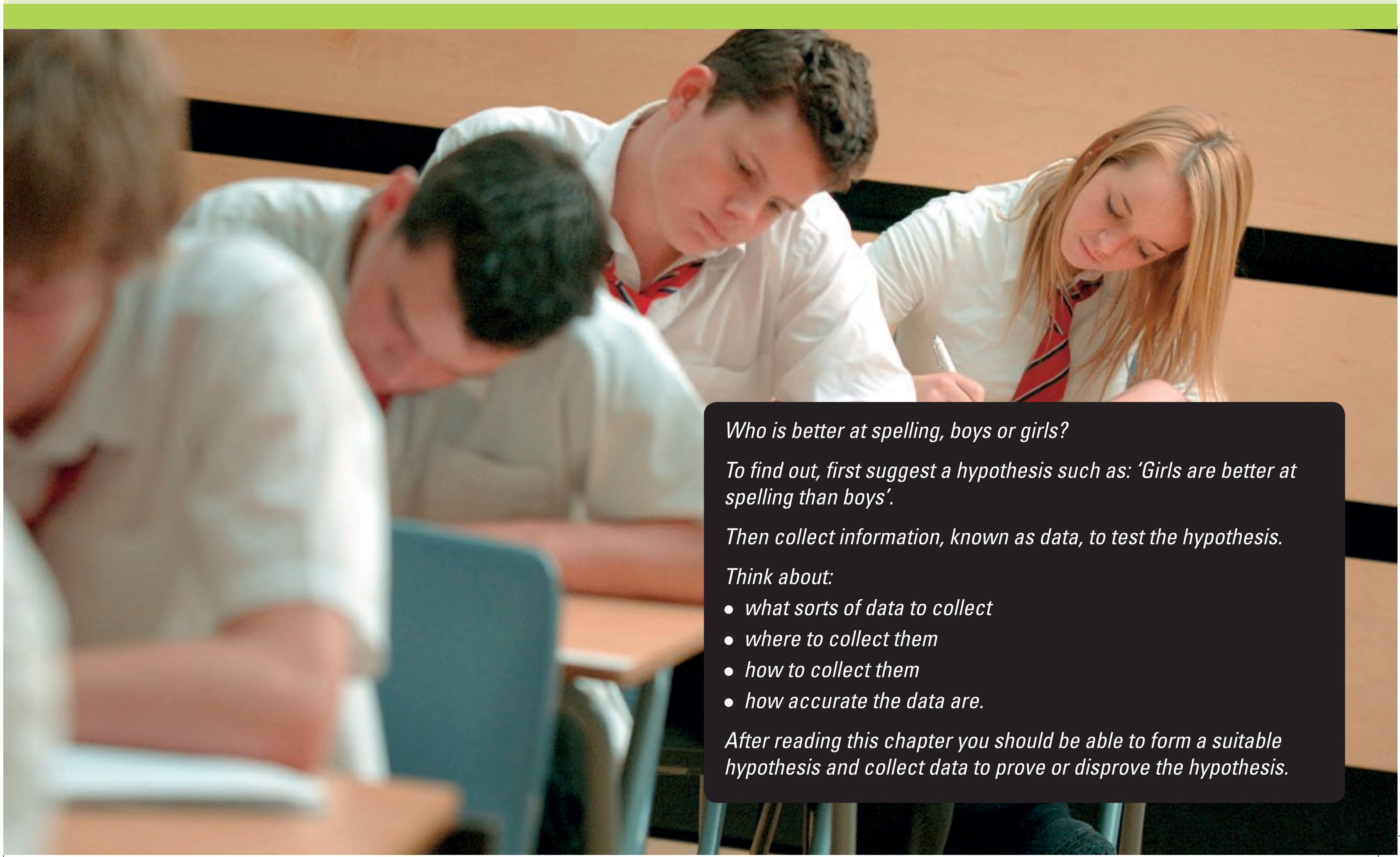
Chapter 1: Collecting data

After completing this chapter you should be able to

- understand what is meant by statistics
- specify a line of enquiry and form a suitable hypothesis
- recognise the difference between quantitative and qualitative variables
- know the difference between discrete and continuous quantitative data and that the measurement of continuous data is subject to some error
- understand the meaning of bivariate data
- know the difference between primary and secondary data
- understand the meaning of the term population
- know what is meant by a census
- understand the reason for sampling and that sampling is used to estimate values in the population
- understand the terms random, randomness and random sample
- understand and use a variety of sampling methods



Who is better at spelling, boys or girls?

To find out, first suggest a hypothesis such as: 'Girls are better at spelling than boys'.

Then collect information, known as data, to test the hypothesis.

Think about:

- *what sorts of data to collect*
- *where to collect them*
- *how to collect them*
- *how accurate the data are.*

After reading this chapter you should be able to form a suitable hypothesis and collect data to prove or disprove the hypothesis.

1.1 The meaning of statistics

Statistics are a way to answer questions using information. The information has to be observed or collected, ordered, represented and then analysed.

The information collected is called raw data. The data collected will depend upon the question you are trying to answer. This question is often called a **hypothesis**.

A **hypothesis** is an assumption made as a starting point for an investigation. It may or may not be true.

'How does the price of a second-hand car change as the car gets older?' The hypothesis to help answer this question could be: 'The price of a second-hand car goes down as it gets older'.

The data needed to answer this question would be the age and price of second-hand cars.

In this chapter you will learn about the different types of data, how to choose samples and how to collect data.

1.2 Types of data

Raw data are obtained by collecting information such as the number of brothers or sisters a person has, or by measuring things such as people's heights.

The things being observed are known as variables because they vary from observation to observation, for example one person's eyes may be blue and the next person's may be brown.

Shoe size, height and eye colour are all variables.

Variable	Observation or measurement
Shoe size	5, 6, 7, 8, 9
Height	169 cm, 178 cm, 183 cm, 185 cm, 179 cm
Eye colour	blue, brown, green, brown, blue

Shoe size and height are numerical measurements (they are written as numbers). They are examples of **quantitative variables**.

