Questions

Q1.

Sara is investigating the variation in daily maximum gust, *t* kn, for Camborne in June and July 1987.

She used the large data set to select a sample of size 20 from the June and July data for 1987. Sara selected the first value using a random number from 1 to 4 and then selected every third value after that.

(a) State the sampling technique Sara used.

(1)

(b) From your knowledge of the large data set, explain why this process may not generate a sample of size 20.

(1)

The data Sara collected are summarised as follows

$$n = 20 \qquad \sum t = 374 \qquad \sum t^2 = 7600$$

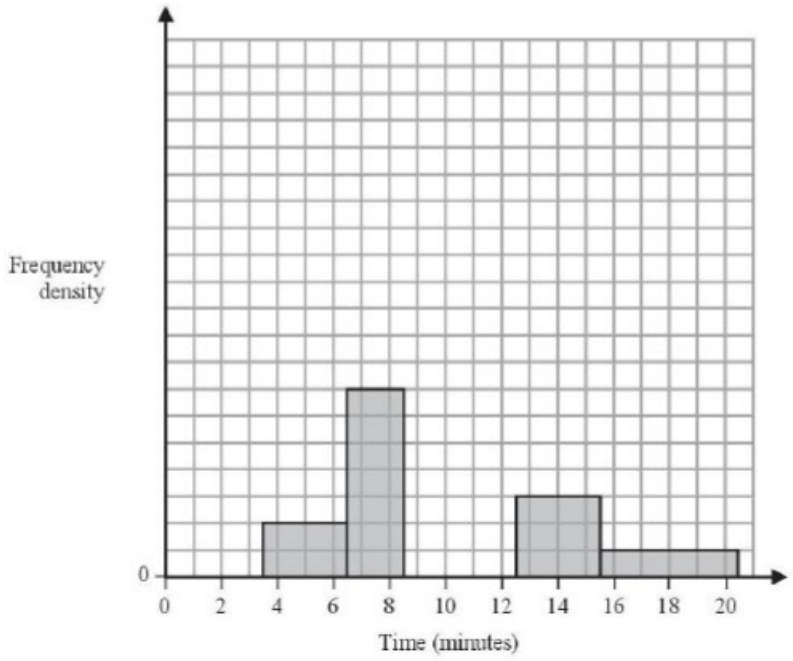
(c) Calculate the standard deviation.

(2)

(Total for question = 4 marks)

Q2.

The partially completed histogram and the partially completed table show the time, to the nearest minute, that a random sample of motorists were delayed by roadworks on a stretch of motorway.



Delay (minutes)	Number of motorists
4 – 6	6
7 – 8	
9	17
10 – 12	45
13 – 15	9
16 – 20	

Estimate the percentage of these motorists who were delayed by the roadworks for between 8.5 and 13.5 minutes.

(5)

(Total for question = 5 marks)

Q3.

Sara was studying the relationship between rainfall, *r* mm, and humidity, *h* %, in the UK. She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

She obtained the following results.

h	93	86	95	97	86	94	97	97	87	97	86
r	1.1	0.3	3.7	20.6	0	0	2.4	1.1	0.1	0.9	0.1

Sara examined the rainfall figures and found

$$Q_1 = 0.1 \qquad Q_2 = 0.9 \qquad Q_3 = 2.4$$

A value that is more than 1.5 times the interquartile range (IQR) above Q_3 is called an outlier.

(a) Show that *r* = 20.6 is an outlier.

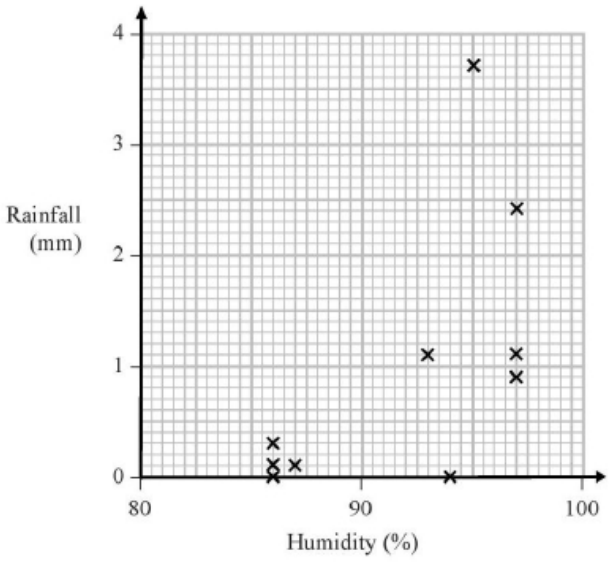
(1)

(b) Give a reason why Sara might

(i) include
(ii) exclude
this day's reading.

(2)

Sara decided to exclude this day's reading and drew the following scatter diagram for the remaining 10 days' values of *r* and *h*.



(c) Give an interpretation of the correlation between rainfall and humidity.

(1)

The equation of the regression line of *r* on *h* for these 10 days is *r* = –12.8 + 0.15*h*

(d) Give an interpretation of the gradient of this regression line.

(1)

(e) (i) Comment on the suitability of Sara's sampling method for this study.

(ii) Suggest how Sara could make better use of the large data set for her study.

(2)

(Total for question = 7 marks)

Q4.

Helen is studying the daily mean wind speed for Camborne using the large data set from 1987.
The data for one month are summarised in Table 1 below.

Windspeed	n/a	6	7	8	9	11	12	13	14	16
Frequency	13	2	3	2	2	3	1	2	1	2

Table 1

(a) Calculate the mean for these data.

(1)

(b) Calculate the standard deviation for these data and state the units.

(2)

The means and standard deviations of the daily mean wind speed for the other months from the large data set for Camborne in 1987 are given in Table 2 below. The data are not in month order.

