Summary of key points

1 The **mode** or **modal class** is the value or class that occurs most often.

2 The **median** is the middle value when the data values are put in order.

3 The **mean** can be calculated using the formula $\bar{x} = \frac{\Sigma x}{n}$.

4 For data given in a frequency table, the mean can be calculated using the formula $\bar{x} = \frac{\Sigma xf}{\Sigma f}$.

5 To find the **lower quartile** for discrete data, divide n by 4. If this is a whole number, the lower quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

6 To find the **upper quartile** for discrete data, find $\frac{3}{4}$ of n. If this is a whole number, the upper quartile is halfway between this data point and the one above. If it is not a whole number, round *up* and pick this data point.

7 The **range** is the difference between the largest and smallest values in the data set.

8 The **interquartile range** (IQR) is the difference between the upper quartile and the lower quartile, $Q_3 - Q_1$.

9 The **interpercentile range** is the difference between the values for two given percentiles.

10 **Variance** $= \frac{\Sigma(x - \bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{S_{xx}}{n}$ where $S_{xx} = \Sigma(x - \bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

11 The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{S_{xx}}{n}}$$

12 You can use these versions of the formulae for variance and standard deviation for grouped data that is presented in a frequency table:

$$\sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{\Sigma f} = \frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2 \qquad \sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{\Sigma f}} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2}$$

where f is the frequency for each group and Σf is the total frequency.

13 If data is coded using the formula $y = \frac{x - a}{b}$

- the mean of the coded data is given by $\bar{y} = \frac{\bar{x} - a}{b}$
- the standard deviation of the coded data is given by $\sigma_y = \frac{\sigma_x}{b}$ where σ_x is the standard deviation of the original data.

Had a look ☐ Nearly there ☐ Nailed it! ☐

Mean

The mean is a measure of **central tendency**. The mean is sometimes written as $\bar{x}$. You might have to calculate an **estimate** of the mean of **grouped** data given in a **frequency table**.

Formulae for the mean

Learn these two formulae for the mean – they're not given in the formulae booklet.

1 For n discrete data values, the mean is:

$$\bar{x} = \frac{\sum x}{n}$$

$\sum x$ – The **sum** of the data values

n – The **number** of data values

2 For data given in a frequency table:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$\sum fx$ – The **sum** of (frequency × data value) or (frequency × midpoint)

$\sum f$ – The **total** frequency

Grouped data

You can use formula 2 on the left with a **grouped frequency distribution**. The value of x for each group is the **midpoint** of that group.

Height, h (cm)	Midpoint
$0 \leqslant t < 5$	2.5
$5 \leqslant t < 7$	6
$7 \leqslant t < 10$	8.5

$\frac{0+5}{2} = 2.5$

$\frac{5+7}{2} = 6$

Watch out!

You can calculate the mean of a set of discrete data quickly using the statistics functions on your calculator. However, you need to understand the formula in case you are given summarised data.

Worked example

A researcher is studying the rainfall in Jacksonville. She uses the large data set to obtain the 24-hour rainfall total, x mm, each day in September 2015. She summarises her results as follows:

$\sum x = 316.2$

Calculate the mean 24-hour rainfall total in Jacksonville in September 2015. **(1 mark)**

$$\bar{x} = \frac{\sum x}{n} = \frac{316.2}{30} = 10.5 \text{ mm (3 s.f.)}$$

There are 30 days in September so $n = 30$.

There is more about the large data set on page 60.

Worked example

A teacher asked her class of 30 students how long, to the nearest hour, they spent on a homework project.

Hours	1–3	4–7	8–10	11–20
Frequency	7	12	8	3

Estimate the mean of the time spent on the project. **(2 marks)**

Midpoint	2	5.5	9	15.5
Frequency	7	12	8	3

$$\bar{x} = \frac{\sum fx}{\sum f}$$

$$= \frac{7 \times 2 + 12 \times 5.5 + 8 \times 9 + 3 \times 15.5}{30}$$

$$= \frac{198.5}{30} = 6.6 \text{ hours (1 d.p.)}$$

Start by working out the midpoint of each class interval. The midpoint of the 4–7 class is $\frac{3.5 + 7.5}{2} = 5.5$. In your exam you can estimate the mean quickly using your calculator.

There is more about using a calculator with grouped frequency tables on page 61.

