SECTION A: STATISTICS

Answer ALL questions. Write your answers in the spaces provided.

1. 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)

(a) Sara used systematic sampling

(b) In the Large Data Set, there are some days with no recorded data, so this process may not generate a sample of size 20.

(c) $\sigma_t = \sqrt{\frac{\Sigma t^2}{n} - \bar{t}^2}$

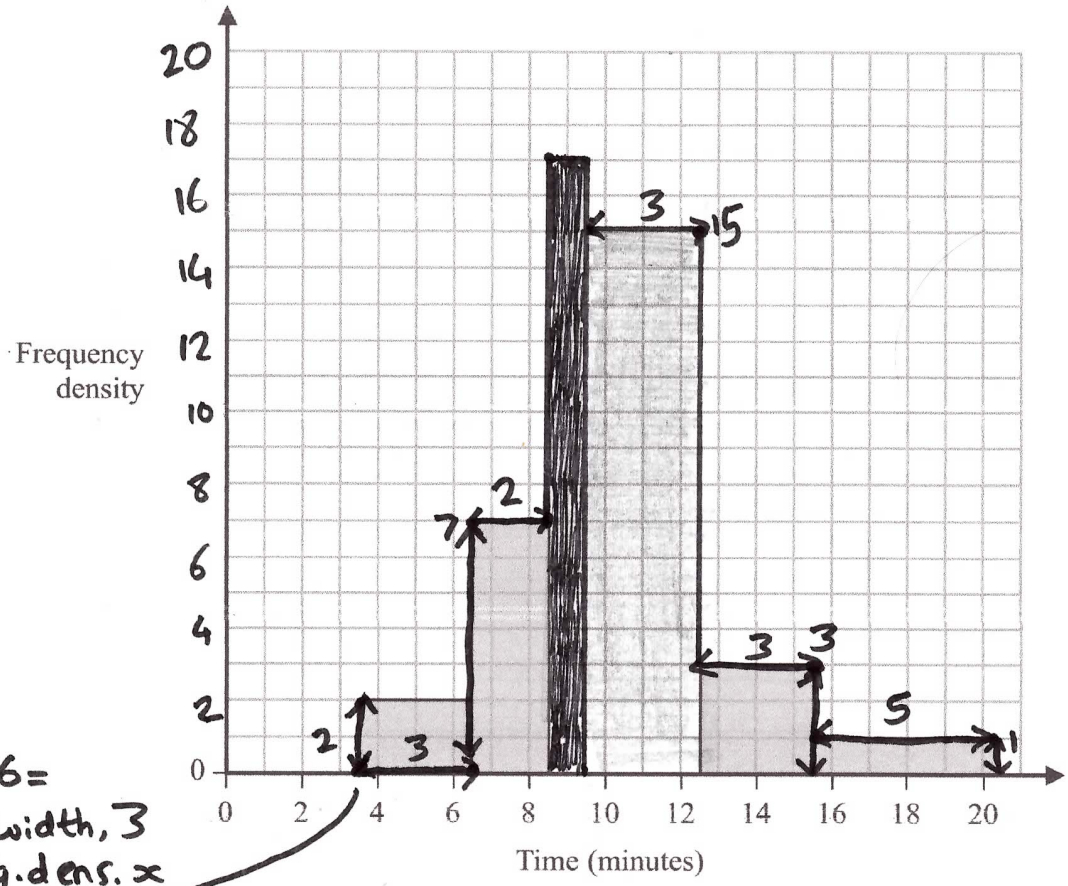
$Q_t = \sqrt{\frac{7600}{20} - \left(\frac{\Sigma t}{n}\right)^2}$

$Q_t = \sqrt{380 - \left(\frac{374}{20}\right)^2}$

$Q_t = \sqrt{380 - 18.7^2}$

$\boxed{Q_t = 5.51 \text{ (to 3s.f.)}}$

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



freq, 6= class width, 3 x freq.dens. x

6= 3x

x=2

Delay (minutes)	Number of motorists
4 – 6	6
7 – 8	14
9	17
10 – 12	45
13 – 15	9
16 – 20	5

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

(5)

frequency = class width x frequency density

Use the freq.density values to fill the table.