Eye colour cannot be written as numbers. It is a quality that people possess. Eye colour is an example of a **qualitative variable**.

Quantitative variables are numerical observations or measurements.

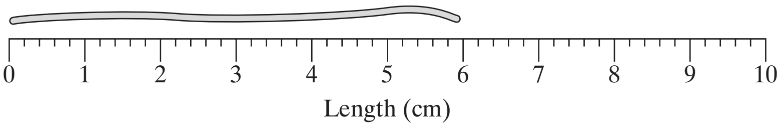
Qualitative variables are non-numerical observations.

ResultsPlus
Watch out!

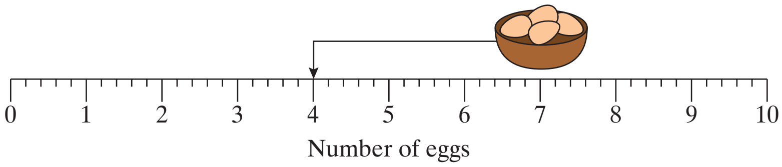
- Some students spell 'quantitative' and 'qualitative' incorrectly in the exam.

Quantitative variables (numerical data) can be **continuous** or **discrete**.

The length of a piece of string could take any value on this scale. It is continuous data.



The number of eggs laid by a chicken can only take particular values. So it is discrete data



Continuous data can take any value on a continuous numerical scale.

Discrete data can only take particular values on a continuous numerical scale.

Weight, length, time, temperature and speed are all examples of continuous data.

Number of sisters, number of doors and shoe size are all examples of discrete data.

Example 1

Mr Jones is going to buy a new car.

He will consider these variables before deciding which car to buy:

miles per gallon	colour	number of passengers	engine size
number of doors	length	type of car	

Write the variables in three lists: qualitative variables, discrete quantitative variables and continuous quantitative variables.

Qualitative: colour, type of car.

Discrete: number of doors, number of passengers.

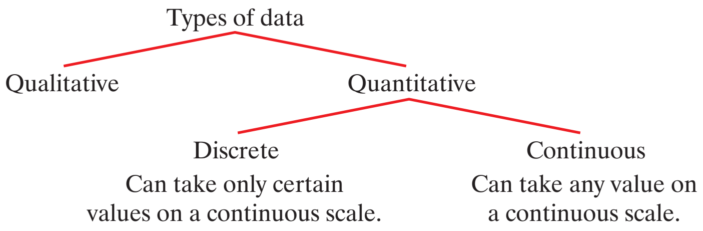
Continuous: miles per gallon, engine size, length.

Ask yourself:

Which cannot be described using numbers?

Which can only take certain numerical values?

Which can take any numerical values?



Can take only certain values on a continuous scale.

Can take any value on a continuous scale.

Edexcel Examiner's Tip

It is important to know the differences between qualitative data, discrete quantitative data and continuous quantitative data because they are represented and analysed in different ways.

1.3 Categorical and ranked data

To make raw data easier to handle and easier to display they may be gathered or ordered in a particular way.

Categorical data and **ranked data** are examples of ordered data.

> A set of data is **categorical** if values or observations belonging to it can be sorted into different categories.

Each piece of categorical data is put into one of a set of non-overlapping categories.

Single-coloured shoes in a cupboard can be sorted according to colour. The characteristic 'colour' can have non-overlapping categories: 'black', 'brown', 'red' and 'other'.

Numerical data can be put into size categories. The characteristic 'engine sizes of cars' could have non-overlapping categories:

$$e \leqslant 1000 \text{ cc}, 1000 < e \leqslant 2000 \text{ cc}, e > 2000 \text{ cc}.$$

> **Ranked data** have values/observations that can be ranked (put in order) or have a rating scale attached. Ranked data can be counted and ordered, but not measured.

The categories for a ranked set of data have a natural order. If judges rank ten dogs on a scale of 1 to 10, 1 would represent the best example of the breed and 10 the worst.

1.4 Bivariate data

In many statistical investigations pairs of variables are examined to find out how they are related or how changes in one variable affects the other variable.

Here are some examples:

- The price and age of second-hand cars.
- The distance and times taken for train journeys.

This is known as **bivariate data**.

> **Bivariate data** are pairs of related variables.

Example 2

Are each of the following categorical data, ranked data or bivariate data?

a Students' year groups.

b The ages and heights of students.

c The league positions of football teams.

a *categorical*

The students are grouped by age, so the category is age.

b *bivariate*

Two sets of data are observed: age and height.

c *ranked*

The teams are ordered according to the number of points they scored.

1.5 Accuracy of continuous data

Continuous data are usually measured and rounded to the nearest sensible unit.

Numbers are usually rounded to the nearest 10, 100, or 1000 depending what is being measured.

The question is whether to round up or down? This is the easy bit!

If the last number is 5 or above, round up.

If the last number is less than 5, round down.

For example, the length of a field would probably be measured to the nearest metre. So, if the exact length is 235.3 m, it would be acceptable to say it is 235 m long.

A measurement given correct to the nearest whole unit can be inaccurate by up to $\pm\frac{1}{2}$ unit.

Age is treated in a slightly different way. A person's age is usually given as their age at their last birthday. So, a person who is 16 years and 11 months old would be called 16 years old. The actual age of a person could be nearly 1 year more than their age at their last birthday, but could not be less.

Example 3

John cycles $4\frac{1}{2}$ miles to the nearest half mile.
What are the longest and shortest distances (upper and lower limits) that John's journey could actually be?

The maximum error will be $\pm\frac{1}{4}$ mile.

The longest distance $= 4\frac{1}{2} + \frac{1}{4} = 4\frac{3}{4}$ miles.

The shortest distance $= 4\frac{1}{2} - \frac{1}{4} = 4\frac{1}{4}$ miles.

John measures to the nearest half mile, so the maximum error will be half of this: $\pm\frac{1}{4}$ mile.

He actually rides any distance between $4\frac{1}{2} \pm \frac{1}{4} = 4\frac{1}{4}$ up to $4\frac{3}{4}$ miles.