Now try this

Jamie recorded the temperature at his school at midday, x °C, each day for 15 days. He summarised his results: $\sum x = 251$

Paul recorded the temperature in °C on the next 5 days.

Here are his results: 19 22 15 21 16

(a) Calculate the mean temperature during the whole 20 days. Answer **(2 marks)**

On the next day, the midday temperature was 16 °C.

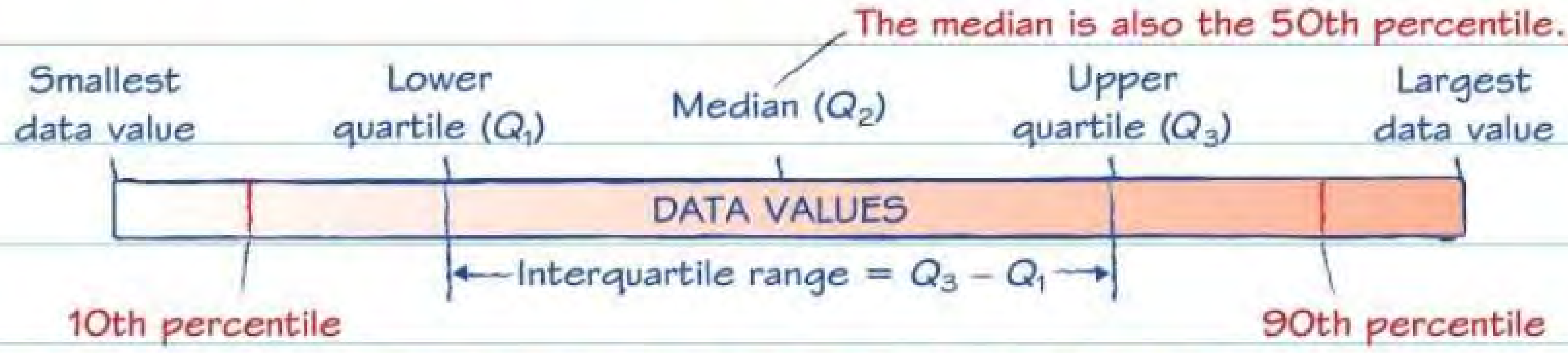
(b) State, giving a reason, the effect this will have on the mean for the whole 21 days. **(1 mark)**

Answer

You know the sum of the first 15 values. Add the next 5 values then divide by the total number of values, 20.

Median and quartiles

The **median** (Q_2), the **lower quartile** (Q_1), the **upper quartile** (Q_3) and the **percentiles** are all measures of **location**.



Discrete data

Follow these steps to work out the median and quartiles for n **discrete** data values.

1. Q_1: Work out $\frac{n}{4}$
2. Q_2: Work out $\frac{n}{2}$
3. Q_3: Work out $\frac{3n}{4}$

Is it a WHOLE NUMBER?

YES → Use the value **half way** between this data value and the **next** one.

NO → Round up to the next **whole number** and use this data value.

You can revise how to find the median and quartiles for **grouped continuous data** on the next page.

Worked example

The marks obtained by 18 people who took the driving theory test are given below:

31 40 18 43 31 20 35 34 43
45 44 43 35 33 29 31 23 31

(a) Write down the modal mark. **(1 mark)**

31

(b) Find the lower quartile, the median and the upper quartile. **(3 marks)**

18 20 23 29 31 31 31 31 33 34
35 35 40 43 43 43 44 45

$\frac{n}{4} = 4.5$, $Q_1 = 31$

$\frac{n}{2} = 9$, $Q_2 = \frac{33 + 34}{2} = 33.5$

$\frac{3n}{4} = 13.5$, $Q_3 = 43$

The **modal** mark is the most frequently occurring mark. 31 appears 4 times, so this is the modal mark.

Write the data in order of size before attempting to find the median and quartiles.
$\frac{n}{4} = 4.5$. This is **not** a whole number, so round up. The lower quartile (Q_1) is the 5th data value.
$\frac{n}{2} = 9$. This is a whole number, so the median (Q_2) is **half way** between the 9th and 10th data values.
$\frac{3n}{4} = 13.5$. This is not a whole number, so round up. The upper quartile (Q_3) is the 14th data value.

