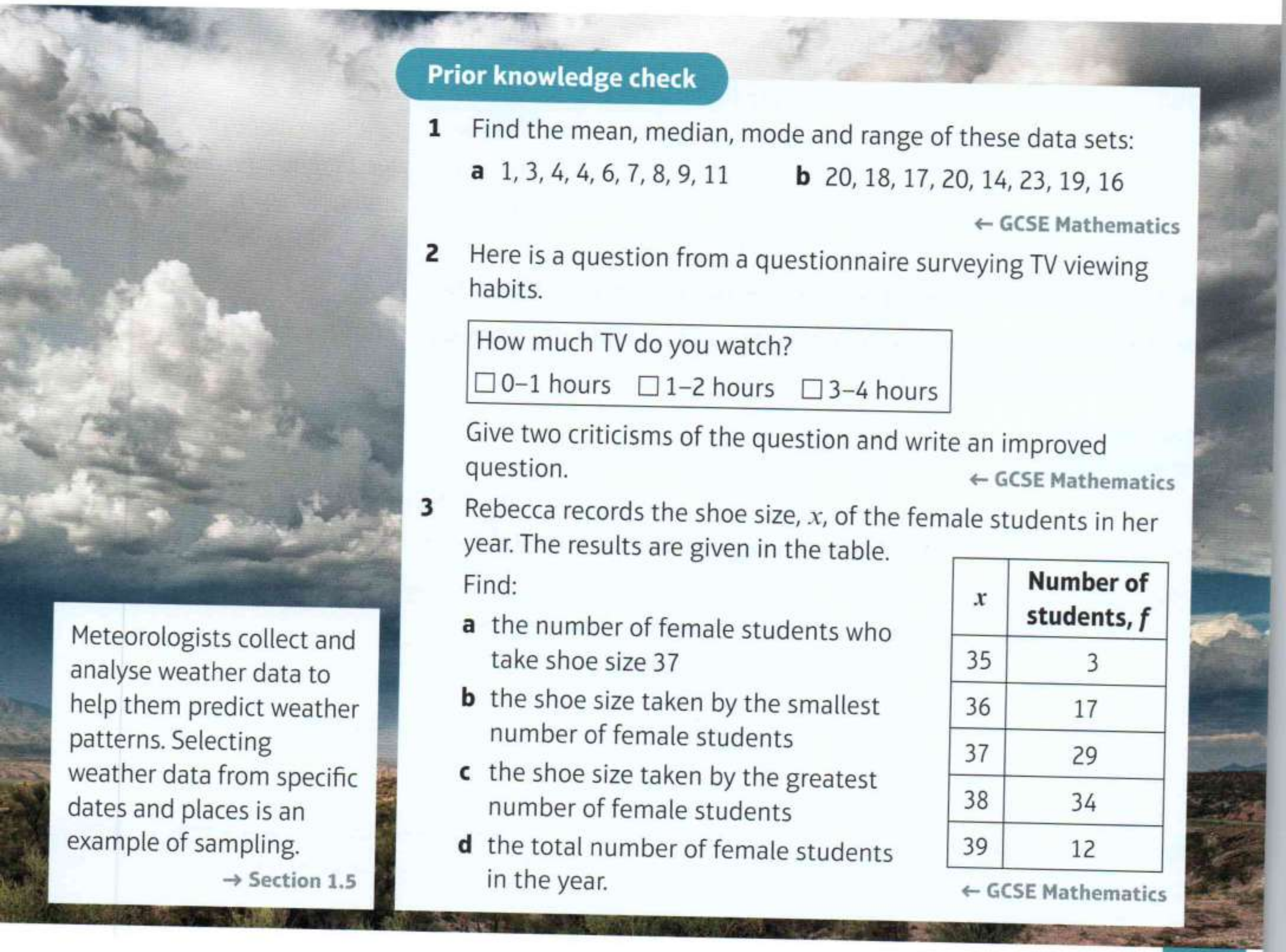
Data collection

1

Objectives

After completing this chapter you should be able to:

- Understand 'population', 'sample' and 'census', and comment on the advantages and disadvantages of each → pages 2–3
- Understand the advantages and disadvantages of simple random sampling, systematic sampling, stratified sampling, quota sampling and opportunity sampling → pages 4–9
- Define qualitative, quantitative, discrete and continuous data, and understand grouped data → pages 9–10
- Understand the large data set and how to collect data from it, identify types of data and calculate simple statistics → pages 11–16



Meteorologists collect and analyse weather data to help them predict weather patterns. Selecting weather data from specific dates and places is an example of sampling.
→ Section 1.5

Prior knowledge check

1 Find the mean, median, mode and range of these data sets:

a 1, 3, 4, 4, 6, 7, 8, 9, 11 **b** 20, 18, 17, 20, 14, 23, 19, 16

← GCSE Mathematics

2 Here is a question from a questionnaire surveying TV viewing habits.

How much TV do you watch?
□ 0–1 hours □ 1–2 hours □ 3–4 hours

Give two criticisms of the question and write an improved question. ← GCSE Mathematics

3 Rebecca records the shoe size, x, of the female students in her year. The results are given in the table.

Find:

a the number of female students who take shoe size 37

b the shoe size taken by the smallest number of female students

c the shoe size taken by the greatest number of female students

d the total number of female students in the year.

x	Number of students, f
35	3
36	17
37	29
38	34
39	12

← GCSE Mathematics

1.1 Populations and samples

- **In statistics, a population is the whole set of items that are of interest.**

For example, the population could be the items manufactured by a factory or all the people in a town. Information can be obtained from a population. Unprocessed information is known as raw data.