Month	*A*	*B*	*C*	*D*	*E*
Mean	7.58	8.26	8.57	8.57	11.57
Standard Deviation	2.93	3.89	3.46	3.87	4.64

Table 2

(c) Using your knowledge of the large data set, suggest, giving a reason, which month had a mean of 11.57

(2)

The data for these months are summarised in the box plots on the opposite page.
They are not in month order or the same order as in Table 2.

(d) (i) State the meaning of the * symbol on some of the box plots.

(ii) Suggest, giving your reasons, which of the months in Table 2 is most likely to be summarised in the box plot marked *Y*.

(3)

.....

.....

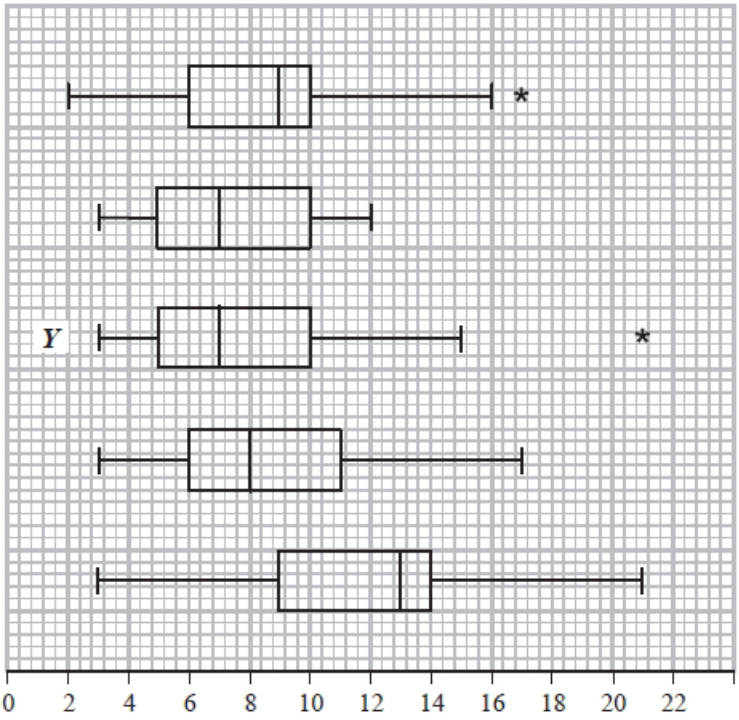
.....

.....

.....

.....

.....



.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

(Total for question = 8 marks)

Q5.

Joshua is investigating the daily total rainfall in Hurn for May to October 2015

Using the information from the large data set, Joshua wishes to calculate the mean of the daily total rainfall in Hurn for May to October 2015

(a) Using your knowledge of the large data set, explain why Joshua needs to clean the data before calculating the mean.

(1)

Using the information from the large data set, he produces the grouped frequency table below.

Daily total rainfall (r mm)	Frequency	Midpoint (x mm)
$0 \leqslant r < 0.5$	121	0.25
$0.5 \leqslant r < 1.0$	10	0.75
$1.0 \leqslant r < 5.0$	24	3.0
$5.0 \leqslant r < 10.0$	12	7.5
$10.0 \leqslant r < 30.0$	17	20.0

You may use $\sum fx = 539.75$ and $\sum fx^2 = 7704.1875$

(b) Use linear interpolation to calculate an estimate for the upper quartile of the daily total rainfall.

(2)

(c) Calculate an estimate for the standard deviation of the daily total rainfall in Hurn for May to October 2015

(2)

(d) (i) State the assumption involved with using class midpoints to calculate an estimate of a mean from a grouped frequency table.

(ii) Using your knowledge of the large data set, explain why this assumption does not hold in this case.

(iii) State, giving a reason, whether you would expect the actual mean daily total rainfall in Hurn for May to October 2015 to be larger than, smaller than or the same as an estimate based on the grouped frequency table.

(3)

(Total for question = 8 marks)

Q6.

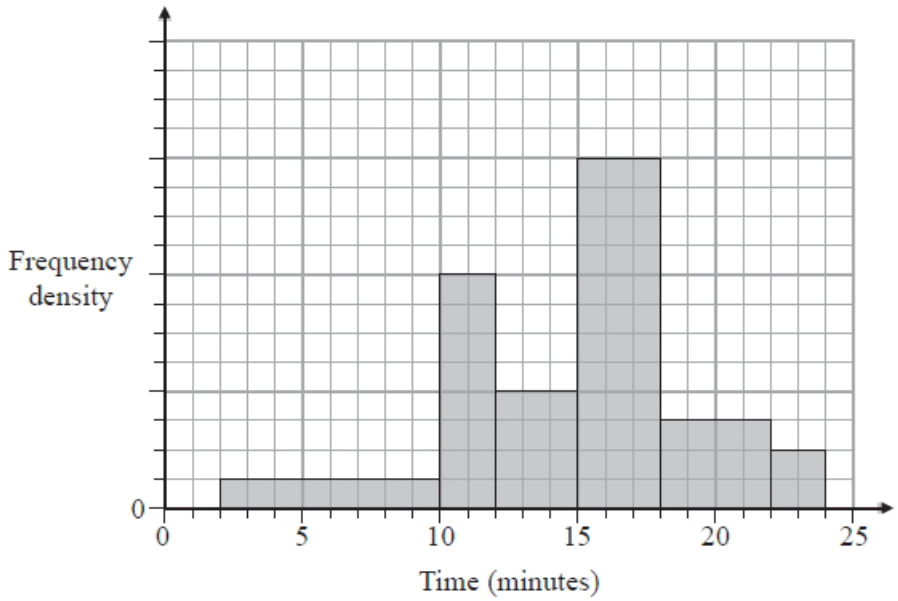


Figure 1

The histogram in Figure 1 shows the times taken to complete a crossword by a random sample of students.

The number of students who completed the crossword in more than 15 minutes is 78.