Exercise 1A

1 A food processing company thinks that males are the main buyers of their products.

It decides to investigate this.

Write a hypothesis the company could use.

2 A researcher wants to investigate whether Drug A has a better cure rate than Drug B.

Write a hypothesis he could use.

3 Are these discrete data or continuous data?

A The weight of a dog.

B The number of flowers in a bouquet.

C The time it takes to bake a cake.

4 Are these quantitative data or qualitative data?

A The makes of cars.

B The number of acorns on an oak tree.

C The weight of acorns on an oak tree.

5 Julita sold raffle tickets at a village fair.

The tickets were red, green, blue and yellow.

discrete **continuous** **qualitative** **quantitative**

Which of the above can be used to describe:

a the number of tickets sold

b the colour of tickets sold?

6 Which of these could be ranked?

A The marks gained in a test by a group of students.

B The position of dogs in a dog show.

C The colours of sweets.

7 Write down **two** ways in which cars could be categorised.

8 Bivariate data are data that are related in some way.

Choose a word that could be used to make each pair bivariate data.

a Height and ____________ of people.

b Hours of work and ____________.

c Car age and ____________.

9 Here are the ages, in years, of some people in a village.

32	40	17	34	58	60	15	14	22	29
44	18	26	31	36	42	18	23	25	38
31	33	28	47	65	72	19	77	30	34
37	34	58	56	60	63	42	15	82	17
40	61	33	42	21	31	42	72	16	22

Group these ages according to the categories: 0 to <10, 10 to <20, 20 to <30, 30 to <40, 40 to <50, 50 to <60, 60 to <70, 70 to <80, 80 to <90.

10 Karen says that her age is 16 years.

Explain why this response is actually *continuous* data but may appear to be *discrete*.

11 **a** A journey is 54.2 km. What is this to the nearest kilometre?

b The weight of tomatoes picked from one plant is 4.7 kg. What is this to the nearest kilogram?

c A water butt holds 60.5 litres. What is this to the nearest litre?

12 Find the upper and lower limits for:

a a TV that costs £200, correct to the nearest £5

b a car that costs £10 000, correct to the nearest £500

c the height of a wall labelled as 4 m, correct to the nearest metre

d a tree whose height is given as 5 m, correct to the nearest metre

e a railway journey of 125 miles, correct to the nearest mile

f a boy's height of 175 cm, correct to the nearest centimetre

g the age of a man of 63 years

h the distance from London to Manchester, which is given as 200 miles, correct to the nearest 5 miles.

1.6 Populations and sampling

The first thing to do in an investigation is to identify the **population**.

A **population** is everything or everybody that could possibly be involved in an investigation.

Populations can vary in character as these two examples show.

The owner of Highfield Motor Company wants information about the number of miles travelled by the cars in the company garage. The population is all the cars in the garage.

Some people who are waiting in a queue for a shop to open are to be asked questions about their spending habits. The population is all people in the queue.

Census data

Sometimes, the entire population will be sufficiently small to include the entire population in a study. This type of research is called a **census** study because data is gathered on every member of the population.

Census data contains information about every member of the population.

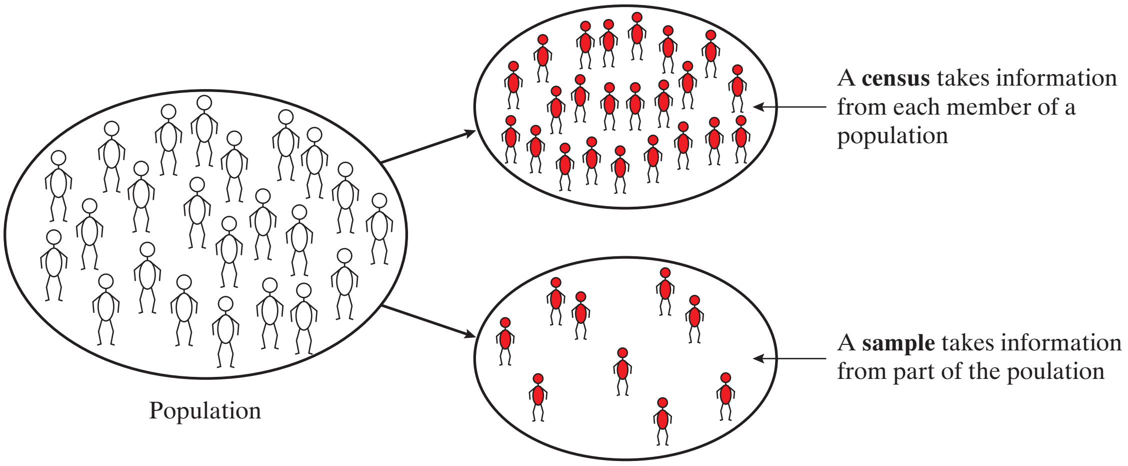
Workers for central and local government, health authorities, police and such organisations often use census data to help allocate resources or plan services for people.

The best known census is the Government's National Census. This takes place every ten years. All people who own or rent a house are identified and asked certain questions, including how many people live in their house.

Sample data

Usually, the population is too large to survey all of its members. A small, but carefully chosen, sample can be used to represent the population. The **sample** must reflect the characteristics of the population from which it is drawn.

A **sample** contains information about part of the population.


Population
A **census** takes information from each member of a population
A **sample** takes information from part of the poulation

	Advantages	Disadvantages
Census	Unbiased Accurate Takes the whole population into account	Time-consuming Expensive Difficult to ensure the whole population is used Lots of data to handle
Sample	Cheaper Less time consuming Less data to be considered	Not completely representative May be biased

The people or items in the population are called **sampling units**.

To take a sample, the population is formed into a list called a **sampling frame**.

The **sampling units** are the people or items that are to be sampled.

The **sampling frame** is a list of the people or items that are to be sampled.

There are two questions to be asked about a sample:

- How big does the sample need to be?
- How is a sample taken so that it represents the population accurately?