- **A census observes or measures every member of a population.**
- **A sample is a selection of observations taken from a subset of the population which is used to find out information about the population as a whole.**

There are a number of advantages and disadvantages of both a census and a sample.

	Advantages	Disadvantages
Census	• It should give a completely accurate result	• Time consuming and expensive • Cannot be used when the testing process destroys the item • Hard to process large quantity of data
Sample	• Less time consuming and expensive than a census • Fewer people have to respond • Less data to process than in a census	• The data may not be as accurate • The sample may not be large enough to give information about small sub-groups of the population

The size of the sample can affect the validity of any conclusions drawn.

- The size of the sample depends on the required accuracy and available resources.
- Generally, the larger the sample, the more accurate it is, but you will need greater resources.
- If the population is very varied, you need a larger sample than if the population were uniform.
- Different samples can lead to different conclusions due to the natural variation in a population.

- **Individual units of a population are known as sampling units.**
- **Often sampling units of a population are individually named or numbered to form a list called a sampling frame.**

Example 1

A supermarket wants to test a delivery of avocados for ripeness by cutting them in half.

a Suggest a reason why the supermarket should not test all the avocados in the delivery.

The supermarket tests a sample of 5 avocados and finds that 4 of them are ripe. They estimate that 80% of the avocados in the delivery are ripe.

b Suggest one way that the supermarket could improve their estimate.

a Testing all the avocados would mean that there were none left to sell.

When testing a product destroys it, a 'census' is not appropriate.

b They could take a larger sample, for example 10 avocados. This would give a better estimate of the overall proportion of ripe avocados.

In general, larger samples produce more accurate predictions about a population.

Exercise 1A

1 A school uses a census to investigate the dietary requirements of its students.

a Explain what is meant by a census.

b Give one advantage and one disadvantage to the school of using a census.

2 A factory makes safety harnesses for climbers and has an order to supply 3000 harnesses. The buyer wishes to know that the load at which the harness breaks exceeds a certain figure.

a Suggest a reason why a census would not be used for this purpose.

The factory tests four harnesses and the load for breaking is recorded:

320 kg   260 kg   240 kg   180 kg

b The factory claims that the harnesses are safe for loads up to 250 kg. Use the sample data to comment on this claim.

c Suggest one way in which the company can improve their prediction.

3 A city council wants to know what people think about its recycling centre. The council decides to carry out a sample survey to learn the opinion of residents.

a Write down one reason why the council should not take a census.

b Suggest a suitable sampling frame.

c Identify the sampling units.

4 A manufacturer of microswitches is testing the reliability of its switches. It uses a special machine to switch them on and off until they break.

a Give one reason why the manufacturer should use a sample rather than a census.

The company tests a sample of 5 switches, and obtains the following results:

23 150   25 071   19 480   22 921   7455

b The company claims that its switches can be operated an average of 20 000 times without breaking. Use the sample data above to comment on this claim.

c Suggest one way the company could improve its prediction.

5 A manager of a garage wants to know what their mechanics think about a new pension scheme designed for them. The manager decides to ask all the mechanics in the garage.

a Describe the population the manager will use.

b Write down the main advantage in asking all of their mechanics.

1.2 Sampling

In random sampling, every member of the population has an equal chance of being selected. The sample should therefore be **representative** of the population. Random sampling also helps to remove **bias** from a sample.

There are three methods of random sampling:

- Simple random sampling
- Systematic sampling
- Stratified sampling

■ **A simple random sample of size n is one where every sample of size n has an equal chance of being selected.**

To carry out a simple random sample, you need a sampling frame, usually a list of people or things. Each person or thing is allocated a unique number and a selection of these numbers is chosen at random.

There are two methods of choosing the numbers: generating random numbers (using a calculator, computer or random number table) and **lottery** sampling.

In lottery sampling, the members of the sampling frame could be written on tickets and placed into a 'hat'. The required number of tickets would then be drawn out.

Example 2

The 100 members of a yacht club are listed alphabetically in the club's membership book.

The committee wants to select a sample of 12 members to fill in a questionnaire.

a Explain how the committee could use a calculator or random number generator to take a simple random sample of the members.

b Explain how the committee could use a lottery sample to take a simple random sample of the members.

a Allocate a number from 1 to 100 to each member of the yacht club. Use your calculator or a random number generator to generate 12 random numbers between 1 and 100.
Go back to the original population and select the people corresponding to these numbers.

b Write all the names of the members on (identical) cards and place them into a hat. Draw out 12 names to make up the sample of members.