Question 2 continued

Between 8.5 and 13.5 minutes:

$$\text{Frequency} = 17 + 45 + \left(\tfrac{1}{3} \times 9\right)$$

$$= 17 + 45 + 3$$

$$= 65$$

← The class width is 3, but we only require the values for 1, so divide by 3

% of motorists between 8.5 and 13.5 minutes:

$$\% = \left(\frac{65}{6+14+17+45+9+5}\right) \times 100$$

$$= \frac{65}{96} \times 100$$

$$= \boxed{67.7\% \text{ (to 3s.f.)}}$$

(Total for Question 2 is 5 marks)

3. 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:

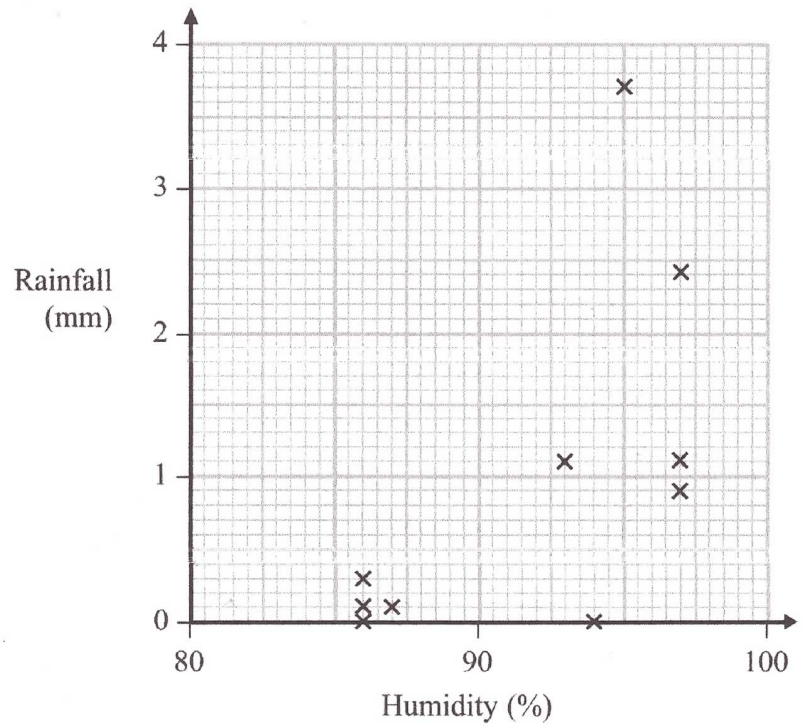
(i) include

(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)

Question continued

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

(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)

(a) IQR = $Q_3 - Q = 2.4 - 0.1$ = 2.3

2.4 + 1.5 × 2.3 = 5.85

r = 20.6 > 5.85, so r = 20.6 is an outlier

(b) (i) She should include it because it is a piece of data and all data should be considered.

(ii) She could exclude it since it is an extreme value and affect the investigation.

(c) As humidity increases, rainfall increases.

(d) The gradient (0.15) represents that there's a 0.15mm increase in rainfall per percentage of humidity.

(e) (i) Sara's sampling method isn't very good since she only uses 11 days out of the whole month, and only uses one specific location.

(ii) Sara could use data from more than one UK location and also use a wider range of months with more days per month.