Sample size can vary but usually the larger the sample, the more reliable the results. It is important to remember what the sample is being used for and to balance the size of the sample with the accuracy needed. In opinion polls, the views of the whole nation are often found by using a sample of only 1500 people.

1.7 Random sampling

To represent the population accurately, the sample should be taken so that it is free from bias. This simply means that the results are not distorted in any way.

Bias can occur in many ways. Examples are:

- using an unrepresentative sample
- poor or misleading questions
- external factors affecting the data collection
- not correctly identifying the whole population.

A **random sample** is chosen without a conscious decision about which items from the population are selected.

Random sampling methods include simple random sampling and stratified sampling.

Simple random sampling

Simple random sampling is the purest form of sampling. Each sample of size *n* has an equal chance of being selected.

There are many ways of taking a simple random sample. Each item of the sampling frame could be given a number and then the items to be included could be selected by:

- using a random number table
- using a random number generator on a calculator
- using a computer to choose numbers
- putting the numbers in a hat and then selecting however many you need for your sample.

Example 4

This is an extract from a random number table.

33 52 21 17 04 51 78 62 73 41

53 27 15 82 38 59 48 20 82 34

Starting at 33, and working across the rows, use the table to give eight numbers between 1 and 50.

33 21 17 04 41 27 15 38

Start with 33 and take pairs of digits.
Ignore any number larger than 50.
04 counts as the number 4.

Stratified sampling

Edexcel Examiner's Tip
'Stratum' means just one sub-population. 'Strata' is the plural word, it means more than one sub-population.

There may often be factors that divide the population into sub-populations (known as 'strata') such as gender (male/female), age, earnings, etc.

This has to be considered when selecting a sample from the population to ensure it is representative of the population.

To take a stratified sample from each stratum, or sub-population, a simple random sample is taken.

The size of each sample must be in proportion to the relative size of the stratum from which it is taken.

Example 5

The headteacher of a school of 1000 students wants to take a sample of 60 students. Here are the numbers of students in each year.

	Year 7	Year 8	Year 9	Year 10	Year 11
Students	250	250	200	150	150

How many students should be included from each year?

The sample for Years 7 and 8 will be $\frac{250}{1000} \times 60 = 15$.

In Year 7 there are 250 out of the total 1000 students, so take $\frac{250}{1000}$ of 60.

The sample for Year 9 will be $\frac{200}{1000} \times 60 = 12$.

The sample for Years 10 and 11 will be $\frac{150}{1000} \times 60 = 9$.

Calculate other numbers (you do not need to calculate Year 8 because it is the same as Year 7, or Year 11 because it is the same as Year 10).

Check: $15 + 15 + 12 + 9 + 9 = 60$

Check the answer by totalling the answers from all year groups.

1.8 Non-random sampling

These are sampling methods in which the selection of the items is not random.

Methods include cluster sampling, quota sampling and systematic sampling.

Cluster sampling

Cluster sampling is used when the population being sampled splits naturally into groups or clusters.

A sample of the groups is randomly selected and the required information is collected from the sampling units within each selected group.

One version of cluster sampling is area sampling. Clusters consist of geographical areas such as towns.

Quota sampling

Quota sampling is a method of sampling widely used in opinion polling and market research. Each member of the population will have certain characteristics such as age, gender, etc. Instructions are given about the quota (or number) to be sampled from each section of the population who have a particular combination of these characteristics.

An interviewer might be told to interview 20 men over twenty years of age, 20 women over twenty years of age, 15 teenage girls and 15 teenage boys.

Systematic sampling

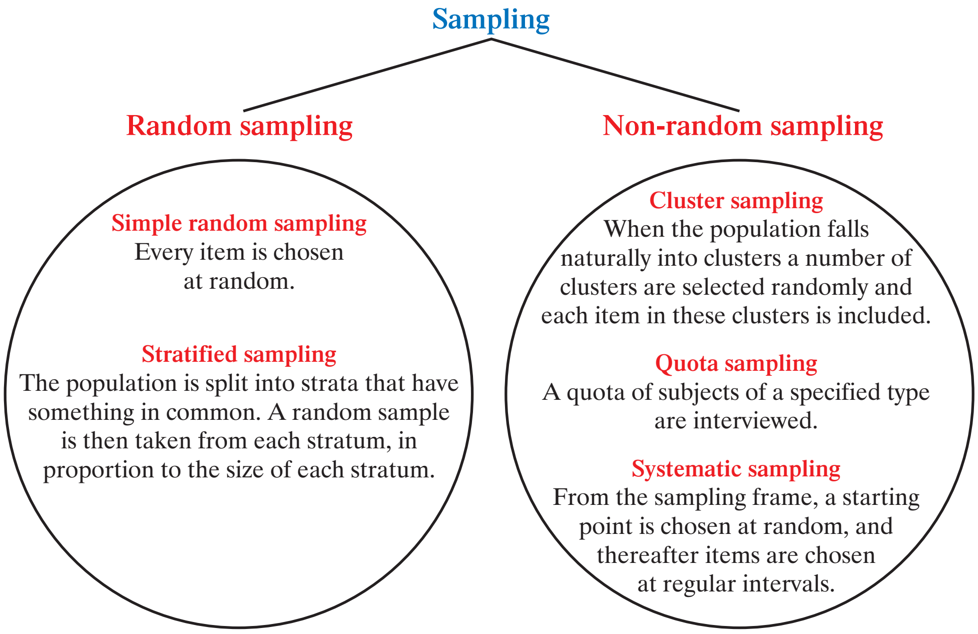
To take a systematic sample, choose a starting point from the sampling frame at random, and then choose items at regular intervals.

This example shows how to sample 20 students from a year group consisting of 100 students.

$\frac{100}{20} = 5$, so every fifth student is chosen after a random starting point between 1 and 5 has been chosen. If the random starting point is 4, then the students selected are numbers 4, 9, 14, 19, …, 94 and 99.

Systematic sampling is used when the population is very large. It is simple to use, but is not always truly representative.

Sampling summary



Sampling

Random sampling

Simple random sampling
Every item is chosen at random.