If your calculator generates a number that has already been selected, ignore that number and generate an extra random number.

■ **In systematic sampling, the required elements are chosen at regular intervals from an ordered list.**

For example, if a sample of size 20 was required from a population of 100, you would take every fifth person since $100 \div 20 = 5$.

The first person to be chosen should be chosen at random. So, for example, if the first person chosen is number 2 in the list, the remaining sample would be persons 7, 12, 17 etc.

- **In stratified sampling, the population is divided into mutually exclusive strata (males and females, for example) and a random sample is taken from each.**

The proportion of each strata sampled should be the same. A simple formula can be used to calculate the number of people we should sample from each stratum:

$$\text{The number sampled in a stratum} = \frac{\text{number in stratum}}{\text{number in population}} \times \text{overall sample size}$$

Example 3

A factory manager wants to find out what his workers think about the factory canteen facilities.

The manager decides to give a questionnaire to a sample of 80 workers. It is thought that different age groups will have different opinions.

There are 75 workers between ages 18 and 32.

There are 140 workers between 33 and 47.

There are 85 workers between 48 and 62.

a Write down the name of the method of sampling the manager should use.

b Explain how he could use this method to select a sample of workers' opinions.

a Stratified sampling.

b There are: $75 + 140 + 85 = 300$ workers altogether.

18–32: $\frac{75}{300} \times 80 = 20$ workers.

33–47: $\frac{140}{300} \times 80 = 37\frac{1}{3} \approx 37$ workers.

48–62: $\frac{85}{300} \times 80 = 22\frac{2}{3} \approx 23$ workers.

Number the workers in each age group. Use a random number table (or generator) to produce the required quantity of random numbers. Give the questionnaire to the workers corresponding to these numbers.

Find the total number of workers.

For each age group find the number of workers needed for the sample.

Where the required number of workers is not a whole number, round to the nearest whole number.

Each method of random sampling has advantages and disadvantages.

Simple random sampling	
Advantages	**Disadvantages**
• Free of bias • Easy and cheap to implement for small populations and small samples • Each sampling unit has a known and equal chance of selection	• Not suitable when the population size or the sample size is large as it is potentially time consuming, disruptive and expensive. • A sampling frame is needed

Systematic sampling	
Advantages	**Disadvantages**
• Simple and quick to use • Suitable for large samples and large populations	• A sampling frame is needed • It can introduce bias if the sampling frame is not random

Stratified sampling	
Advantages	**Disadvantages**
• Sample accurately reflects the population structure • Guarantees proportional representation of groups within a population	• Population must be clearly classified into distinct strata • Selection within each stratum suffers from the same disadvantages as simple random sampling

Exercise 1B

1 a The head teacher of an infant school wishes to take a stratified sample of 20% of the pupils at the school. The school has the following numbers of pupils.

Year 1	Year 2	Year 3
40	60	80

Work out how many pupils in each age group will be in the sample.

b Describe one benefit to the head teacher of using a stratified sample.

Problem-solving

When describing advantages or disadvantages of a particular sampling method, always refer to the context of the question.

2 A survey is carried out on 100 members of the adult population of a city suburb. The population of the suburb is 2000. An alphabetical list of the inhabitants of the suburb is available.

a Explain one limitation of using a systematic sample in this situation.

b Describe a sampling method that would be free of bias for this survey.

3 A gym wants to take a sample of its members. Each member has a 5-digit membership number, and the gym selects every member with a membership number ending 000.

a Is this a systematic sample? Give a reason for your answer.

b Suggest one way of improving the reliability of this sample.

4 A head of sixth form wants to get the opinion of year 12 and year 13 students about the facilities available in the common room. The table shows the numbers of students in each year.

	Year 12	Year 13
Male	70	50
Female	85	75

a Suggest a suitable sampling method that might be used to take a sample of 40 students.

b How many students from each gender in each of the two years should the head of sixth form ask?

5 A factory manager wants to get information about the ways their workers travel to work. There are 480 workers in the factory, and each has a clocking-in number. The numbers go from 1 to 480. Explain how the manager could take a systematic sample of size 30 from these workers.

6 The director of a sports club wants to take a sample of members. The members each have a unique membership number. There are 121 members who play cricket, 145 members who play hockey and 104 members who play squash. No members play more than one sport.

a Explain how the director could take a simple random sample of 30 members and state one disadvantage of this sampling method.

The director decides to take a stratified sample of 30 members.

b State one advantage of this method of sampling.

c Work out the number of members who play each sport that the director should select for the sample.

1.3 Non-random sampling

There are two types of non-random sampling that you need to know:

- Quota sampling
- Opportunity sampling

■ **In quota sampling, an interviewer or researcher selects a sample that reflects the characteristics of the whole population.**

The population is divided into groups according to a given characteristic. The size of each group determines the proportion of the sample that should have that characteristic.