Estimate the percentage of students who took less than 11 minutes to complete the crossword.

(Total for question = 4 marks)

Q7.

Jerry is studying visibility for Camborne using the large data set June 1987.

The table below contains two extracts from the large data set.

It shows the daily maximum relative humidity and the daily mean visibility.

Date	Daily Maximum Relative Humidity	Daily Mean Visibility
Units	%	
10/06/1987	90	5300
28/06/1987	100	0

(The units for Daily Mean Visibility are deliberately omitted.)

Given that daily mean visibility is given to the nearest 100,

(a) write down the range of distances in metres that corresponds to the recorded value 0 for the daily mean visibility.

(1)

Jerry drew the following scatter diagram, Figure 2, and calculated some statistics using the June 1987 data for Camborne from the large data set.

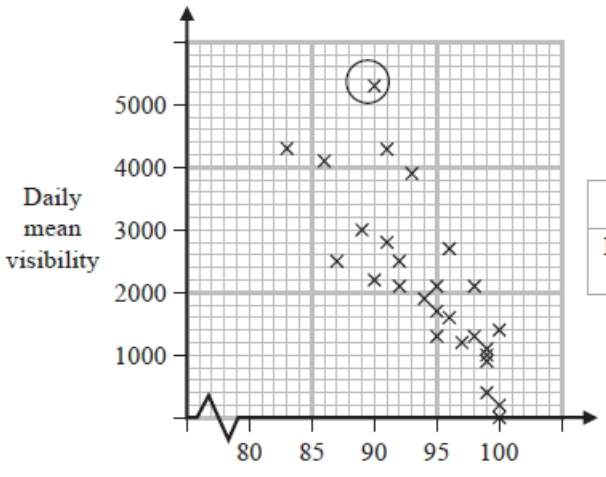


Figure 2

	Q_1	IQR
Daily mean visibility	1100	1600
Daily maximum relative humidity (%)	92	8

Jerry defines an outlier as a value that is more than 1.5 times the interquartile range above Q_3 or more than 1.5 times the interquartile range below Q_1.

(b) Show that the point circled on the scatter diagram is an outlier for visibility.

(2)

(c) Interpret the correlation between the daily mean visibility and the daily maximum relative humidity.

(1)

Jerry drew the following scatter diagram, Figure 3, using the June 1987 data for Camborne from the large data set, but forgot to label the *x*–axis.

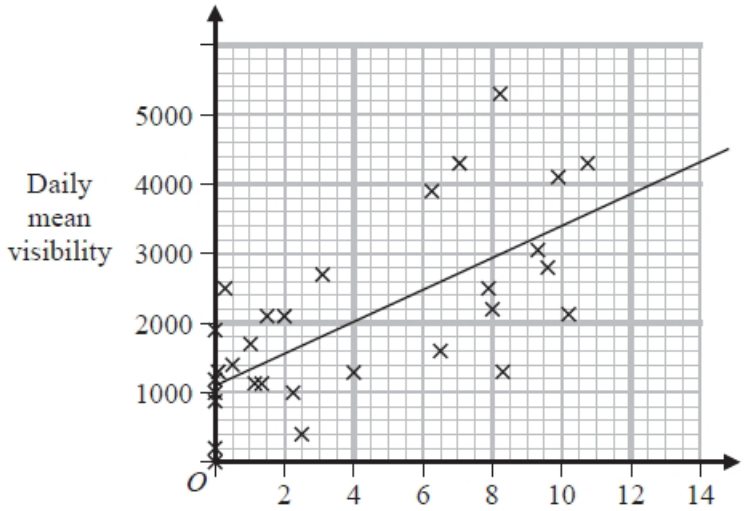


Figure 3

(d) Using your knowledge of the large data set, suggest which variable the *x*-axis on this scatter diagram represents.

(1)

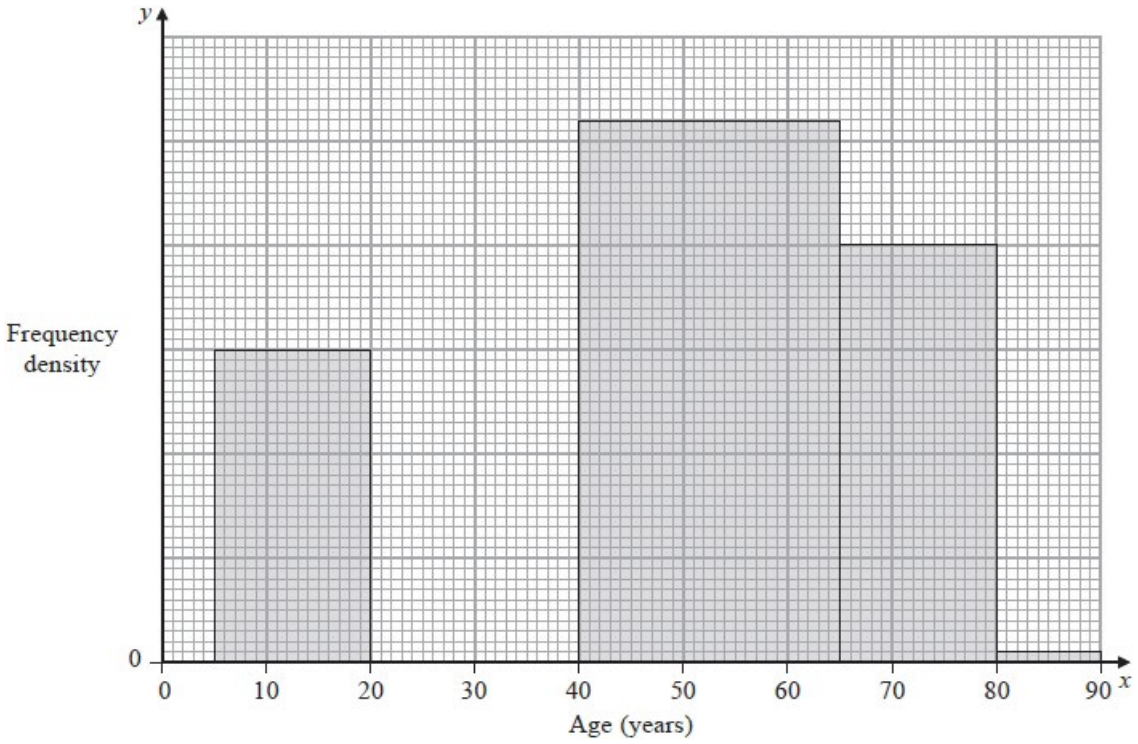
(Total for question = 5 marks)

Q8.