Stratified sampling
The population is split into strata that have something in common. A random sample is then taken from each stratum, in proportion to the size of each stratum.

Non-random sampling

Cluster sampling
When the population falls naturally into clusters a number of clusters are selected randomly and each item in these clusters is included.

Quota sampling
A quota of subjects of a specified type are interviewed.

Systematic sampling
From the sampling frame, a starting point is chosen at random, and thereafter items are chosen at regular intervals.

Exercise 1B

1 A new canteen is going to open at Edewell College.
The canteen manager want to find out what students would like on the menu. He decides to ask the students.

a Write down the population he should use.

b Describe a sampling unit.

c Give **one** advantage and **one** disadvantage of him using a census.

2 An estate agent wants to get information about house prices in the city where she works.

a What is the population she will use?

b Why would she not use a census of the house prices?

She decides to use a sample. She also decides to use the price of all houses on her list of houses for sale.

c Give a reason why this might be a poor sample.

3 A college decides to investigate the number of hours that students study per week.

a Describe the sampling frame the college will use.

b Describe a sampling unit.

4 **a** Give **two** advantages of using a sample rather than a census.

b Give **two** advantages of using a census rather than a sample.

5 Jack wishes to find out how much people in Britain are prepared to spend on a weekend break. He asked people in his village.

a Identify Jack's population.

b Explain why Jack's sample is likely to be biased.

6 Write the name of the sampling method that is being used in each of these cases.

A Yves needs a sample of 20 people from a numbered list of 100. He generates 20 random numbers and uses those numbered people.

B A factory manager requires a sample of 20 from his work force of 60 men and 40 women. He randomly selects 12 men and eight women.

7 A nursery school has three age groups. The first age group has 60 children, the second has 40 children, and the third has 20 children.

Describe how you would get a sample of 30 children, stratified by age.

ResultsPlus
Build Better Answers

Question: A market research company is going to do a national opinion poll.
They want to find out what people think about the European Union.
The company is going to do a telephone poll.
First they will pick 10 towns at random.
Then they will pick 10 telephone numbers from the telephone book for each town.
They will ring these 100 telephone numbers.
The people who answer will form the sample.
Discuss whether this will form a satisfactory sample for the poll. (2 marks)

🟥 **Zero marks**
Over a quarter of candidates scored no marks for this question. The examiners were looking for two reasons why the sample would not be satisfactory.

● **Good**
For a good answer you need to give one reason, for example point out that the number involved (100 or less) is too small to represent the whole country.

▲ **Excellent**
For an excellent answer you need to give two reasons. A second reason might be that using only 10 towns to represent the country could lead to bias.

8 A university decides to investigate the use of the common room facilities.
It wants to ask a sample of 50 students in total from three year groups.

Year group 1 has 540 males and 420 females.

Year group 2 has 600 males and 660 females.

Year group 3 has 360 males and 420 females.

It decides to use a stratified sample.

a Describe the strata it will use.

b Work out the number of males and females in each stratum that will be used.

c Describe how it will choose the individual members of the strata.

ResultsPlus
Watch out!

🟥 Many students forget that the sample from each stratum should be taken randomly.

9 Here is an extract from a table of random numbers.

335217	045178	627341	532715	823859	482082	342173
451739	936415	526338	127642	137284	463919	394821
264519	143857	012653	628491	558317	316832	229103

a Select 10 random numbers each less than 50. Start at the top left-hand corner and work across from left to right.

b Select 10 random numbers each less than 50. Start at the top left-hand corner and work down in pairs, and from left to right.

10 Tim is asked to take a random sample of 25 students from the registration roll at his school.

He attempts to do so by:

- listing all their names in order
- rolling a dice
- selecting the student shown by the number on the dice (i.e. if the dice shows a 4, he selects the student numbered four in the list)
- rolling the dice again. If it shows a 3, he selects the 4th + 3rd = 7th name on the list, and so on.

Give **two** reasons why Tim's method will not give him a random sample.

11 Write the name of the sampling method that is being used in each of these cases.

A A health centre is interested in which of their facilities are most appreciated by patients. They send a questionnaire to every 20th person on their patient list starting at a random number between 1 and 20.

B A market research company wants some information about the use of parking bays in a supermarket car park. They question 20 people in total from four different age groups of the population.

C A company director wants to know what his workers think about the company pension plan. There are 20 departments in the factory. He asks people in eight of the departments.

1.9 Collecting data

Primary and secondary data

Primary data is collected by, or for, the person who is going to use it.

Secondary data has been collected by someone else.

Examples of **primary data** include:

- measuring hand spans of students in your class
- observing and tallying the colours of all the cars passing your house on a certain morning

Secondary data may be collected from

- websites
- magazines, newspapers
- databases
- research articles

see page XX

Both forms of data collection have advantages and disadvantages.

Data	Advantages	Disadvantages
Primary	You know how the data was obtained. Accuracy is known.	Time-consuming. Expensive.
Secondary	Easy to obtain. Cheap to obtain.	Method of collection unknown. The data might be out of date. May contain mistakes.

Surveys

Primary data may be collected using a **survey**.

A **survey** is the collection of data from a given population. The data is used to analyse a particular issue.

There are a number of methods of collecting primary data in a survey. The main methods are

- questionnaires
- interviews
- observations

Other methods include

- experiments
- data logging

Pilot surveys

A **pilot survey** is conducted on a small sample to test the design and methods of that survey.

A **pilot survey** should identify any problems with questions, such as wording, likely responses, etc.

Questionnaires

Questionnaires are a popular means of collecting data, but are difficult to design. Many rewrites are often needed to produce an acceptable questionnaire.

Sometimes the questions can be factual, such as ‘How old are you?’, or they might ask for an opinion, such as ‘What is your favourite colour?’.

A **questionnaire** is a set of questions designed to obtain data.

Anyone answering a questionnaire is called a respondent.

It is important to follow these rules for writing a questionnaire.