As an interviewer, you would meet people, assess their group and then, after interview, allocate them into the appropriate quota.

This continues until all quotas have been filled. If a person refuses to be interviewed or the quota into which they fit is full, then you simply ignore them and move on to the next person.

■ **Opportunity sampling consists of taking the sample from people who are available at the time the study is carried out and who fit the criteria you are looking for.**

Notation Opportunity sampling is sometimes called **convenience sampling**.

This could be the first 20 people you meet outside a supermarket on a Monday morning who are carrying shopping bags, for example.

There are advantages and disadvantages of each type of sampling.

Quota sampling	
Advantages	**Disadvantages**
• Allows a small sample to still be representative of the population • No sampling frame required • Quick, easy and inexpensive • Allows for easy comparison between different groups within a population	• Non-random sampling can introduce bias • Population must be divided into groups, which can be costly or inaccurate • Increasing scope of study increases number of groups, which adds time and expense • Non-responses are not recorded as such

Opportunity sampling	
Advantages	**Disadvantages**
• Easy to carry out • Inexpensive	• Unlikely to provide a representative sample • Highly dependent on individual researcher

Exercise 1C

1 Interviewers in a shopping centre collect information on the spending habits from a total of 40 shoppers.

a Explain how they could collect the information using:
i quota sampling **ii** opportunity sampling

b Which method is likely to lead to a more representative sample?

2 Describe the similarities and differences between quota sampling and stratified random sampling.

3 An interviewer asks the first 50 people he sees outside a fish and chip shop on a Friday evening about their eating habits.

a What type of sampling method did he use?

b Explain why the sampling method may not be representative.

c Suggest two improvements he could make to his data collection technique.

4 A researcher is collecting data on the radio-listening habits of people in a local town. She asks the first 5 people she sees on Monday morning entering a supermarket. The number of hours per week each person listens is given below:

4 7 6 8 2

a Use the sample data to work out a prediction for the average number of hours listened per week for the town as a whole.

b Describe the sampling method used and comment on the reliability of the data.

c Suggest two improvements to the method used.

5 In a research study on the masses of wild deer in a particular habitat, scientists catch the first 5 male deer they find and the first 5 female deer they find.

a What type of sampling method are they using?

b Give one advantage of this method.

The masses of the sampled deer are listed below.

Male (kg)	75	80	90	85	82
Female (kg)	67	72	75	68	65

c Use the sample data to compare the masses of male and female wild deer.

d Suggest two improvements the scientists could make to the sampling method.

6 The heights, in metres, of 20 ostriches are listed below:

1.8, 1.9, 2.3, 1.7, 2.1, 2.0, 2.5, 2.7, 2.5, 2.6, 2.3, 2.2, 2.4, 2.3, 2.2, 2.5, 1.9, 2.0, 2.2, 2.5

a Take an opportunity sample of size five from the data.

b Starting from the second data value, take a systematic sample of size five from the data.

c Calculate the mean height for each sample.

d State, with reasons, which sampling method is likely to be more reliable.

Hint An example of an opportunity sample from this data would be to select the first five heights from the list.

1.4 Types of data

- **Variables or data associated with numerical observations are called quantitative variables or quantitative data.**

For example, you can give a number to shoe size so shoe size is a quantitative variable.

- **Variables or data associated with non-numerical observations are called qualitative variables or qualitative data.**

For example, you can't give a number to hair colour (blonde, red, brunette). Hair colour is a qualitative variable.

- **A variable that can take any value in a given range is a continuous variable.**

For example, time can take any value, e.g. 2 seconds, 2.1 seconds, 2.01 seconds etc.

- **A variable that can take only specific values in a given range is a discrete variable.**

For example, the number of girls in a family is a discrete variable as you can't have 2.65 girls in a family.

Large amounts of data can be displayed in a frequency table or as grouped data.

- **When data is presented in a grouped frequency table, the specific data values are not shown. The groups are more commonly known as classes.**
  - **Class boundaries tell you the maximum and minimum values that belong in each class.**
  - **The midpoint is the average of the class boundaries.**
  - **The class width is the difference between the upper and lower class boundaries.**

Example 4

The lengths, x mm, to the nearest mm, of the forewings of a random sample of male adult butterflies are measured and shown in the table.

Length of forewing (mm)	Number of butterflies, f
30–31	2
32–33	25
34–36	30
37–39	13

a State whether length is

i quantitative or qualitative

ii discrete or continuous.

b Write down the class boundaries, midpoint and class width for the class 34–36.

a i Quantitative

ii Continuous

b Class boundaries 33.5 mm, 36.5 mm

Midpoint $= \frac{1}{2}(33.5 + 36.5) = 35$ mm

Class width $= 36.5 - 33.5 = 3$ mm

Watch out Be careful when finding class boundaries for continuous data. The data values have been rounded to the nearest mm, so the upper class boundary for the 30–31 mm class is 31.5 mm.

