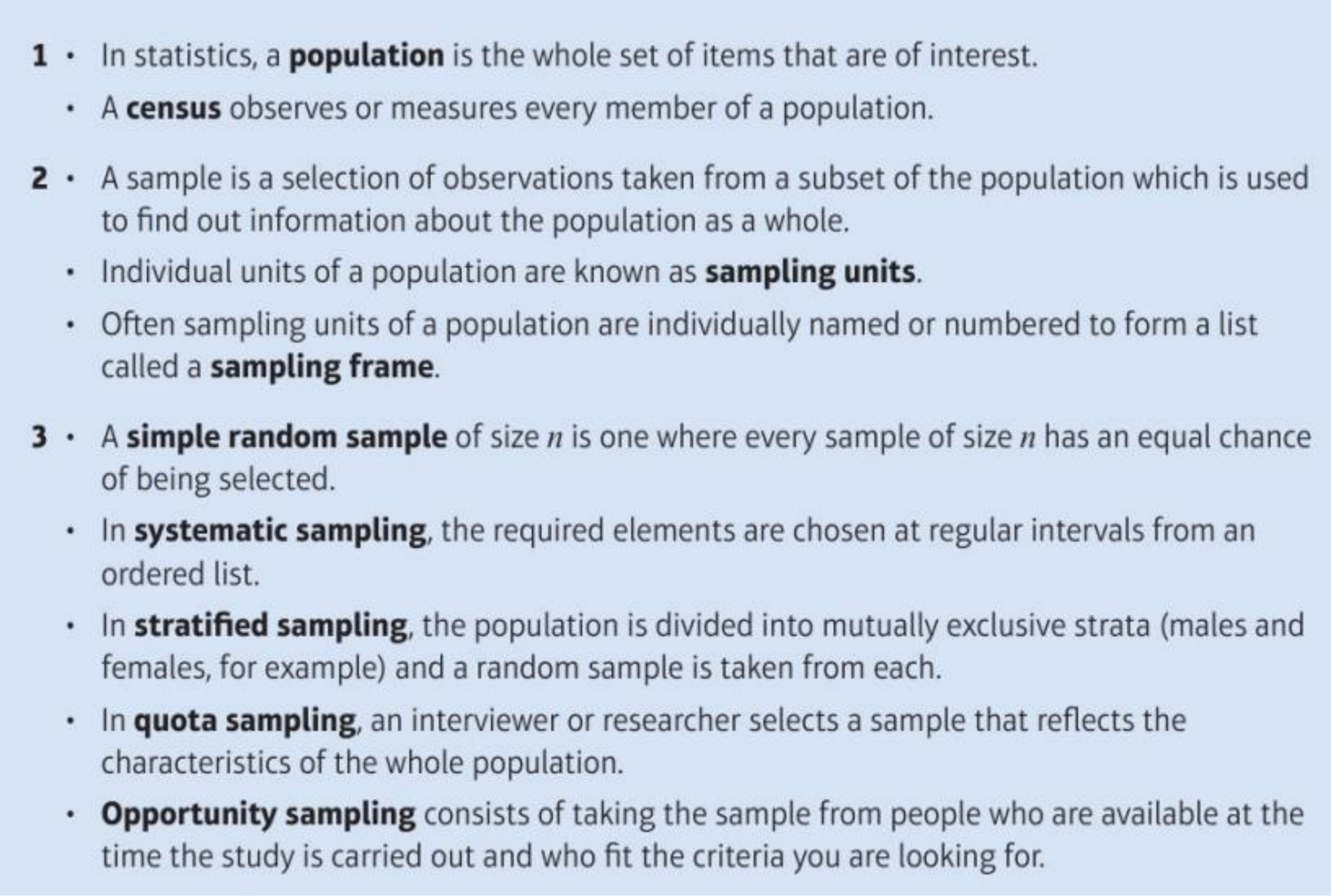
Stats1 Chapter 1: Data Collection

Chapter Practice

Key Points

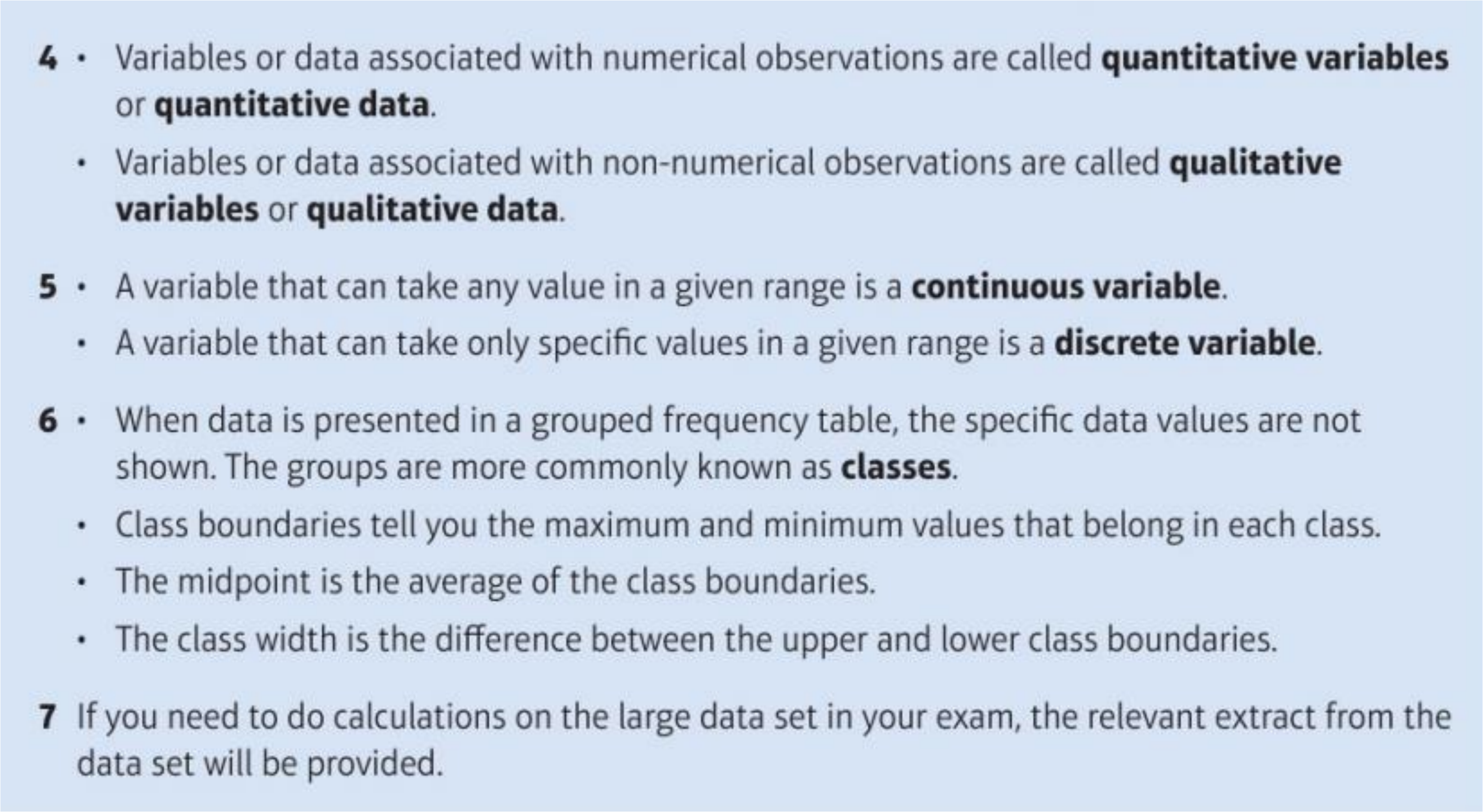


1
- In statistics, a **population** is the whole set of items that are of interest.
- A **census** observes or measures every member of a population.

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.