You can use the statistics mode on your calculator to enter **1-variable** data like this and calculate Q_1, Q_2 and Q_3 quickly. There are different conventions for calculating the quartiles. If your calculator gives a different answer, either one is valid.

Now try this

The daily mean air pressure (hPa) at Heathrow for the first 22 days in May 1987 was measured and recorded in the large data set. The data are given below.

1017 1014 1023 1032 1033 1032 1031 1026 1016 1015 1011
1002 1011 1002 1008 1014 1012 1012 1022 1030 1026 1021

For this period, find

(a) the median air pressure Answer **(1 mark)**

(b) the interquartile range. Answer **(2 marks)**

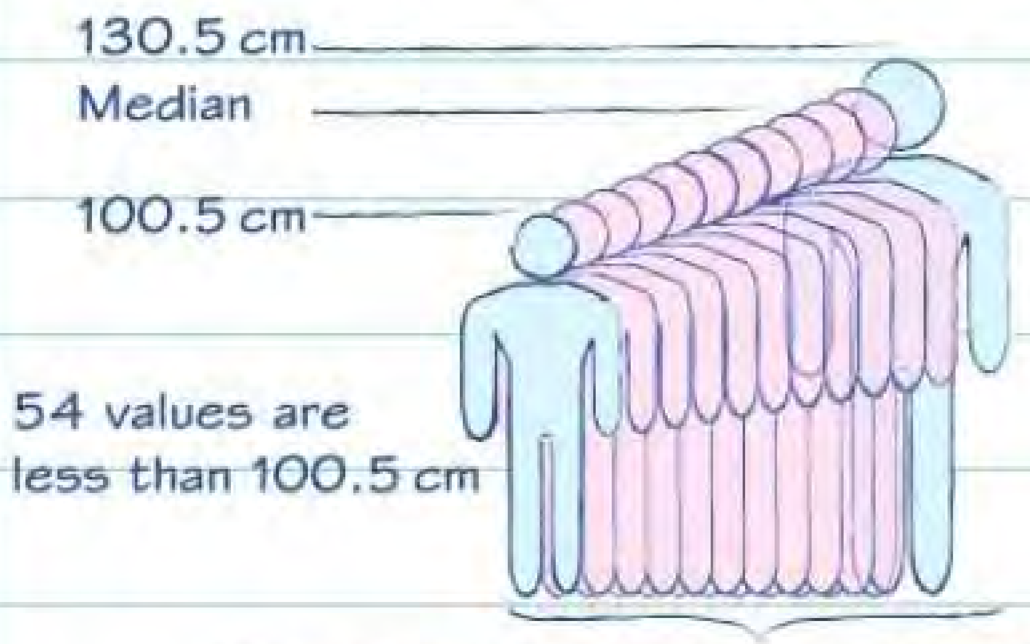
Had a look ☐ Nearly there ☐ Nailed it! ☐

Linear interpolation

You can use linear interpolation to estimate the **median**, **quartiles** and **percentiles** of grouped continuous data. You don't know the exact data values, so you assume the data values are **evenly distributed** within each group.

Using proportion

In the Worked example on the right, the median is in the 101–130 cm group. To use interpolation, you assume that the heights in this group are **evenly distributed** between the **lower** and **upper class boundaries**.



There are 125 students in the 101–130 group, so each student represents 30 ÷ 125 = 0.24 cm

There are 54 values below 100.5 cm so the median is 145 − 54 = 91 values into the group. Each value represents 0.24 cm, so the median is 100.5 + 91 × 0.24 = 122.34 cm.

Worked example

This table shows the heights, to the nearest cm, of 290 students in a school.

Height (cm)	Number of students
71–100	54
101–130	125
131–160	87
161–190	24

Use interpolation to estimate the median height. **(2 marks)**

$\frac{290}{2} = 145$

$100.5 + (145 - 54) \times \frac{30}{125} = 122.34$ cm

Work out $\frac{n}{2}$ to find the position of the median.

Choosing class boundaries

You need to be careful when you are working out lower and upper class boundaries.

Time (mins)
1–5
6–8
9–12
13–20

These data have been rounded to the nearest minute and there are GAPS between the groups. The lower boundary for this group is 8.5 minutes, and the upper boundary is 12.5 minutes.