Exercise 1D

1 State whether each of the following variables is qualitative or quantitative.

a Height of a tree

b Colour of car

c Time waiting in a queue

d Shoe size

e Names of pupils in a class

2 State whether each of the following quantitative variables is continuous or discrete.

a Shoe size

b Length of leaf

c Number of people on a bus

d Weight of sugar

e Time required to run 100 m

f Lifetime in hours of torch batteries

3 Explain why:

a 'Type of tree' is a qualitative variable

b 'The number of pupils in a class' is a discrete quantitative variable

c 'The weight of a collie dog' is a continuous quantitative variable.

4 The distribution of the masses of two-month-old lambs is shown in the grouped frequency table.

Mass, m (kg)	Frequency
$1.2 \leqslant m < 1.3$	8
$1.3 \leqslant m < 1.4$	28
$1.4 \leqslant m < 1.5$	32
$1.5 \leqslant m < 1.6$	22

Hint The class boundaries are given using inequalities, so the values given in the table are the actual class boundaries.

a Write down the class boundaries for the third group.

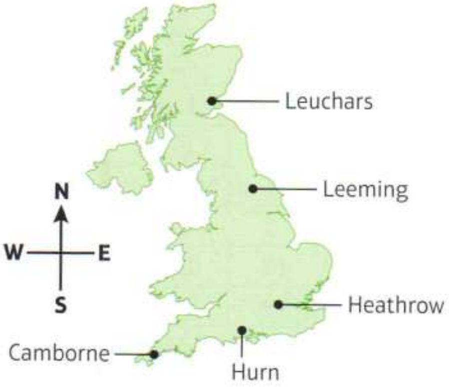
b Work out the midpoint of the second group.

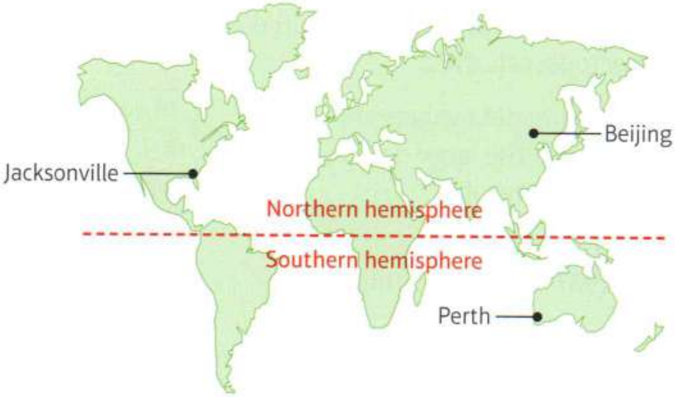
c Work out the class width of the first group.

1.5 The large data set

You will need to answer questions based on real data in your exam. Some of these questions will be based on weather data from the **large data set** provided by Edexcel.

The data set consists of weather data samples provided by the Met Office for five UK weather stations and three overseas weather stations over two set periods of time: May to October 1987 and May to October 2015. The weather stations are labelled on the maps below.





The large data set contains data for a number of different variables at each weather station:

- **Daily mean temperature** in °C – this is the average of the hourly temperature readings during a 24-hour period.
- **Daily total rainfall** including solid precipitation such as snow and hail, which is melted before being included in any measurements – amounts less than 0.05 mm are recorded as 'tr' or 'trace'
- **Daily total sunshine** recorded to the nearest tenth of an hour
- **Daily mean wind direction and windspeed** in knots, averaged over 24 hours from midnight to midnight. Mean wind directions are given as bearings and as cardinal (compass) directions. The data for mean windspeed is also categorised according to the **Beaufort scale**

Beaufort scale	Descriptive term	Average speed at 10 metres above ground
0	Calm	Less than 1 knot
1–3	Light	1 to 10 knots
4	Moderate	11 to 16 knots
5	Fresh	17 to 21 knots

Notation A **knot** (kn) is a 'nautical mile per hour'. 1 kn = 1.15 mph.

- **Daily maximum gust** in knots – this is the highest instantaneous windspeed recorded. The direction from which the maximum gust was blowing is also recorded
- **Daily maximum relative humidity**, given as a percentage of air saturation with water vapour. Relative humidities above 95% give rise to misty and foggy conditions

Watch out For the overseas locations, the only data recorded are:

- Daily mean temperature
- Daily total rainfall
- Daily mean pressure
- Daily mean windspeed

- **Daily mean cloud cover** measured in 'oktas' or eighths of the sky covered by cloud
- **Daily mean visibility** measured in decametres (Dm). This is the greatest horizontal distance at which an object can be seen in daylight
- **Daily mean pressure** measured in hectopascals (hPa)

Any missing data values are indicated in the large data set as n/a or 'not available'.

Data from Hurn for the first days of June 1987 is shown to the right.

You are expected to be able to take a sample from the large data set, identify different types of data and calculate statistics from the data.