Key Points



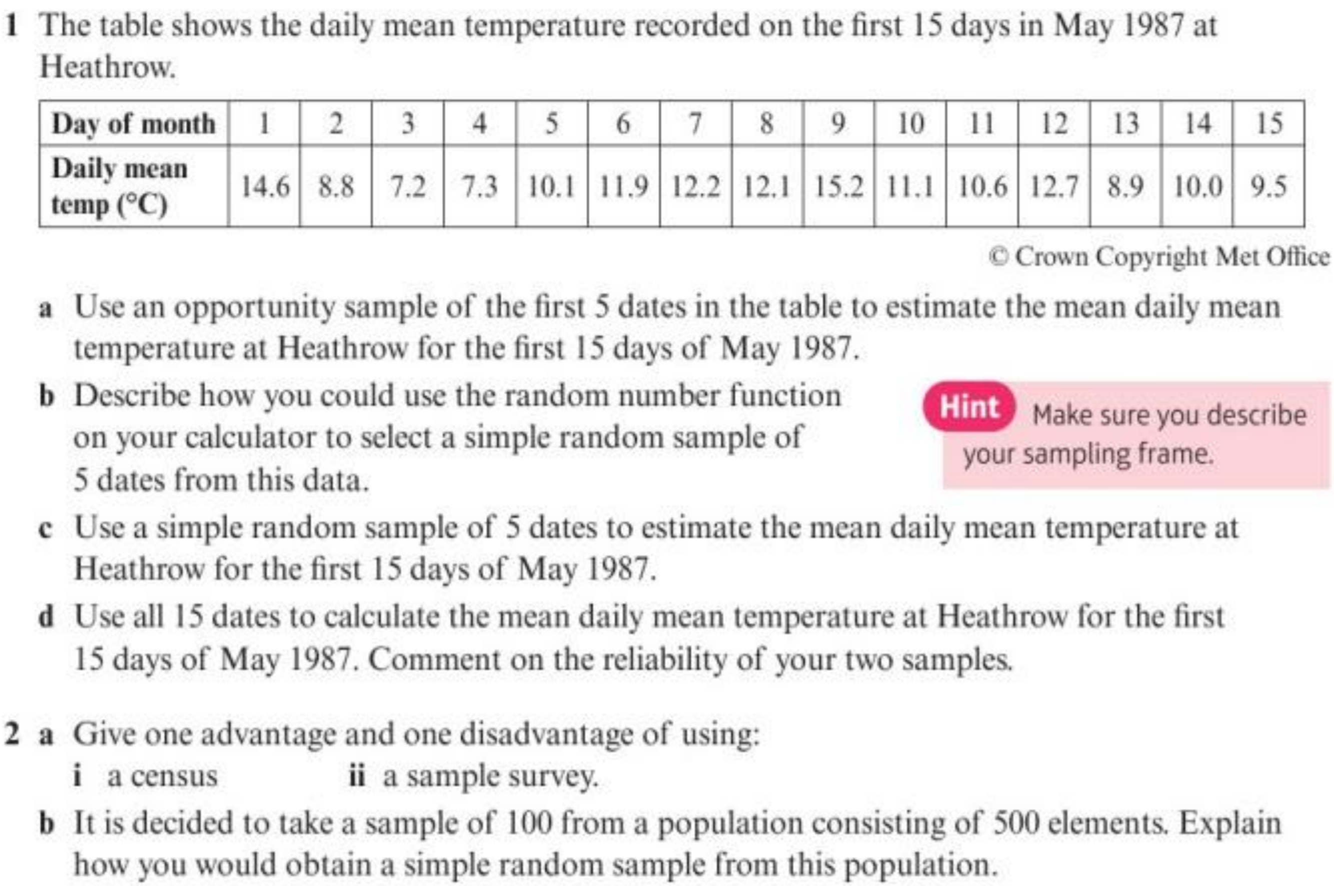
4
- Variables or data associated with numerical observations are called **quantitative variables** or **quantitative data**.
- Variables or data associated with non-numerical observations are called **qualitative variables** or **qualitative data**.

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.

Chapter Exercises



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

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.