Distance, x (cm)
$100 \leqslant x < 150$
$150 \leqslant x < 250$
$250 \leqslant x < 350$
$350 \leqslant x < 500$

There are NO GAPS between the groups here. The lower boundary for this group is 100 cm and the upper boundary is 150 cm.

Now try this

This table summarises the hand spans, to the nearest cm, of a group of 178 people.

Hand span (cm)	Frequency
10–14	29
15–17	64
18–19	55
20–22	21
23–30	9

Use interpolation to estimate the median Q_2, the lower quartile Q_1 and the 90th percentile of these data. Answer **(4 marks)**

You can use interpolation to estimate the quartiles in exactly the same way as you estimate the median.

$\frac{n}{4} = 44.5$, so Q_1 is $44.5 - 29 = 15.5$ values into the 15–17 group. This group is $17.5 - 14.5 = 3$ cm wide, so each value represents $\frac{3}{64}$ cm. The estimate for Q_1 is

$14.5 + (44.5 - 29) \times \frac{3}{64}$

Round your answers to 2 decimal places.

Standard deviation 1

Standard deviation and **variance** are both measures of **spread** (or **dispersion**). Standard deviation is used more frequently because it has the **same units** as the data.

Discrete data

For n discrete data values:

$$\text{Variance} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2$$

$$\text{Standard deviation} = \text{SD} = \sqrt{\text{Variance}}$$

You can revise how to find the variance and standard deviation for **grouped continuous data** on the next page.

Notation

You need to be familiar with the notations used in statistics formulae.

- $\sum x^2$ means the sum of the squares of each of the data values.
- σ is sometimes used for the standard deviation.
- σ^2 is sometimes used for the variance.

Worked example

The ages, x years, of 50 passengers on a coach were recorded.

Given that $\sum x = 1756$ and $\sum x^2 = 69\,942$, calculate the standard deviation of the ages. **(2 marks)**

$$\text{Variance} = \frac{69942}{50} - \left(\frac{1756}{50}\right)^2 = 165.4256$$

$$\text{SD} = \sqrt{165.4256} = 12.9 \text{ years (3 s.f.)}$$

Summary statistics

You will often be given summary statistics like $\sum x$, $\sum x^2$ and S_{xx} in your exam. You can use these to simplify your calculations:

$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$, so another formula for the **variance** is $\frac{S_{xx}}{n}$.

The large data set

Some of the questions in your statistics exam will be based on meteorological data from the large data set. You don't need to learn any data from the large data set, but you will need to be **familiar** with it. This means:

- Knowing the approximate locations of each of the eight weather stations
- Understanding the different categories of data and the approximate values they can take
- Recognising coastal and non-coastal locations
- Knowing that high humidity can give rise to misty and foggy conditions
- Identifying likely locations based on given statistics (such as knowing that it is generally warmer in Beijing than in UK locations)

Now try this

The daily maximum temperature, t °C, in Camborne for the month of May 1987 was recorded and the data were summarised as follows:

$\sum t = 397.3$ and $\sum t^2 = 5165.39$

Answer

(a) Calculate the mean and standard deviation, giving both answers correct to 1 d.p. **(2 marks)**

The mean and standard deviation of the daily maximum temperature in Leuchars for the same period were 13.6 °C and 1.2 °C respectively.

(b) State, with a reason, whether these results support the conclusion that southern locations are warmer than northern locations. Answer **(2 marks)**

Standard deviation 2

You can **estimate** the variance and standard deviation of **grouped continuous data** given in a frequency table using the **midpoints**, x, of each group.

Grouped continuous data

For data given in a frequency table:

$$\text{Variance} = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{\sum fx^2}{n} - (\bar{x})^2$$

$$\text{Standard deviation} = \text{SD} = \sqrt{\text{Variance}}$$

Learn these formulae – they are **not** in the booklet.

Notation

$\sum fx^2$ is the **sum** of (frequency × midpoint^2).

This is **not** the same as $\left(\sum fx\right)^2$.

For a reminder about the other notation used in the formula for variance, have a look at pages 57 and 60.

Worked example

This table shows the times a random sample of 150 people took to complete a puzzle.

Time, t (mins)	Number of people	Midpoint
$0 \leqslant t < 5$	23	2.5
$5 \leqslant t < 7$	82	6
$7 \leqslant t < 10$	36	8.5
$10 \leqslant t < 15$	9	12.5

Use your calculator to estimate the mean and the standard deviation of the length of time taken. **(3 marks)**

$\bar{t} = 3.34$ mins (3 s.f.)