- **If you need to do calculations on the large data set in your exam, the relevant extract from the data set will be provided.**
- **You need to be familiar with the types and ranges of data in the large data set, and with the characteristics of each location. You may need to recall trends from within the data set, or identify a location based on given data.**

HURN	© Crown Copyright Met Office 1987					
Date	Daily mean temperature (°C)	Daily total rainfall (mm)	Daily total sunshine (hrs)	Daily mean windspeed (kn)	Daily mean windspeed (Beaufort conversion)	Daily maximum gust (kn)
01/6/1987	15.1	0.6	4.5	7	Light	19
02/6/1987	12.5	4.7	0	7	Light	22
03/6/1987	13.8	tr	5.6	11	Moderate	25
04/6/1987	15.5	5.3	7.8	7	Light	17
05/6/1987	13.1	19.0	0.5	10	Light	33
06/6/1987	13.8	0	8.9	19	Fresh	46
07/6/1987	13.2	tr	3.8	11	Moderate	27
08/6/1987	12.9	1	1.7	9	Light	19
09/6/1987	11.2	tr	5.4	6	Light	19
10/6/1987	9.2	1.3	9.7	4	Light	n/a
11/6/1987	12.6	0	12.5	6	Light	18
12/6/1987	10.4	0	11.9	5	Light	n/a
13/6/1987	9.6	0	8.6	5	Light	15
14/6/1987	10.2	0	13.1	5	Light	18
15/6/1987	9.2	3.7	7.1	4	Light	25
16/6/1987	10.4	5.6	8.3	6	Light	25
17/6/1987	12.8	0.1	5.3	10	Light	27
18/6/1987	13.0	7.4	3.2	9	Light	24
19/6/1987	14.0	tr	0.4	12	Moderate	33
20/6/1987	12.6	0	7.7	6	Light	17

Example 5

Look at the extract from the large data set given above.

a Describe the type of data represented by daily total rainfall.

Alison is investigating daily maximum gust. She wants to select a sample of size 5 from the first 20 days in Hurn in June 1987. She uses the first two digits of the date as a sampling frame and generates five random numbers between 1 and 20.

b State the type of sample selected by Alison.

c Explain why Alison's process might not generate a sample of size 5.

a Continuous quantitative data.

b Simple random sample

c Some of the data values are not available (n/a).

Watch out Although you won't need to recall specific data values from the large data set in your exam, you will need to know the limitations of the data set and the approximate range of values for each variable.

Example 6

Using the extract from the large data set on the previous page, calculate:

a the mean daily mean temperature for the first five days of June in Hurn in 1987

b the median daily total rainfall for the week of 14th June to 20th June inclusive.

The median daily total rainfall for the same week in Perth was 19.0 mm. Karl states that more southerly countries experience higher rainfall during June.

c State with a reason whether your answer to part **b** supports this statement.

a $15.1 + 12.5 + 13.8 + 15.5 + 13.1 = 70.0$

$70.0 \div 5 = 14.0$ °C (1 d.p.)

b The values are: 0, 3.7, 5.6, 0.1, 7.4, tr, 0

In ascending order: 0, 0, tr, 0.1, 3.7, 5.6, 7.4

The median is the middle value so 0.1 mm.

c Perth is in Australia, which is south of the UK, and the median rainfall was higher (19.0 mm > 0.1 mm). However, this is a very small sample from a single location in each country so does not provide enough evidence to support Karl's statement.

The mean is the sum of the data values divided by the number of data values. The data values are given to 1 d.p. so give your answer to the same degree of accuracy.

Trace amounts are slightly larger than 0. If you need to do a numerical calculation involving a trace amount you can treat it as 0.

Online Use your calculator to find the mean and median of discrete data.

Problem-solving Don't just look at the numerical values. You also need to consider whether the sample is large enough, and whether there are other geographical factors which could affect rainfall in these two locations.

Exercise 1E

1 From the eight weather stations featured in the large data set, write down:

a the station which is furthest north

b the station which is furthest south

c an inland station

d a coastal station

e an overseas station.

2 Explain, with reasons, whether daily maximum relative humidity is a discrete or continuous variable.

Questions 3 and 4 in this exercise use the following extracts from the large data set.

LEEMING

© Crown Copyright Met Office 2015

Date	Daily mean temperature (°C)	Daily total rainfall (mm)	Daily total sunshine (hrs)	Daily mean windspeed (kn)
01/06/2015	8.9	10	5.1	15
02/06/2015	10.7	tr	8.9	17
03/06/2015	12.0	0	10.0	8
04/06/2015	11.7	0	12.8	7
05/06/2015	15.0	0	8.9	9
06/06/2015	11.6	tr	5.4	17
07/06/2015	12.6	0	13.9	10
08/06/2015	9.4	0	9.7	7
09/06/2015	9.7	0	12.1	5
10/06/2015	11.0	0	14.6	4