- Keep the questions short, simple and to the point.
- Use words and phrases that are easily understood.
- Avoid leading questions that suggest a certain answer. For example, ‘Don’t you agree that Sudso is the best washing powder?’ is a leading question that is asking for the answer ‘Yes’.
- Write questions that only address a single issue. For example, questions such as ‘Does your car run on diesel fuel?’ should be broken down into two stages. Firstly find out if the respondent has a car and then secondly, if they own a car, find out if it is a diesel car.
- State units required but do not aim for too high a degree of accuracy. Use an interval, such as £10 to £15, 3–5 or to the nearest metre, rather than an exact figure.
- Avoid embarrassing words.

There are two types of question to use in a questionnaire: **open questions** or **closed questions**.

An **open question** is one that has no suggested answers.

An open question could reveal an answer or response not considered by the person asking the question.

The main problem with open questions is that many different answers have to be summarised to enable them to be analysed.

ResultsPlus
Exam Question Report

A town council plans to build a swimming pool.
It is going to carry out a survey to find out what people think of the plan.
The council should carry out a pilot study (pre-test).
(e) Give **two** reasons why.

How students answered

Poor 51%

Most students answered this question poorly. Common incorrect answers were: ‘to check the results from the main survey’; ‘to see if it is worth doing the actual survey’.

Good 31%

Giving one good reason gained one mark. For example: ‘to see if there are any problems with the question wording or response boxes’.

Excellent 18%

Not many students gave excellent answers. Giving two good reasons gained the full 2 marks. A second good reason is: ‘Makes sure the survey is designed and planned to collect the information needed’.

ResultsPlus
Watch out!

● When asked to give an **advantage** of using closed questions, some students incorrectly give an **example** of a closed question and gain no marks.

A **closed question** has a set of answers for the respondent to choose from.

A closed question has the advantage that it is easier to summarise the data.

Closed questions will often use an **opinion scale**.

For example:

Read the following statement and then tick to show whether you strongly agree, agree, disagree or strongly disagree with the statement.

	Strongly agree	Agree	Disagree	Strongly disagree
Statistics is an easy subject				

One problem with opinion scales is that most people will answer somewhere near the middle. They are unlikely to indicate a strong opinion either way as they do not wish to seem extreme.

Sometimes the respondent may be asked to tick a box.

For example:

Tick one box to indicate your age group.

□	□	□	□
Under 20	21 to 40	41 to 60	over 61

Make sure that the possible answers are clear, do not overlap and cover all possibilities.

Example 6

State what is wrong with each of the following questions.

a 'How many brothers do you have?'

□	□	□	□
1	2 to 3	3 to 4	5

b 'You do support the idea of a school uniform, don't you?'

c 'Do you take drugs?' Tick the correct box.

□	□	□
Sometimes	Usually	A lot

d In what age range are you? Tick the correct box.

□	□	□	□
under 10	10 to 20	20 to 30	40 to 60

a There isn't a box for someone who has no brothers or more than five brothers. The option for three brothers appears twice.

b This is forcing the respondent to answer 'Yes'.
There isn't a space for an answer.

c This is a sensitive question and people are unlikely to give a correct answer.
The terms are vague.
There isn't a box for 'Never'.

d 20 can be ticked in two boxes.
There isn't a box for 31 to 39.
There isn't a box for over 60s.

Look for:

- boxes that do not cover all possibilities
- boxes that cover one option more than once
- biased questions that try to persuade you to 'agree'
- questions that people are unlikely to answer honestly
- open questions that allow for personal opinions and do not have tick boxes

Interviews

Interviews may be conducted in different ways, for example:

- Sending questions to people by post or email.
- Calling people on the telephone.
- Face-to-face personal questioning.

Each of these has advantages and disadvantages.

- Postal surveys are cheap but the least likely to provide a response.
- Emails are cheap but they may often be ignored if people think they are spam messages.
- Telephone surveys cost more but the response rate is better than for postal surveys. There is a small possibility that the interviewer may cause bias.
- Personal interviews are the most costly and there is a greater chance of the interviewer causing bias. However, it is more suitable for complex questions where the interviewer can explain the question to the respondent.

Investigations and experiments

Data can be collected from experiments or by direct observation.

For example, flipping a coin many times and seeing how many times it came down heads would identify if the coin is biased.

Here are five possible ways of completing an experiment as part of a statistical investigation.

Before and after experiments

For example, looking at the number of accidents on a stretch of road before and after a 30 miles per hour speed restriction is enforced.

Control groups

A control group is often used to test things such as the effectiveness of drugs. People often feel better because they are receiving attention rather than because of the effect of the drug they are given. The control group and the group to be tested are randomly selected. The control group is then given an inactive substance and the other group is given the actual drug. The effect of the drug can then be assessed by comparing the two groups.

Matched pairs

Two groups of people are used to test the effects of a particular factor. Each individual in one group is paired with an individual in the second group who has everything in common with him/her except the factor being studied. Identical twins can be very important in this type of experiment.

Data logging

This is a mechanical or electronic method of automatically collecting primary data. The instrument is programmed to take readings at set intervals.

Capture–recapture method

This is a method for estimating the size of a population.

Suppose the population is of size N, so that N is the number to be estimated.

The first step is to capture *M* members of the population, mark (or tag) them and release them back into the population.

After waiting some time so that captured members have had time to mix, *n* members of the population are then captured, and the number, *m*, of these that are marked is recorded.

The ratio of the members originally captured and marked, *M*, to the whole population, *N*, is assumed to be the same as the ratio of marked members who were recaptured, *m*, to the total number of members who were recaptured, *n*.

So $\frac{M}{N} = \frac{m}{n}$.

The population size can be found using the formula $N = \frac{n}{Mm}$.

ResultsPlus

Watch out!

- Many students forget the formula $N = \frac{n}{Mm}$.

The following assumptions are made in the capture-recapture method:

i The population has not changed, that is there have been no members entering the population or leaving the population and no births or deaths between the release and recapture times.

ii The probability of being caught is equal for all individuals.

iii Marks (or tags) are not lost and are always recognisable.