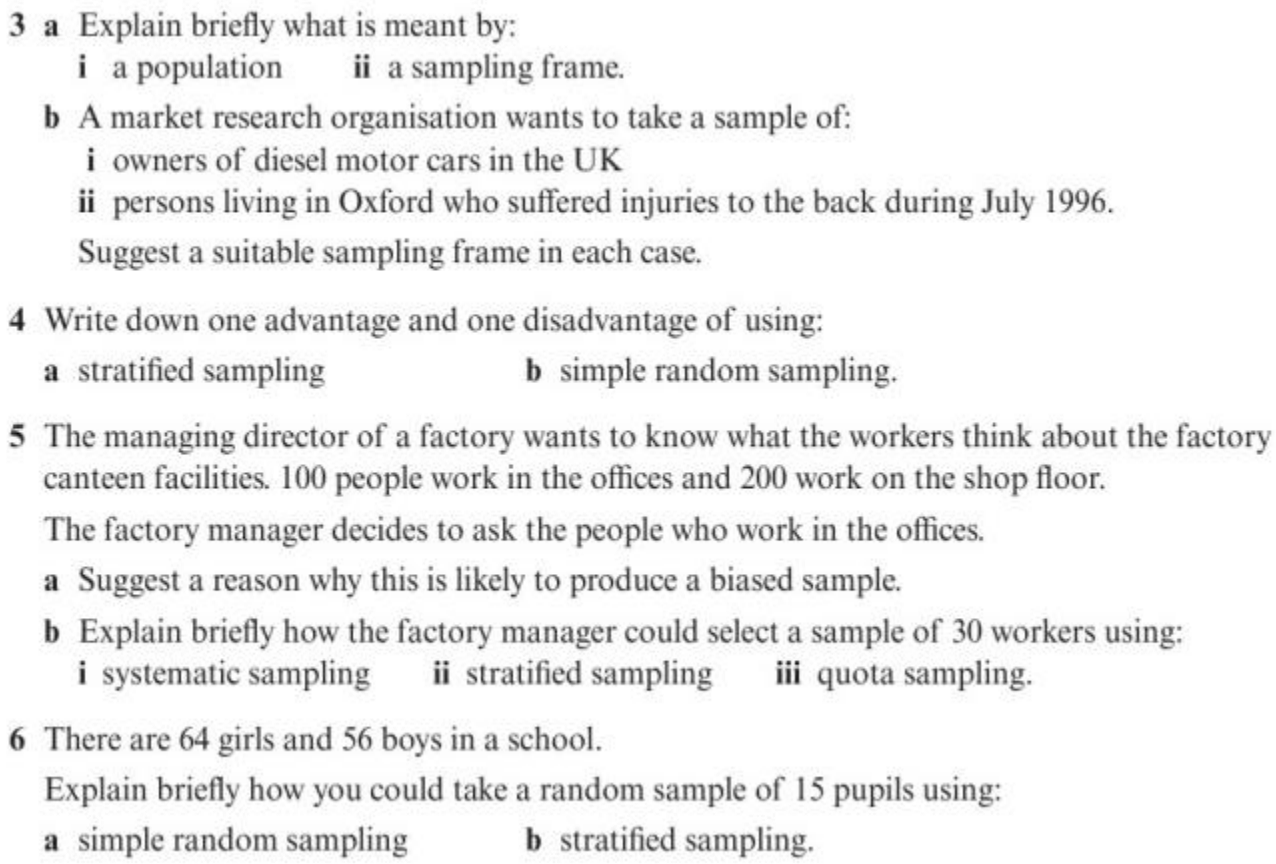
d Use all 15 dates to calculate the mean daily mean temperature at Heathrow for the first 15 days of May 1987. Comment on the reliability of your two samples.

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.

Chapter Exercises



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:

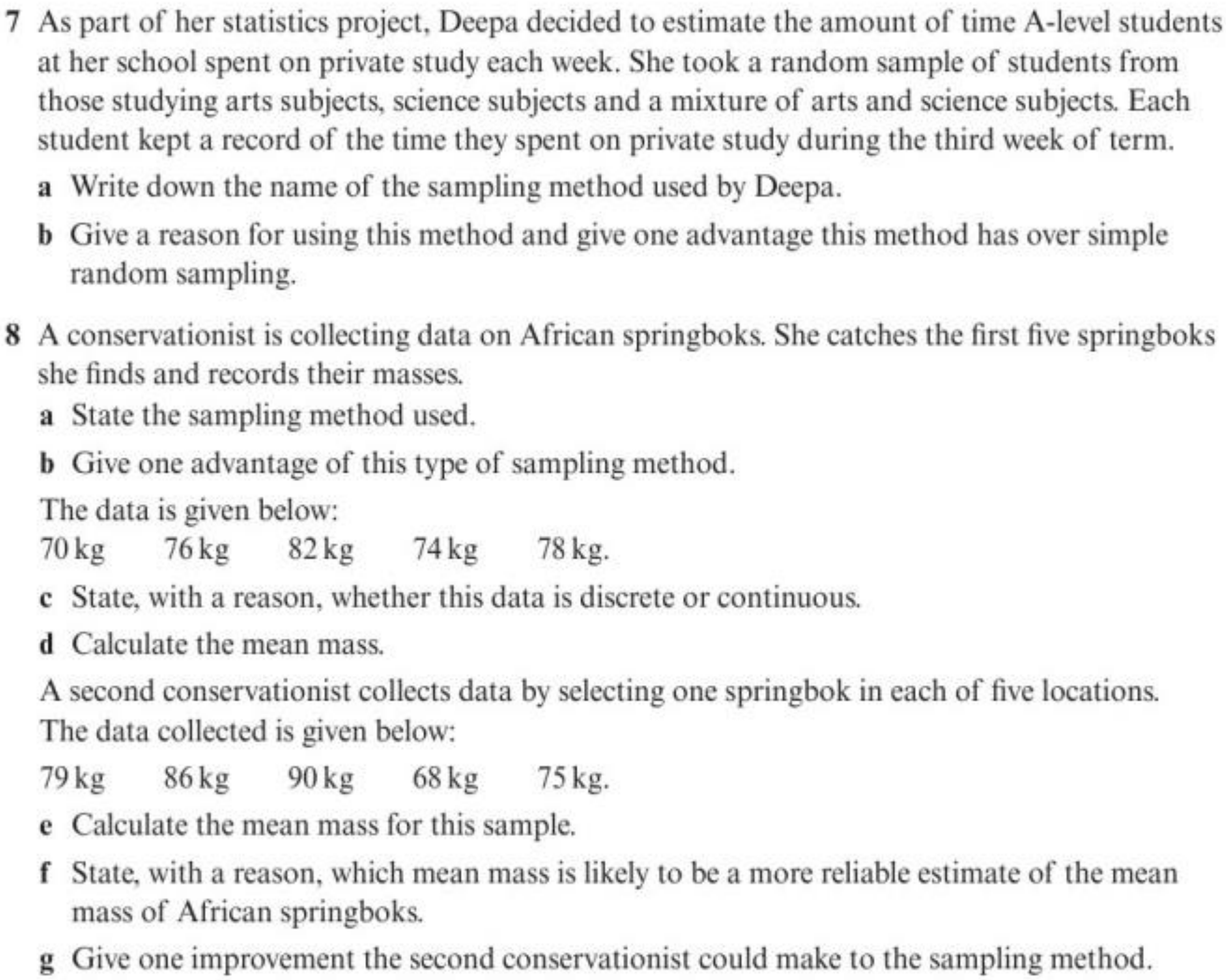
i systematic sampling **ii** stratified sampling **iii** quota sampling.