HEATHROW

© Crown Copyright Met Office 2015

Date	Daily mean temperature (°C)	Daily total rainfall (mm)	Daily total sunshine (hrs)	Daily mean windspeed (kn)
01/06/2015	12.1	0.6	4.1	15
02/06/2015	15.4	tr	1.6	18
03/06/2015	15.8	0	9.1	9
04/06/2015	16.1	0.8	14.4	6
05/06/2015	19.6	tr	5.3	9
06/06/2015	14.5	0	12.3	12
07/06/2015	14.0	0	13.1	5
08/06/2015	14.0	tr	6.4	7
09/06/2015	11.4	0	2.5	10
10/06/2015	14.3	0	7.2	10

Ⓟ **3 a** Work out the mean of the daily total sunshine for the first 10 days of June 2015 in:

i Leeming

ii Heathrow.

b Work out the range of the daily total sunshine for the first 10 days of June 2015 in:

i Leeming

ii Heathrow.

c Supraj says that the further north you are, the fewer the number of hours of sunshine. State, with reasons, whether your answers to parts **a** and **b** support this conclusion.

Hint State in your answer whether Leeming is north or south of Heathrow.

Ⓟ **4** Calculate the mean daily total rainfall in Heathrow for the first 10 days of June 2015. Explain clearly how you dealt with the data for 2/6/2015, 5/6/2015 and 8/6/2015.

Ⓟ **5** Dominic is interested in seeing how the average monthly temperature changed over the summer months of 2015 in Jacksonville. He decides to take a sample of two days every month and average the temperatures before comparing them.

a Give one reason why taking two days a month might be:

i a good sample size

ii a poor sample size.

b He chooses the first day of each month and the last day of each month. Give a reason why this method of choosing days might not be representative.

c Suggest a better way that he can choose his sample of days.

Ⓟ **6** The table shows the mean daily temperatures at each of the eight weather stations for August 2015:

	Camborne	**Heathrow**	**Hurn**	**Leeming**	**Leuchars**	**Beijing**	**Jacksonville**	**Perth**
Mean daily mean temp (°C)	15.4	18.1	16.2	15.6	14.7	26.6	26.4	13.6

© Crown Copyright Met Office

a Give a geographical reason why the temperature in August might be lower in Perth than in Jacksonville.

b Comment on whether this data supports the conclusion that coastal locations experience lower average temperatures than inland locations.

Ⓟ **7** Brian calculates the mean cloud coverage in Leeming in September 1987. He obtains the answer 9.3 oktas. Explain how you know that Brian's answer is incorrect.

Ⓟ **8** The large data set provides data for 184 consecutive days in 1987. Marie is investigating daily mean windspeeds in Camborne in 1987.

a Describe how Marie could take a systematic sample of 30 days from the data for Camborne in 1987. **(3 marks)**

b Explain why Marie's sample would not necessarily give her 30 data points for her investigation. **(1 mark)**

Large data set

You will need access to the large data set and spreadsheet software to answer these questions.

1 a Find the mean daily mean pressure in Beijing in October 1987.

b Find the median daily rainfall in Jacksonville in July 2015.

c i Draw a grouped frequency table for the daily mean temperature in Heathrow in July and August 2015. Use intervals $10 \leqslant t < 15$, etc.

ii Draw a histogram to display this data.

iii Draw a frequency polygon for this data.

Hint You can use the **COUNTIFS** command in a spreadsheet to work out the frequency for each class.

2 a i Take a simple random sample of size 10 from the data for daily mean windspeed in Leeming in 1987.

ii Work out the mean of the daily windspeeds using your sample.

b i Take a sample of the last 10 values from the data for daily mean windspeed in Leuchars in 1987.

ii Work out the mean of the daily mean windspeeds using your sample.

c State, with reasons, which of your samples is likely to be more representative.

d Suggest two improvements to the sampling methods suggested in part **a**.

e Use an appropriate sampling method and sample size to estimate the mean windspeeds in Leeming and Leuchars in 1987. State with a reason whether your calculations support the statement 'Coastal locations are likely to have higher average windspeeds than inland locations'.

Mixed exercise 1

1 The table shows the daily mean temperature recorded on the first 15 days in May 1987 at Heathrow.

Day of month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Daily mean temp (°C)	14.6	8.8	7.2	7.3	10.1	11.9	12.2	12.1	15.2	11.1	10.6	12.7	8.9	10.0	9.5

© Crown Copyright Met Office

a Use an opportunity sample of the first 5 dates in the table to estimate the mean daily mean temperature at Heathrow for the first 15 days of May 1987.

b Describe how you could use the random number function on your calculator to select a simple random sample of 5 dates from this data.

Hint Make sure you describe your sampling frame.

c Use a simple random sample of 5 dates to estimate the mean daily mean temperature at Heathrow for the first 15 days of May 1987.

d Use all 15 dates to calculate the mean daily mean temperature at Heathrow for the first 15 days of May 1987. Comment on the reliability of your two samples.