The partially completed table and partially completed histogram give information about the ages of passengers on an airline.

There were no passengers aged 90 or over.

Age (x years)	$0 \leqslant x < 5$	$5 \leqslant x < 20$	$20 \leqslant x < 40$	$40 \leqslant x < 65$	$65 \leqslant x < 80$	$80 \leqslant x < 90$
Frequency	5	45	90			1



(a) Complete the histogram.

(3)

(b) Use linear interpolation to estimate the median age.

(4)

An outlier is defined as a value greater than Q_3 + 1.5 × interquartile range.

Given that Q_1 = 27.3 and Q_3 = 58.9

(c) determine, giving a reason, whether or not the oldest passenger could be considered as an outlier.

(2)

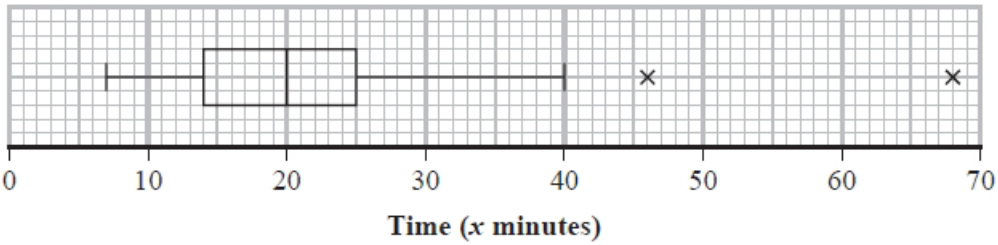
(Total for question = 9 marks)

Q9.

Each member of a group of 27 people was timed when completing a puzzle.

The time taken, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



(a) Find the range of the times.

(1)

(b) Find the interquartile range of the times.

(1)

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17623.25$

(c) calculate the mean time taken to complete the puzzle,

(1)

(d) calculate the standard deviation of the times taken to complete the puzzle.

(2)

Taruni defines an outlier as a value more than 3 standard deviations above the mean.

(e) State how many outliers Taruni would say there are in these data, giving a reason for your answer.

(1)

Adam and Beth also completed the puzzle in a minutes and b minutes respectively, where $a > b$.

When their times are included with the data of the other 27 people

- the median time increases
- the mean time does not change

(f) Suggest a possible value for a and a possible value for b, explaining how your values satisfy the above conditions.

(3)

(g) Without carrying out any further calculations, explain why the standard deviation of all 29 times will be lower

than your answer to part (d).

(1)

(Total for question = 10 marks)

Q10.

Stav is studying the large data set for September 2015

He codes the variable Daily Mean Pressure, x, using the formula $y = x - 1010$

The data for all 30 days from Hurn are summarised by

$$\sum y = 214 \qquad \sum y^2 = 5912$$

(a) State the units of the variable x

(1)

(b) Find the mean Daily Mean Pressure for these 30 days.

(2)

(c) Find the standard deviation of Daily Mean Pressure for these 30 days.

(3)

Stav knows that, in the UK, winds circulate

- in a **clockwise** direction around a region of **high** pressure
- in an **anticlockwise** direction around a region of **low** pressure

The table gives the Daily Mean Pressure for 3 locations from the large data set on 26/09/2015

Location	Heathrow	Hurn	Leuchars
Daily Mean Pressure	1029	1028	1028
Cardinal Wind Direction			

The Cardinal Wind Directions for these 3 locations on 26/09/2015 were, in random order,

W NE E

You may assume that these 3 locations were under a single region of pressure.

(d) Using your knowledge of the large data set, place each of these Cardinal Wind Directions in the correct location in the table.
Give a reason for your answer.

(2)

(Total for question = 8 marks)

Q11.

Charlie is studying the time it takes members of his company to travel to the office.
He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

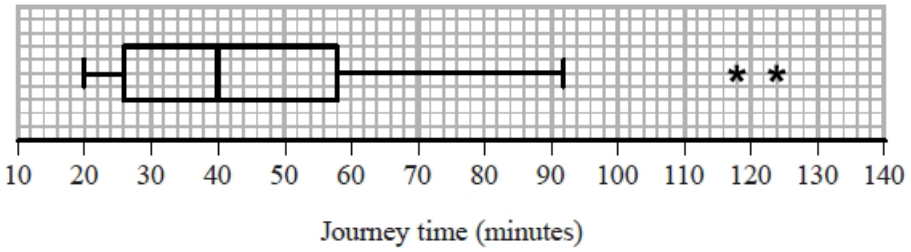
(2)

Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202\,294$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

(Total for question = 13 marks)

Q12.

A lake contains three different types of carp.

There are an estimated 450 mirror carp, 300 leather carp and 850 common carp.

Tim wishes to investigate the health of the fish in the lake.

He decides to take a sample of 160 fish.

(a) Give a reason why stratified random sampling cannot be used.

(1)

(b) Explain how a sample of size 160 could be taken to ensure that the estimated populations of each

type of carp are fairly represented.
You should state the name of the sampling method used.

(2)

As part of the health check, Tim weighed the fish.

His results are given in the table below.