Example 7

Twenty birds from a bird colony were captured, ringed and released. Subsequently 100 birds were caught, of which 10 were already ringed. Estimate the number of birds in the colony.

$$\frac{20}{N} = \frac{10}{100}$$

$$N = \frac{20 \times 100}{10} = 200 \text{ birds}$$

Edexcel Examiner's Tip
Use whichever formula you can remember. They both give the same answer.

Replication

Experimental replication means repeating experiments. If similar results are obtained each time, it is reasonable to rely on them more.

Direct observation

Investigations can also be done by direct observation. This involves recording the behaviour patterns of people, objects and events in a systematic manner. For example, recording the number of cars passing in every 10-minute interval.

Example 8

James wishes to see which class in Years 10 and 11 is better at arriving on time in the mornings.

How should he collect suitable data?

He could use class registers to record how many students were late.

Example 9

A creative design company is given a contract to design and market a lifestyle magazine for younger women.

a Explain how they could use both primary and secondary data.

b Explain a possible method of collecting primary data.

a Primary – survey, in the form of a questionnaire, of younger women to find out features of a magazine that they find most and least attractive.

Secondary – refer to other publications to see what they contain.

b By post, email, personal interview or telephone.

Exercise 1C

1 Is the data collected in these examples primary data or secondary data?

A Banji decides to investigate the amount of rainfall his garden gets in one month. He uses a measuring cylinder to collect the rainfall each day.

B A research student decides to investigate the sales of books. He collects data from several websites.

C A student decides to do a project on the milk yield of dairy cows. He gets his data from a local farm's records.

D A council decides to investigate the use of a waste disposal site. A council member goes to the site and collects data by questioning the people using the site.

2 James and Colin wish to predict the winners of the next Football World Cup.

James looks at the World Cup results from 2006, when Italy were the winners.

Colin looks at the table of all World Cup results starting from when the competition began in 1930 to the present day.

a What types of data are they considering?

b Whose opinion would you trust the most and why?

3 Kerry is writing a questionnaire about people's ages.

In it she asks the question 'How old are you?'

Young □ Middle-aged □ Old □

a What is wrong with the question and answers?

b Rewrite the question and answers to improve them.

4 Leslie is carrying out a survey about cricket teams.

He uses this question ‘How often do you watch a cricket match?’

Never □ Once a week □ Whenever I can □

a This is not a good question. Explain why.

b Rewrite the question in a better form.

ResultsPlus
Watch out!

● When asked to comment on a question for a questionnaire, some students don’t state clearly whether the question is unsuitable or not, and lose marks.

5 Which of these questions are open and which are closed?

A Do you like the food? Yes/No

B What do you think about the proposed new town hall?

C How old are you? Under 30 □ 30 to 60 □ Over 60 □

6 A hotel leaves a questionnaire in the hotel rooms for guests to complete.

One of the questions on the questionnaire is ‘Do you agree that this hotel has an excellent dinner menu?’

What is wrong with this question?

7 Write **three** things you need to think about when writing a question for a questionnaire.

8 Write **two** advantages and **two** disadvantages of using an interview to get information.

9 A large chemist company with 170 shops wants to obtain information about sales.

They decide to send out a questionnaire to all shops, but first they carry out a pilot survey.

What are the advantages of doing a pilot survey?

10 You are carrying out a survey to see how much money people will spend buying a car. Give **one** reason why you would choose to conduct a personal interview rather than a postal survey. Also explain why you might not choose to conduct personal interviews.

11 Twenty birds in a large aviary are caught and tagged. They are then returned to the aviary. Later, 40 birds are caught and two are found to have tags. Estimate the number of birds in the aviary.

12 Forty fish in a lake are caught, marked and returned to the lake. A second sample of 100 fish is later caught. Of these 100 fish, 10 are marked.

a Estimate the number of fish in the lake.

b Give **two** assumptions you made before estimating the number of fish in the lake.

13 Describe what is meant by a control group.

Chapter 1 review

1 There are three horses in a field.

★☆☆☆☆ challenge

Use **one** of these words to complete each sentence.

discrete quantitative qualitative continuous cumulative

a The colour of the horses is ____________ data.

b The number of horses is ____________ data.

2 Which of these are continuous data and which are discrete data?

★☆☆☆☆ challenge

A Time

B Number of dogs

C Quantity of milk

3 Which of these are qualitative data and which are quantitative data?

★☆☆☆☆ challenge

A Number of pets

B Height

C Make of car

4 Which of these is primary data and which is secondary data?

★☆☆☆☆ challenge

A Data collected from a car magazine.

B Data from the BBC website.

C Data collected by asking questions of people at a supermarket.

5 A council included this question in a questionnaire:

★☆☆☆☆ challenge

‘Do you agree that the new roundabout is an improvement?’

Describe what is wrong with the question.

6 Give **one** disadvantage of each of these:

★☆☆☆☆ challenge

A using a census

B using a sample.

7 The headteacher in a primary school took a random sample of 10 boys and 12 girls from all the children in the school.

★★☆☆☆ challenge

a What is the sampling frame used by the headteacher?

The headteacher asked each of these 22 children this question as part of a questionnaire:

‘You go to bed before 9 pm, don’t you?’

b Give **one** reason why the headteacher should not have asked the question in this way.

8 A hospital decides to investigate whether they have more women than men visiting their Accident and Emergency Department.

★★☆☆☆ challenge

Write a suitable hypothesis they could use for the investigation.

9 Here are some examples of different types of data.

A The number of people in a café.

B The time it takes to go to work.

C The colour of a dress.

a Which **one** of these is continuous data?

b Which **one** of these is qualitative data?

In a survey it is decided to use secondary data.

c Write **one** advantage and **one** disadvantage of using secondary data.

edexcel *past paper question*

★★☆☆☆ challenge

10 A hotel owner wants to give his guests information about the number of hours of sunshine they can expect in June.

a Write **one** way that he can collect the information if he wants to use primary data.

b Write **one** way that he can collect the information if he wants to use secondary data.

c Is the data he collects in qualitative or quantitative?

edexcel *past paper question*

★★★☆☆ challenge

11 A town council plans to build a swimming pool.

It is going to carry out a survey to find out what people think of the plan.