2 a Give one advantage and one disadvantage of using:

i a census **ii** a sample survey.

b It is decided to take a sample of 100 from a population consisting of 500 elements. Explain how you would obtain a simple random sample from this population.

3 **a** Explain briefly what is meant by:
i a population **ii** a sampling frame.

b A market research organisation wants to take a sample of:
i owners of diesel motor cars in the UK
ii persons living in Oxford who suffered injuries to the back during July 1996.

Suggest a suitable sampling frame in each case.

4 Write down one advantage and one disadvantage of using:

a stratified sampling **b** simple random sampling.

5 The managing director of a factory wants to know what the workers think about the factory canteen facilities. 100 people work in the offices and 200 work on the shop floor.

The factory manager decides to ask the people who work in the offices.

a Suggest a reason why this is likely to produce a biased sample.

b Explain briefly how the factory manager could select a sample of 30 workers using:
i systematic sampling **ii** stratified sampling **iii** quota sampling.

6 There are 64 girls and 56 boys in a school.

Explain briefly how you could take a random sample of 15 pupils using:

a simple random sampling **b** stratified sampling.

7 As part of her statistics project, Deepa decided to estimate the amount of time A-level students at her school spent on private study each week. She took a random sample of students from those studying arts subjects, science subjects and a mixture of arts and science subjects. Each student kept a record of the time they spent on private study during the third week of term.

a Write down the name of the sampling method used by Deepa.

b Give a reason for using this method and give one advantage this method has over simple random sampling.

8 A conservationist is collecting data on African springboks. She catches the first five springboks she finds and records their masses.

a State the sampling method used.

b Give one advantage of this type of sampling method.

The data is given below:

70 kg 76 kg 82 kg 74 kg 78 kg.

c State, with a reason, whether this data is discrete or continuous.

d Calculate the mean mass.

A second conservationist collects data by selecting one springbok in each of five locations. The data collected is given below:

79 kg 86 kg 90 kg 68 kg 75 kg.

e Calculate the mean mass for this sample.

f State, with a reason, which mean mass is likely to be a more reliable estimate of the mean mass of African springboks.

g Give one improvement the second conservationist could make to the sampling method.

Ⓔ **9** Data on the daily total rainfall in Beijing during 2015 is gathered from the large data set. The daily total rainfall (in mm) on the first of each month is listed below:

May 1st	9.0
June 1st	0.0
July 1st	1.0
August 1st	32.0
September 1st	4.1
October 1st	3.0

a State, with a reason, whether or not this sample is random. **(1 mark)**

b Suggest two alternative sampling methods and give one advantage and one disadvantage of each in this context. **(2 marks)**

c State, with a reason, whether the data is discrete or continuous. **(1 mark)**

d Calculate the mean of the six data values given above. **(1 mark)**

e Comment on the reliability of this value as an estimate for the mean daily total rainfall in Beijing during 2015. **(1 mark)**

Large data set

You will need access to the large data set and spreadsheet software to answer these questions.

a Take a systematic sample of size 18 for the daily maximum relative humidity in Camborne during 1987.

b Give one advantage of using a systematic sample in this context.

c Use your sample to find an estimate for the mean daily maximum relative humidity in Camborne during 1987.

d Comment on the reliability of this estimate. Suggest one way in which the reliability can be improved.

Summary of key points

1 - In statistics, a **population** is the whole set of items that are of interest.
  - A **census** observes or measures every member of a population.

2 - A sample is a selection of observations taken from a subset of the population which is used to find out information about the population as a whole.
  - Individual units of a population are known as **sampling units**.
  - Often sampling units of a population are individually named or numbered to form a list called a **sampling frame**.

3 - A **simple random sample** of size n is one where every sample of size n has an equal chance of being selected.
  - In **systematic sampling**, the required elements are chosen at regular intervals from an ordered list.
  - In **stratified sampling**, the population is divided into mutually exclusive strata (males and females, for example) and a random sample is taken from each.
  - In **quota sampling**, an interviewer or researcher selects a sample that reflects the characteristics of the whole population.
  - **Opportunity sampling** consists of taking the sample from people who are available at the time the study is carried out and who fit the criteria you are looking for.

4 - Variables or data associated with numerical observations are called **quantitative variables** or **quantitative data**.
  - Variables or data associated with non-numerical observations are called **qualitative variables** or **qualitative data**.

5 - A variable that can take any value in a given range is a **continuous variable**.
  - A variable that can take only specific values in a given range is a **discrete variable**.

6 - When data is presented in a grouped frequency table, the specific data values are not shown. The groups are more commonly known as **classes**.
  - Class boundaries tell you the maximum and minimum values that belong in each class.
  - The midpoint is the average of the class boundaries.
  - The class width is the difference between the upper and lower class boundaries.

7 If you need to do calculations on the large data set in your exam, the relevant extract from the data set will be provided.

8 You need to be familiar with the types and ranges of data in the large data set, and with the characteristics of each location. You may need to recall trends from within the data set, or identify a location based on given data.