Weight (w kg)	Frequency (f)	Midpoint (m kg)
$2 \leqslant w < 3.5$	8	2.75
$3.5 \leqslant w < 4$	32	3.75
$4 \leqslant w < 4.5$	64	4.25
$4.5 \leqslant w < 5$	40	4.75
$5 \leqslant w < 6$	16	5.5

(You may use $\sum fm = 692$ and $\sum fm^2 = 3053$)

(c) Calculate an estimate for the standard deviation of the weight of the carp.

(2)

Tim realised that he had transposed the figures for 2 of the weights of the fish.

He had recorded in the table 2.3 instead of 3.2 and 4.6 instead of 6.4.

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 3.2 instead of 2.3
(ii) using the correct figure of 6.4 instead of 4.6

would have on your estimated standard deviation.
Give a reason for each of your answers.

(2)

(Total for question = 7 marks)

Mark Scheme

Q1.

Question	Scheme	Marks	AOs
(a)	Systematic (sample)	B1cao	1.2
(b)	In LDS some days have gaps because the data was not recorded	B1	2.4
(c)	$\left[\bar{t} = \frac{374}{20} = 18.7\right]$ $\sigma_t = \sqrt{\frac{7600}{20} - \bar{t}^2}$ $[=\sqrt{30.31}]$	M1	1.1a
	$= 5.5054\ldots$ awrt **5.51** (Accept use of $s_t = \sqrt{\frac{7600 - 20\bar{t}^2}{19}}$ $= 5.6484\ldots$)	A1	1.1b
			(4 marks)

Part	Notes
(b)	B1 a correct explanation
(c)	M1 for a correct expression for $\bar{t}$ and σ_t or s_t. Ft an incorrect evaluation of $\bar{t}$
	A1 for σ_t = awrt 5.51 or s_t = awrt 5.65

Q2.

Question	Scheme	Marks	AOs
	$17 + 45 + \frac{1}{3}\times 9$ $[= 65]$	M1	2.2a
	$(7-8)$ **14** or $(16-20)$ **5** [Values may be seen in the table]	M1 A1	3.1a 1.1b
	Percentage of motorists is $\frac{"65"}{6+"14"+17+45+9+"5"}\times 100$	M1	3.1b
	$=$ **67.7%**	A1	1.1b
			(5 marks)

Part	Notes
	1st M1 for a fully correct expression for the number of motorists in the interval
	2nd M1 for clear use of frequency density in (4-6) or (13-15) cases to establish the fd scale. Then use of area to find frequency in one of the missing cases.
	1st A1 for both correct values seen
	3rd M1 for realising that total is required and attempting a correct expression for %
	2nd A1 for awrt 67.7%

Q3.

Question	Scheme	Marks	AOs
(a)	IQR = 2.3 and $20.6 \gg 2.4 + 1.5\times 2.3$ (= 5.85) (Compare correct values)	B1	1.1b
		(1)	
(b)(i)	e.g. it is a piece of data and we should consider all the data (o.e.)	B1	2.4
(ii)	e.g. it is an extreme value and could unduly influence the analysis or it could be a mistake	B1	2.4
		(2)	
(c)	e.g. "as humidity increases rainfall increases"	B1	2.2b
		(1)	
(d)	e.g. a 10% increase in humidity gives rise to a 1.5 mm increase in rainfall or represents 0.15mm of rainfall per percentage of humidity	B1	3.4
		(1)	
(e)(i)	Not a good method since only uses 11 days from one location in one month.	B1	2.4
(ii)	e.g. She should use data from more of the UK locations and more of the months or using a spreadsheet or computer package she could use all of the available UK data	B1	2.4
		(2)	
		(7 marks)	

Part	Notes
(a)	B1 for sight of the correct calculation and suitable comparison with 20.6
(b)(i)	B1 for a suitable reason for including the data point
(ii)	B1 for a suitable reason for excluding the data point
(c)	B1 for a suitable interpretation of positive correlation mentioning humidity and rainfall
(d)	B1 for a suitable description of the rate: rainfall per percentage of humidity including reference to values.
(e)(i)	B1 for a comment that supports the idea that her sampling method was not a good one
(ii)	B1 for some sensible suggestions that would give a better representation of the data across the UK. Must show some awareness of the fact that LDS has different locations and more months of data available but must be clear they are NOT using any overseas locations. NB B0 for a comment that says use more than one location without specifying that only UK locations are required.

Q4.

Qu	Scheme	Marks	AO
(a)	$\bar{x} = 10.2\ (2222\ldots)$ awrt **10.2**	B1 (1)	1.1b
(b)	$\sigma_x = 3.17\ (20227\ldots)$ awrt **3.17**	B1ft	1.1b
	Sight of "knots" or "kn" (condone knots/s etc)	B1 (2)	1.2
(c)	October ….. since	B1	2.2b
	it is windier in the autumn or month of the hurricane or latest month in the year	B1 (2)	2.4
(d)(i)	They represent outliers	B1	1.2
(ii)	*Y* has low median so expect lowish mean (but outlier so > 7) and	M1	2.4
	Y has big range/IQR or spread so expect larger st.dev Suggests *B*	A1 (3)	2.2b
		(8 marks)	