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.

Chapter Exercises



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:

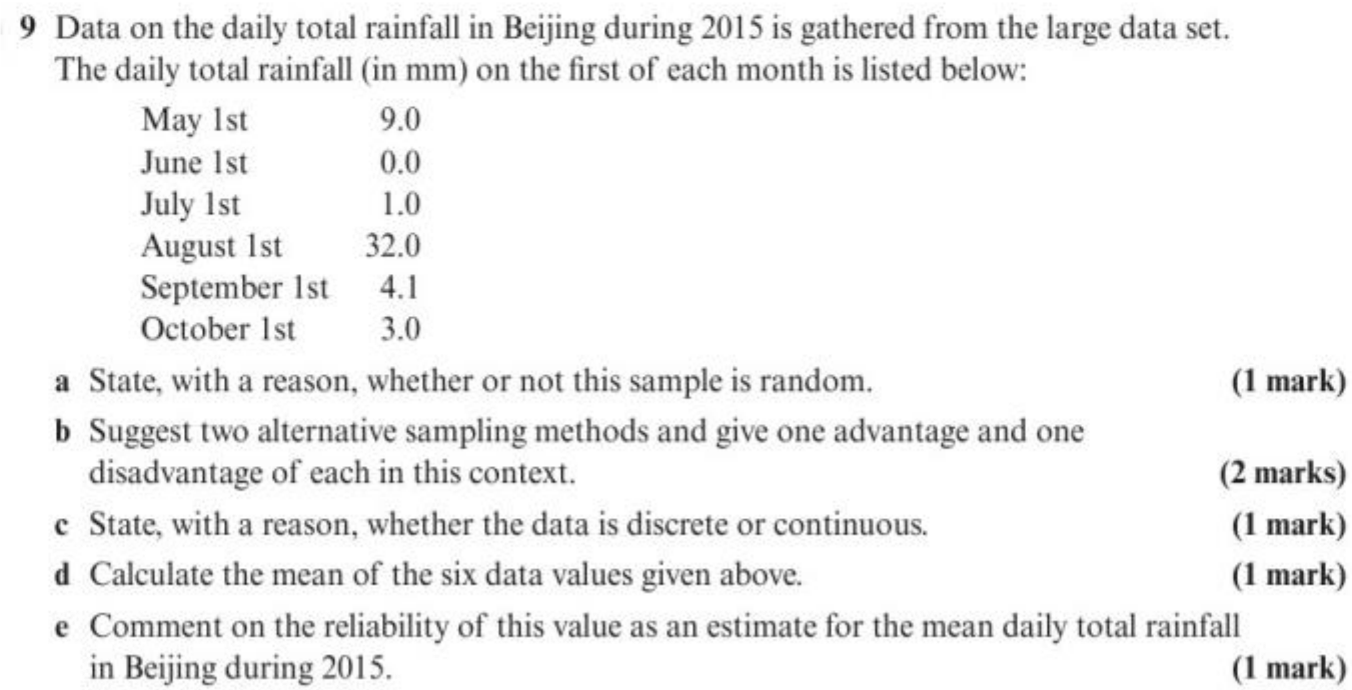
79 kg 86 kg 90 kg 68 kg 75 kg.

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.