$\sigma = 2.39$ mins (3 s.f.)

Use the midpoint of each class interval as your x-value for that class. You could work out the standard deviation using the formulae given above:

$\sum f = 290$

$\sum fx = 23 \times 2.5 + 82 \times 6 + 36 \times 8.5 + 9 \times 12.5 = 968$

$\sum fx^2 = 23 \times 2.5^2 + 82 \times 6^2 + 36 \times 8.5^2 + 9 \times 12.5^2 = 7103$

$\sigma = \sqrt{\frac{7103}{150} - \left(\frac{968}{150}\right)^2} = 2.39$ (3 s.f.)

However, it is quicker to work it out using the statistical functions on your calculator.

Calculator skills

You need to be confident using the statistical functions on your calculator. If you have to enter a grouped frequency distribution like the one above into your calculator, you should enter the **midpoints** of each class interval as the x-values. You might need to make sure you have switched on **frequency mode** on your calculator to enter a frequency distribution.

	x	Freq
1	2.5	23
2	6	82
3	8.5	36
4	12.5	9

Enter data values into your calculator very carefully!

Now try this

A researcher timed how long, to the nearest minute, a group of shoppers spent in a supermarket checkout queue.

Minutes	1–3	4–6	7–10	11–20
Frequency	48	31	15	6

Use your calculator to estimate the mean and standard deviation. Answer **(3 marks)**

Coding

Coding is a way of making statistics calculations easier. Each data value is **coded** to make a new set of data values. It is especially useful when you are dealing with **very large** or **very small** data values.

Transforming data

You can use this formula to transform a data value x into a new data value y:

$$y = \frac{x - a}{b}$$

The numbers a and b are chosen to make the final y-values easy to work with.

Effects of coding

If data is coded, different statistics will change in different ways.

	Before	After
Data values	x	$\frac{x-a}{b}$
Mean	$\bar{x}$	$\frac{\bar{x}-a}{b}$
Median	m	$\frac{m-a}{b}$
Standard deviation	s	$\frac{s}{b}$

Worked example

A scientist measures the temperature, x, in a nuclear reactor.

She uses the coding $y = \frac{x - 300}{10}$ to code her data.

The mean of the coded data is $\bar{y} = 3$ and the standard deviation of the coded data is $\sigma_y = 1.72$.

Calculate the mean and standard deviation of the original data. **(2 marks)**

$3 = \frac{\bar{x} - 300}{10}$ so $\bar{x} = 30 + 300$

$= 330\,°\text{C}$

$1.72 = \frac{\sigma_x}{10}$ so $\sigma_x = 17.2\,°\text{C}$ (3 s.f.)

Problem solved!

You will often have to calculate the mean or standard deviation of the **original** data, based on information about the **coded** data. You need to rearrange the rules given above:

$\bar{y} = \frac{\bar{x} - a}{b}$ so $\bar{x} = b\bar{y} + a$

$\sigma_y = \frac{\sigma_x}{b}$ so $\sigma_x = n\sigma_y$

You will need to use problem-solving skills throughout your exam – **be prepared!**

Watch out! The **standard deviation** measures **spread** so it is only affected when the data are multiplied or divided by a constant. Adding or subtracting a constant has no effect.

Now try this

1 From the large data set, the daily mean visibility, x metres, in Leeming between May and October 1987 was recorded.

The data is coded using the coding $y = \frac{x}{100}$ and the following summary statistics are obtained:

$n = 184$, $\sum y = 4380$, $S_{yy} = 2727.3$

Find the mean and standard deviation of the daily mean visibility in Leeming between May and October 1987. Answer **(4 marks)**

Write down a coding formula that corresponds to a reduction of 8 followed by a 10% increase.

2 A group of 10 students took a computer-marked maths test. Their scores, x, are summarised as follows:

$\sum x = 668$ $\quad \sum x^2 = 47\,870$

(a) Find the mean and standard deviation of the marks. Answer **(3 marks)**

An instructor discovered that a computer error had caused one question to be marked correct on all the students' tests. He reduced each student's mark by 8, and then increased it by 10%.

(b) Find the mean and standard deviation of the adjusted marks. **(2 marks)**

Answer

Summary of key points

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.

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						
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