	Notes
NB	$\bar{x} = \frac{184}{18}$ and $\sigma_x = \sqrt{\frac{2062}{18} - \bar{x}^2}$
(a)	B1 for $\bar{x} = 10.2$ (allow exact fraction)
(b)	1st B1ft allow 3.2 from a correct expr' accept $s = 3.26(3984\ldots)$ [ft use of n/a] Treating n/a as 0 May see $n = 31$ or $\bar{x} = 5.9354\ldots$ which is B0 in (a) but here in (b) it gives $\sigma_x = 5.59(34\ldots)$ or $s = 5.6858\ldots$(awrt 5.69) and scores 1st B1 2nd B1 accept kn accept in (a) or (b) (allow nautical miles/hour)
(c)	1st B1 choosing October but accept September. 2nd B1 for stating that (Camborne) is windier in autumn/winter months "because it is winter/autumn/windier/colder in "month" " Sep ⩽ "month" ⩽ Mar scores B1B1 for "month" = Sep or Oct and B0B1 for other months in range
(d)(i)	B1 for outlier or the idea of an extreme value allow "anomaly"
(ii)	M1 for a comment relating to location that mentions both median and mean and a comment relating to spread that mentions both range/IQR and standard deviation and leads to choosing *B*, *C* or *D* **Choosing *A* or *E* is M0** Incorrect/false statements score M0 e.g. Q_3 = (mean + σ) or identify Q_2 = mean or *Y* has small spread
ALT	**Use of outliers:** outlier is (mean + 3σ) (*B* = 19.9), (*C* = 18.95), (*D* = 20.2) Must see at least one of these values and compare to *Y*'s outlier[leads to *D* or *B*]
	A1 for suitable inference i.e. *B* (accept *D* or *B* or *D*) M1 **must** be scored

Q5.

Question	Scheme	Marks	AOs
(a)	Tr(ace) (data needs to be converted to numbers before the calculation can be carried out)	B1	2.4
		(1)	
(b)	$[1+]\frac{138-131}{24}\times 4$	M1	2.1
	$= 2.1666\ldots$ awrt **2.17**	A1	1.1b
		(2)	
(c)	$\sigma = \sqrt{\frac{7704.1875}{184} - \left(\frac{539.75}{184}\right)^2} = 5.7676...$ σ = awrt **5.77**	M1 A1	1.1b 1.1b
		(2)	
(d)(i)	Using class midpoints to estimate the mean assumes that the values are uniformly distributed **within the class(es)**.	B1	2.4
(ii)& (iii)	This is not the case here as the majority of the data (in the first class) are 0.	B1	2.3
	The actual mean is likely to be smaller than the estimate (since the first group has more values at 0 and close to 0)	dB1	2.2b
		(3)	
			(8 marks)

Notes	
(a)	**B1:** Identifying tr(ace) data Ignore comments about n/a, missing data, anomalies, etc.
(b)	**M1:** Correct fraction $\frac{7}{24}\times 4$ allow working down $[5]-\frac{155-138}{24}\times 4$ allow a correct equation leading to a correct fraction e.g. $\frac{x-1}{5-1}=\frac{138-131}{155-131}$ for M1 Use of $(n+1)$ with 138.75 allow $\frac{7.75}{24}\times 4$ **A1:** awrt 2.17 (condone $\frac{13}{6}$) awrt 2.29 from $(n+1)$ (condone $\frac{55}{24}$)
(c)	**M1:** Correct expression for standard deviation (allow mean = awrt 2.93) **A1:** awrt 5.77 correct answer only scores M1A1 (allow $s = 5.78$) **SC:** 5.76 with no working scores M1A0
(d)(i)	**B1:** Explaining that data assumed to be spread evenly across each class (o.e.) e.g. The midpoint of each class is the mean of each class or all the values in the class are located at the midpoint condone normally distributed within each class
Mark together (ii)&(iii)	**B1:** Demonstrating an understanding of the LDS that the majority of data values (in the first class) are at 0 or close to 0 (trace). **dB1:** (dependent upon 2nd B1) Correct inference based on knowledge of the LDS **SC:** If B1 is scored in (i) for 'The data are spread evenly across each class,' then in (ii) 'The data are not evenly distributed in the classes' scores B1 but in (iii) 'the actual mean is smaller' with no further justification scores B0

Q6.

Question	Scheme	Marks	AOs
	1 square is $\frac{78}{12\times3+3\times4+2\times2}=\left[\frac{78}{52}=1.5\right]$ **and** $(8\times1+1\times8)\times"1.5"$	M1	3.1a
	24 students took less than 11 minutes	A1	1.1b
	Percentage of students $=\frac{"24"}{78+"24"+1\times8\times"1.5"+3\times4\times"1.5"}\times100$	M1	3.1b
	$=$ 18.18… awrt 18%	A1	1.1b
		(4)	
			Total 4

Notes		
	M1:	For clear use of frequency density to establish the fd scale and then use the area to find frequency of <11 minutes. Allow maximum of 3 errors in either the heights or widths in total if working shown. They may calculate the area using other size squares. Allow for realising they need to find the total number of squares (88) maximum of 4 errors in either the heights or widths and number < 11 minutes(16) - must have a maximum of 1 error in either the heights or widths (and not use the 78 as part of calulation)
	A1:	For correct values seen. Allow for 88 and 16
	M1:	For realising the need to find the total and calculating a percentage. (with "their 24" as the numerator). Allow $(8\times1+2\times8)\times"1.5"$ instead of $"24"+1\times8\times"1.5"$ If working shown can allow maximum of 2 errors in either the heights or widths in the calculation of the total. Allow "their 24" / 132 oe
	A1:	awrt 18

Q7.

Question	Scheme	Marks	AOs
(a)	0 to 500 m	B1	1.2
		(1)	
(b)	$1100+1600+1.5\times1600$ [= 5100]	M1	2.1
	5300 > 5100 therefore outlier	A1	1.1b
		(2)	
(c)	As the humidity increases the mean visibility decreases	B1	2.4
		(1)	
(d)	(Hours of) sunshine	B1	2.2b
		(1)	
			(5 marks)

Notes		
(a)	**B1:**	For realising it is the maximum distance and distance given with correct units. Allow 0 to 50dm or < 500m or < 50dm
(b)	**M1:**	Attempt to find Q_3 and the upper limit
	A1:	5100, if a value for the point is stated it must be above 5100 otherwise it is A0. For a statement comparing and conclusion it is an outlier or it is above Q_3 +1.5IQR. Allow accept the point circled is greater than 5100 oe
(c)	**B1:**	For a suitable interpretation of a negative correlation mentioning humidity and visibility
(d)	**B1:**	A correct deduction that the unlabelled variable is the hours of sunshine. Condone missing hours. Do not allow if more than one variable given. Must be quantative variable Not cloud cover since values bigger than 8 Not wind speed since values not integers Not daily mean temperature since mean temperature near to zero are unlikely in June

Q8.