Chapter Exercises



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)**

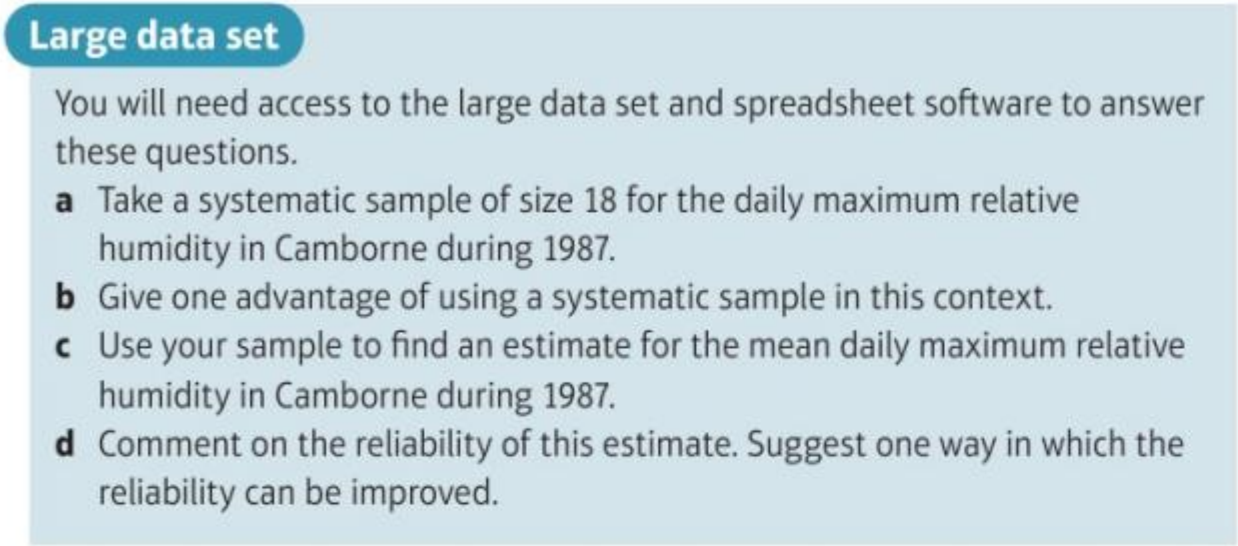
b Suggest two alternative sampling methods and give one advantage and one disadvantage of each in this context. **(2 marks)**

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)**

Chapter Exercises



Large data set

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

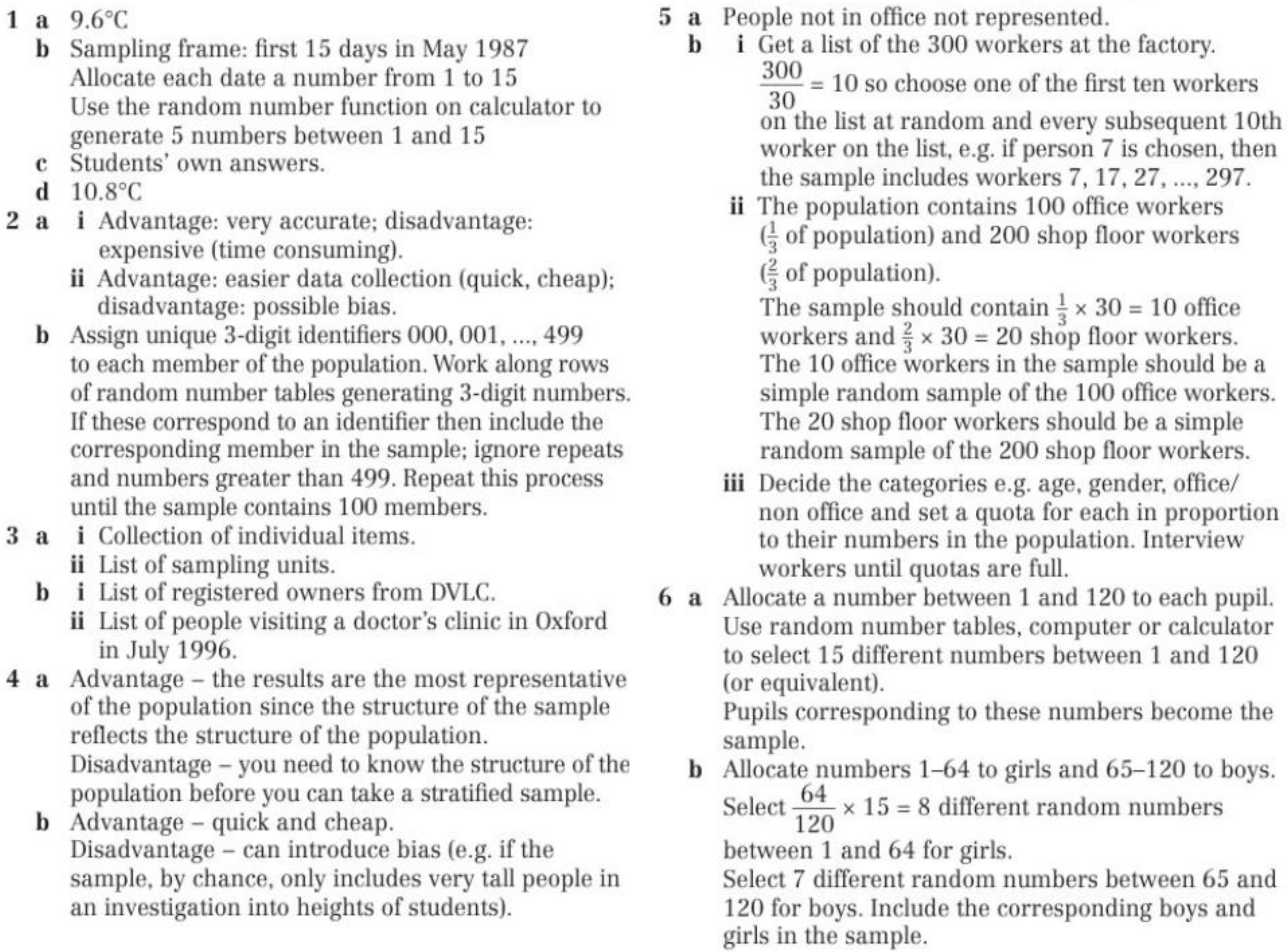
a Take a systematic sample of size 18 for the daily maximum relative humidity in Camborne during 1987.

