Data Collection, Mixed exercise 1

1 a Mean daily maximum temperature $= \dfrac{14.6 + 8.8 + 7.2 + 7.3 + 10.2}{5} = \dfrac{48}{5} = 9.6°\text{C}$

b The sampling frame is the first 15 days in May 1987. Allocate each date a number from 1 to 15. Use the random number function on a calculator to generate 5 random numbers between 1 and 15 and choose the dates with these numbers.

c For example: The 5 random numbers generated are 12, 8, 13, 6 and 15. These give the temperatures 12.7, 12.1, 8.9, 11.9 and 9.5.

Mean daily max temperature $= \dfrac{55.1}{5} = 11.0°\text{C}$ (to 1 d.p.)

d Mean daily max temperature $= \dfrac{162.2}{15} = 10.8°\text{C}$ (to 1 d.p.)

Using data for the whole period gives a completely accurate result. A sample of one third of the dates will give a less reliable one. This variation is apparent in the answers to a and c, above.

2 a i Advantage: very accurate.
Disadvantages: expensive, time consuming and difficult to process.

ii Advantages: easier data collection, quick and cheap.
Disadvantages: less accurate, less representative and possibly biased.

b Assign unique 3-digit identifiers 000, 001, …, 499 to each member of the population. Use random number tables, a computer or a calculator to generate 3-digit numbers. If these correspond to an identifier then include the corresponding member in the sample, being careful to 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, with each unit given an identifying name or number.

b i List of registered owners from the DVLA.

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

4 a Advantages:

The results are the most representative of the population since the structure of the sample reflects the structure of the population.
It guarantees proportional representation of groups within a population.

Disadvantages:

You need to know the structure of the population before you can take a stratified sample.
Classification into mutually exclusive strata may be difficult to implement.
The sampling within each strata may suffer from the disadvantages of simple random sampling.

4 **b** Advantages:

Quick
Cheap.
All units have an equal chance of selection.

Disadvantages:

Can introduce bias (e.g. if the sample, by chance, only includes very tall people in an investigation into heights of students).
A sampling frame is needed first.

5 **a** People on the shop floor are 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, …, 287 and 297.

ii The sample should contain $\frac{1}{3}\times 30=10$ office workers and $\frac{2}{3}\times 30=20$ shop floor workers, as those are the corresponding proportions in the whole population. The 10 office workers in the sample should be selected by a simple random sample of the 100 office workers. The 20 shop floor workers should be selected by 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 (the total number of pupils) to each pupil. Use random number tables, a computer or a calculator to select 15 different whole numbers between 1 and 120.

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 the remaining 7 sampling units using random numbers between 65 and 120 for boys.

Include the corresponding boys and girls in the sample.

7 **a** Stratified sampling.

b This method uses naturally occurring groupings (strata). The results are more likely to represent the views of the whole population since the sample reflects its structure.

8 **a** Opportunity sampling

b Any one of:

It is easy to carry out.
It is inexpensive.

c The data is continuous, as weight can take any positive value.

8 **d** Mean weight = $\frac{70+76+82+74+78}{5}=\frac{380}{5}$ = 76 kg

e Mean weight = $\frac{79+86+90+68+75}{5}=\frac{398}{5}$ = 80 kg (to the nearest whole number)

f The second conservationist is likely to have a more reliable estimate as opportunity sampling is unlikely to provide a representative sample of the whole population, as it does not necessarily reflect its structure.

g The second conservationist could select more springboks at each location.

9 **a** This sample is not entirely random as the dates are selected at regular intervals. It is actually a systematic sample.

b A systematic sample: select the first date at random and then the same date each month. An advantage of a systematic sample is that each month is covered. A disadvantage of a systematic sample is that there may be patterns, and therefore bias, in the sample data.

A simple random sample: select the six days completely at random. An advantage of a simple random sample is that it avoids the likelihood of patterns and unintentional bias. A disadvantage of a simple random sample is that it may not cover the full range of months.

c The data is continuous as temperature can take any value.

d Mean daily maximum temperature = $\frac{59.1}{6}$= 9.9 °C (to 1 d.p.)

e Any suitable reason e.g. this estimate is unlikely to be reliable as it does not include the winter months / the data is very variable.

Large data set

a $\frac{184}{18}\approx 10$, so take every 10th day, choosing a starting place by getting a calculator to generate a random number, for example, 10. The sample is then:

79, 99, 99, 86, 95, 99, 93, 99, 100, 93, 99, 98, 98, 100, 89, 99, 95, 95

b Any of: easy, quick, more suitable for this relatively large population..

c Mean = $\frac{1715}{18}$ = 95% (to the nearest whole number)

d The sampling frame is not random (it is in date order) so systematic sampling could introduce bias. This bias could be counteracted by using simple random sampling.