Qu	Scheme	Marks	AO
(a)	From [5,20) fd = 3 or 1 large square = 2.5 passengers o.e.	M1	2.2a
	Correct bar above [0, 5)	A1	1.1b
	Correct bar above [20, 40)	A1	1.1b
		(3)	
(b)	For [40, 65) **130** passengers or for [65, 80) **60** passengers	M1	2.1
	For attempt to find total number of passengers = **331**	A1ft	1.1b
	[Median =] $40+\frac{\frac{1}{2}("331")-140}{"130"}\times 25$ or $65-\frac{270-\frac{1}{2}("331")}{"130"}\times 25$ (o.e.)	M1	1.1b
	= 44.9038… = awrt **44.9**	A1	1.1b
		(4)	
(c)	Upper outlier limit = 58.9 + 1.5 × (58.9 – 27.3) = 106 (.3) > 90	M1	2.4
	So oldest passenger is not an outlier	A1	2.2a
		(2)	
		(9 marks)	

	Notes
(a)	M1 for attempt at fd or a suitable method to deduce the scale for the histogram May be implied by one correct bar. 1st A1 for first bar [0, 5) with fd = 1 or 2 large squares high 2nd A1 for third bar with fd = 4.5 or 9 large squares high
(b)	1st M1 for an attempt using their fd to find the missing frequencies. May be in table 1st A1ft for a clear attempt to find the total number of passengers (ft their 130 and 60) 2nd M1 for any expression/equation leading to correct Q_2 Must be using 40-65 class 2nd A1 for awrt 44.9 (allow $(n + 1)$ leading to 45)
(c)	M1 for finding the upper outlier limit (expression or awrt 106) and stating or implying > 90 A1 dep on M1 seen for deducing NOT an outlier

Q9.

	Scheme	Marks	AO
(a)	[68 – 7 =] **61** (only)	B1 **(1)**	1.1b
(b)	[25 – 14] = **11**	B1 **(1)**	1.1b
(c)	$\left[\mu \text{ or } \bar{x} = \frac{607.5}{27} = \right]$ = **22.5**	B1 **(1)**	1.1b
(d)	$\sigma = \sqrt{\frac{17\,623.25}{27} - "22.5"^2}$ or $\sqrt{146.4629...}$	M1	1.1b
	= 12.10218… awrt **12.1**	A1 **(2)**	1.1b
(e)	$\mu + 3\sigma = "22.5" + 3 \times "12.1..." =$ awrt 59 so only **one** outlier	B1ft **(1)**	1.1b
(f)	Median increases implies that both values must be > 20	M1	3.1b
	Mean is the same means that $a + b = 45$	M1	1.1b
	So possible values are: e.g. $b = 21$ and $a = 24$ (o.e.)	A1 **(3)**	2.2b
(g)	Both values will be less than 1 standard deviation from the mean and so the standard deviation of all 29 values will be smaller	B1 **(1)**	2.4
		(10 marks)	

	Notes
(a)	B1 for correctly interpreting the box plot to find the range (more than 1 answer is B0)
(b)	B1 for correct understanding of IQR and answer of 11
(c)	B1 for 22.5 only (or exact equivalent such as $\frac{45}{2}$). Allow 22 mins and 30 secs.
(d)	M1 for a correct expression including square root. Allow $\sqrt{146}$ or better. Ft their mean A1 for awrt 12.1 NB Allow use of $s = 12.3327\ldots$ or awrt 12.3
(e)	B1ft for a correct calculation or value based on their μ and σ and compatible conclusion
(f)	1st M1 Correct start to the problem and a correct statement about the values based on median Allow if their final two values are both >20 2nd M1 for a correct explanation leading to equation $a + b = 45$ (o.e. e.g. equidistant from mean) Allow if their final two values sum to 45 A1 for a correct pair of values (both > 20 with a sum of 45) **and** at least some attempt to explain how their values satisfy at least one of the conditions (both > 20 or $a + b = 45$). Ignore $a =$ or $b =$ labels **NB** The values for a and b do not need to be integers.
(g)	B1 for a correct explanation. Must mention that both values are less than 1 sd (ft their answer to (d)) from the mean

Q10.

	Scheme	Marks	AO
(a)	Hectopascal or hPa	B1 (1)	1.2
(b)	$\bar{x} = \bar{y} + 1010$ or $\frac{214}{30} + 1010$	M1	1.1b
	= 1017.1333… awrt **1017**	A1 (2)	1.1b
(c)	$\sigma_x = \sigma_y$ (or statement that standard deviation is not affected by this type of coding)	M1	3.1b
	$\left[\sigma_y =\right] \sqrt{\frac{5912}{30} - (\text{"7.13[33...]"})^2}$ or $\sqrt{146.1822...}$	M1	1.1b
	= 12.0905… awrt **12.1**	A1 (3)	1.1b
(d)	High pressure (since approx. mean + sd) so clockwise Locations are (from North to South): Leuchars, Heathrow, Hurn	B1	2.4
	Wind direction is direction wind blows from So: Heathrow (**NE**) Hurn (**E**) Leuchars (**W**)	B1 (2)	2.2a
		(8 marks)	