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.

Chapter Answers



1 **a** 9.6°C

b Sampling frame: first 15 days in May 1987
Allocate each date a number from 1 to 15
Use the random number function on calculator to generate 5 numbers between 1 and 15

c Students' own answers.

d 10.8°C

2 **a** **i** Advantage: very accurate; disadvantage: expensive (time consuming).

ii Advantage: easier data collection (quick, cheap); disadvantage: possible bias.

b Assign unique 3-digit identifiers 000, 001, ..., 499 to each member of the population. Work along rows of random number tables generating 3-digit numbers. If these correspond to an identifier then include the corresponding member in the sample; ignore repeats and numbers greater than 499. Repeat this process until the sample contains 100 members.

3 **a** **i** Collection of individual items.

ii List of sampling units.

b **i** List of registered owners from DVLC.

ii List of people visiting a doctor's clinic in Oxford in July 1996.

4 **a** Advantage – the results are the most representative of the population since the structure of the sample reflects the structure of the population.
Disadvantage – you need to know the structure of the population before you can take a stratified sample.

b Advantage – quick and cheap.
Disadvantage – can introduce bias (e.g. if the sample, by chance, only includes very tall people in an investigation into heights of students).

5 **a** People not in office not represented.

b **i** Get a list of the 300 workers at the factory.
$\frac{300}{30} = 10$ so choose one of the first ten workers on the list at random and every subsequent 10th worker on the list, e.g. if person 7 is chosen, then the sample includes workers 7, 17, 27, ..., 297.