The council should take a sample rather than a census.

a Give **one** reason why.

b What sampling frame could the council use?

c Describe how the council could take a random sample.

★★★☆☆ challenge

12 There are 90 black cows and 10 brown cows in a herd of 100 cows.

a Write the best word from the list to complete each sentence below.

qualitative continuous quantitative primary

i) The colour of the cows is ____________ data.

ii) The number of cows is ____________ data.

The farmer wants to work out the average amount of milk produced per cow for the whole herd of cows. He will take a 10% stratified sample.

b Write down how could he do this?

edexcel *past paper question*

★★★☆☆ challenge

Edexcel Examiner's Tip

Don't forget to say how you pick within each stratum.

13 In a study about smoking, a doctor selected a sample of adult patients from those registered with him. He looked at their records to see to what degree they claimed to smoke.

a What is the population being studied?

b The doctor wishes to get the number of males and the number of females in proportion to the numbers on his register. What sampling method should he use?

★★★☆☆ challenge

c The doctor's information on smoking was obtained by asking the patients at the time of setting up a database. Give **two** reasons why the data obtained may be unreliable.

edexcel *past paper question*

14 A manufacturer makes two types of ropes – Twineasy and Plasuper.

★★★☆☆ challenge

The manager of the company thinks that Twineasy is the stronger rope. He decides to investigate this.

a Write a suitable hypothesis he could use.

b What would form his population?

c Why would he use a sample to test the hypothesis?

d Describe a sampling unit he will use.

15 The table shows the number of students in each of the four Year 11 maths classes in a school.

★★★☆☆ challenge

Maths class	Number of pupils
Class 1	35
Class 2	25
Class 3	20
Class 4	10

A sample of size 30 is to be taken from Year 11. Omar suggests that three of the classes are chosen at random and 10 students selected at random from each class.

a Would this method give a random sample? Explain your answer.

Nesta suggests a stratified sample of size 36 from the whole of Year 11 using classes as the strata.

b How many students from Class 1 should be in the sample?

16 A city council wishes to know what people think about the plan to build a new ring road.

★★★☆☆ challenge

The council will carry out an opinion poll of residents' views.

a Give **one** reason why the council should not use a census.

b Write down the population from which the sample should be taken.

The council will use a questionnaire.

c It will use closed questions. Give **one** reason why.

One question suggested for the questionnaire was 'You do agree with building a new road, don't you?'

d This is not a good way to find out what people think about the plan to build a new road. Write down **one** reason why.

e Design a suitable question the council could use to find out what people think about the plan to build a new road.

edexcel *past paper question*

★★★☆☆ challenge

17 A market research company is going to carry out a country-wide poll.

They wish to find out people's opinions about the European Union.

a Give **two** reasons why it would not be a good idea to carry out a census.

The company intend to do a telephone poll by randomly picking 10 towns in the country. The company will then randomly select 10 telephone numbers from each town's telephone book. These one hundred telephone numbers will be rung and the people who answer will form the sample.

b Discuss whether you think that this will form a satisfactory sample for the poll.

Two of the questions suggested for asking over the telephone are:

A Do you think that the European Union is working well?

Yes □ No □

B What do you think of the European Union?

c Which question is most suitable? Discuss reasons for your choice.

edexcel *past paper question*

Edexcel Examiner's Tip
Think about open and closed questions.

Chapter 1 summary

Data

1 **Statistics** are a way to answer questions using information. The information has to be observed or collected, ordered, represented and then analysed.

2 A **hypothesis** is an assumption made as a starting point for an investigation. It may or may not be true.

3 **Quantitative variables** are numerical observations or measurements.

Qualitative variables are non-numerical observations.

4 **Continuous data** can take any value on a continuous numerical scale.

Discrete data can only take particular values on a continuous numerical scale.

5 A set of data is **categorical** if values or observations belonging to it can be sorted into different categories.

6 **Ranked data** has values/observations that can be ranked (put in order) or have a rating scale attached. Ranked data can be counted and ordered, but not measured.

7 **Bivariate data** are pairs of related variables.

8 A measurement given correct to the nearest whole unit can be inaccurate by up to ±$\frac{1}{2}$ unit.

Sampling

9 A **population** is everything or everybody that could possibly be involved in an investigation.

10 **Census** data contains information about every member of the population.

11 A **sample** contains information about part of the population.

12 The **sampling units** are the people or items that are to be sampled.

The **sampling frame** is a list of the people or items that are to be sampled.

13 A **random sample** is chosen without a conscious decision about which items from the population are selected.

14 **Primary data** is collected by, or for, the person who is going to use it.

Secondary data has been collected by someone else.

Surveys

15 A **survey** is the collection of data from a given population. The data is used to analyse a particular issue.

16 A **pilot survey** is conducted on a small sample to test the design and methods of that survey.

17 A **questionnaire** is a set of questions designed to obtain data.

18 An **open question** is one that has no suggested answers.

19 A **closed question** has a set of answers for the respondent to choose from.

Test yourself

1 *continuous* *discrete* *quantitative* *qualitative* *primary* *secondary*

Which of the above words can be used to describe the following data?

a Height **b** Colour

c Number of aunts **d** Time

e Census information on a website **f** A tally you make of car types.

2 Give **two** advantages and **two** disadvantages of using:

a primary data **b** secondary data.

3 Describe briefly the meaning of:

a population **b** census.

4 Write **two** advantages of using a sample rather than a census.

5 What is the name given to a sample that allows everyone or everything to have an equal chance of selection?

6 Is this a closed or an open question?

'What do you think about the new hall?'

Give a reason for your answer.

7 Explain **two** ways of selecting a set of random numbers.

8 Write **two** advantages of using a pilot survey.

9 Explain what is meant by a 'control group'.

10 Write the name of the sampling method for each of these.

a Maeve forms her sample by picking boys and girls from her class in proportion to the numbers of boys and girls.

b Jack has a list of 50 students. He uses every 5th student to form a sample.