	Notes
FYI	1 hPa = 100 Pa; 10hPa = 1 kPa; 1Pa = 1 Nm^{-2}
(a)	B1 for "hectopascal" or hPa (condone pascals, allow millibars or mb) o.e. Do NOT allow kPa or kilopascals or Pa on its own
(b)	M1 for a strategy to find $\bar{x}$ Allow an attempt to find $\sum x$ that gets as far as $\sum x = \sum y + 30 \times 1010$ [= 30 514] A1 for awrt 1017 (accept 1020) [Ignore incorrect units]
(c)	1st M1 for an overall strategy using the fact $\sigma_x = \sigma_y$ (can be implied by correct final ans) or for $\sum x = 30\,514$ and $\sum x^2 = 31\,041\,192$ (both seen and correct) 2nd M1 for a correct expression (with $\sqrt{}$)(ft their $\bar{y}$ to 3sf) allow awrt 146 for 146.1822.. or for correct expression in x can ft their $\sum x > 30\,000$ or their answer to (b) A1 (dep on 2nd M1) for awrt 12.1 [Ignore incorrect units]
Final answer	Final ans of awrt 12.1 scores 3/3 **but** if they then adjust for x e.g. add 1010 (M0M1A1)
(d)	1st B1 for at least one of these reasons (these 2 lines) clearly stated (may see diagram) Need "high pressure" **and** "clockwise" to score on 1st line Contradictory statements B0 e.g. correct N~S list but say "anticlockwise" 2nd B1 (indep of 1st B1) for deducing the 3 correct directions either in the table or stated as above If the answers in table and text are different we take the table (as question says)

Q11.

Qu	Scheme	Marks	AO
(a)	Convenience or opportunity [sampling]	B1 (1)	1.2
(b)	Quota [sampling] e.g. Take 4 people every 10 minutes	B1 B1 (2)	1.1a 1.1b
(c)	Census	B1 (1)	1.2
(d)	[58 – 26 =] **32** (min)	B1 (1)	1.1b
(e)	$\mu = \frac{4133}{95} = 43.505263\ldots$ **awrt 43.5 (min)**	B1	1.1b
	$\sigma_x = \sqrt{\frac{202\,294}{95} - \mu^2} = \sqrt{236.7026\ldots}$	M1	1.1b
	= 15.385… **awrt 15.4 (min)**	A1 (3)	1.1b
(f)	There are outliers in the data (or data is skew) which will affect mean and sd Therefore use median and IQR	B1 dB1 (2)	2.4 2.4
(g)	Value of 20, LQ at 26 and outliers will not change or state that median and upper quartile are the values that do change More values now below 40 than above so Q_2 or Q_3 will change and be lower Both Q_2 and Q_3 will be lower	B1 M1 A1 (3)	1.1b 2.1 2.4
		(13 marks)	

	Notes
(b)	1st B1 for quota (sampling) mentioned ("Stratified" or "systematic" or "random" are B0B0) 2nd B1 for a description of how such a system might work, requires suitable strata or categories e.g. time slots, departments, gender, age groups, distance travelled etc Suggestion of randomness is B0
(e)	B1 for a correct mean (awrt 43.5) M1 for a correct expression for the sd (including $\sqrt{}$)ft their mean A1 for awrt 15.4 (Allow s = 15.4667… awrt 15.5)
(f)	1st B1 for acknowledging outliers or skewness are a problem for mean and sd "extreme values"/"anomalies" OK May be implied by saying median and IQR not affected by.. We need to see mention of "outliers", "skewness" and the problem so "data is skewed so use median and IQR" is B0 unless mention that they are not affected by extreme values or mean and standard deviation can be "inflated" by the positive skew etc 2nd dB1 dep on 1st B1 for therefore choosing median and IQR
(g)	B1 for identifying 2 of these 3 groups of unchanged values or stating only Q_2 and Q_3 change M1 for explaining that median or UQ should be lower. E.g. the 2 values have moved to below 40 (or 58) and therefore more than 50% below 40 or (more than 75% below 58) or an argument to show that the other 3 values are the same. (o.e.) Allow arrows on box plot provided statement in words about increased % below 40 or 58 etc A1 for stating median and UQ are both lower with clear evidence of M1 scored [If lots of values on 40 then median might not change but, since two values do change then UQ would change. If this meant that 92 became an outlier then we would have a new value for upper whisker and an extra outlier so effectively 3 values are altered. So median changes]

Q12.

Question	Scheme	Marks	AOs
(a)	It is not possible to have a sampling frame	B1	2.3
		(1)	
(b)	Quota sampling **and** (catch 85 common carp, 45 mirror carp and 30 leather carp) **or** (ignore any fish caught of a type where the quota is full)	M1	1.1a
	Quota sampling **and** catch 85 common carp, 45 mirror carp and 30 leather carp **and** ignore any fish caught of a type where the quota is full	A1	1.1b
		(2)	
(c)	$\sigma = \sqrt{\frac{3053}{160} - \left(\frac{692}{160}\right)^2}$	M1	1.1b
	$= 0.6129\ldots$ awrt 0.613	A1	1.1b
		(2)	
(d)(i)	This would have no effect as the piece of data would remain in the same class	B1	2.2a
(ii)	This would increase the standard deviation as change in mean is small and $6.4 - 4.6 \approx 3\sigma$ therefore estimate of standard deviation will increase	B1	2.2a
		(2)	
			(7 marks)

Notes		
(a)	**B1:**	For the idea there cannot be a sampling frame/list
(b)	**M1:**	Quota sampling **and** either for the correct numbers of each type **or** for the idea that if quota full ignore the fish.
	A1:	Quota sampling **and** both the correct numbers of each type **and** for the idea that if quota full ignore the fish or sample until all quotas are full
(c)	**M1:**	A correct expression for σ
	A1:	Awrt 0.613 allow s = awrt 0.615
(d)	**B1:**	Correct deduction with suitable explanation Allow range for class. Do not allow there is no differences
	B1:	Correct deduction with suitable explanation. so would increase the standard deviation and a suitable reason. Allow the value is bigger than any others in the table **oe**