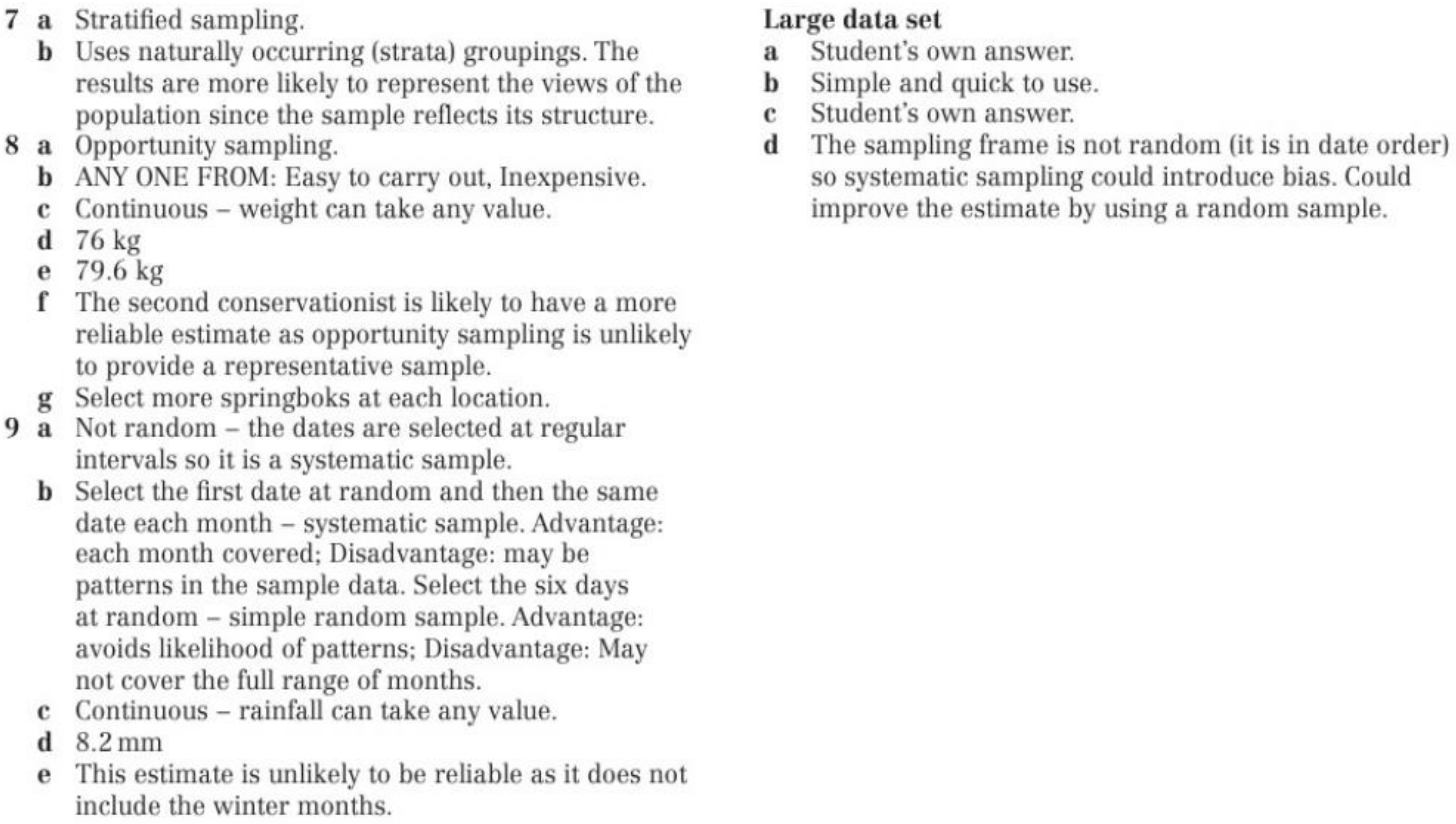
ii The population contains 100 office workers ($\frac{1}{3}$ of population) and 200 shop floor workers ($\frac{2}{3}$ of population).
The sample should contain $\frac{1}{3} \times 30 = 10$ office workers and $\frac{2}{3} \times 30 = 20$ shop floor workers.
The 10 office workers in the sample should be a simple random sample of the 100 office workers.
The 20 shop floor workers should be a simple random sample of the 200 shop floor workers.

iii Decide the categories e.g. age, gender, office/non office and set a quota for each in proportion to their numbers in the population. Interview workers until quotas are full.

6 **a** Allocate a number between 1 and 120 to each pupil.
Use random number tables, computer or calculator to select 15 different numbers between 1 and 120 (or equivalent).
Pupils corresponding to these numbers become the sample.

b Allocate numbers 1–64 to girls and 65–120 to boys.
Select $\frac{64}{120} \times 15 = 8$ different random numbers between 1 and 64 for girls.
Select 7 different random numbers between 65 and 120 for boys. Include the corresponding boys and girls in the sample.

Chapter Answers



7 **a** Stratified sampling.
b Uses naturally occurring (strata) groupings. The results are more likely to represent the views of the population since the sample reflects its structure.

8 **a** Opportunity sampling.
b ANY ONE FROM: Easy to carry out, Inexpensive.
c Continuous – weight can take any value.
d 76 kg
e 79.6 kg
f The second conservationist is likely to have a more reliable estimate as opportunity sampling is unlikely to provide a representative sample.
g Select more springboks at each location.

9 **a** Not random – the dates are selected at regular intervals so it is a systematic sample.
b Select the first date at random and then the same date each month – systematic sample. Advantage: each month covered; Disadvantage: may be patterns in the sample data. Select the six days at random – simple random sample. Advantage: avoids likelihood of patterns; Disadvantage: May not cover the full range of months.
c Continuous – rainfall can take any value.
d 8.2 mm
e This estimate is unlikely to be reliable as it does not include the winter months.

Large data set

a Student's own answer.
b Simple and quick to use.
c Student's own answer.
d The sampling frame is not random (it is in date order) so systematic sampling could introduce bias. Could improve the estimate by using a random sample.