(Total for Question is 7 marks)

4. 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)

4a) $\bar{x} = \frac{\Sigma fx}{n}$

$= \frac{184}{18}$

$= 10.222$

$\simeq 10.2$

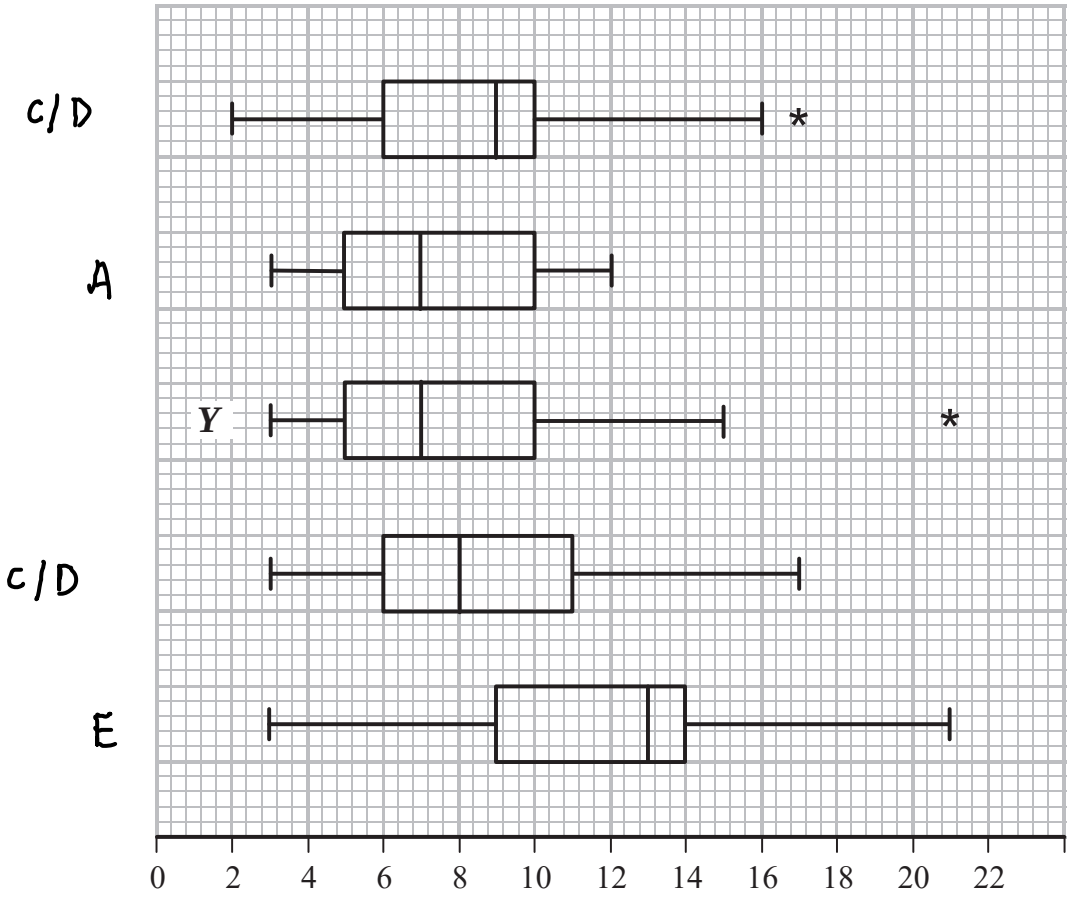
b) $\sigma = \sqrt{\frac{\Sigma fx^2}{n} - \left(\frac{\Sigma fx}{n}\right)^2}$

$= \sqrt{\frac{2062}{18} - \left(\frac{184}{18}\right)^2}$

$= \sqrt{\frac{815}{81}}$

$\simeq 3.17$ knots

Question 4 continued



c) October as it is the latest month included in the dataset.

di) outlier

ii) Y has a low median and big range so the mean should be low and the standard deviation is large.

∴ B

(Total for Question 4 is 8 marks)

5. 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)

a. Trace data must be converted to numbers to allow calculations to be carried out.

b. $1 + \frac{138 - 131}{24} \times 6 = 2.17$

Question continued

c. $\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2} = \sqrt{\frac{77061875}{184} - \left(\frac{539.75}{184}\right)^2}$

$= 5.77$

d. i. Using midpoints assumes the data is distributed uniformly throughout each class.

ii. Most of the data in the first class are 0.

iii. A true mean is likely to be smaller as the first group has more values at and close to 0 than the calculated mean takes into account.

6.

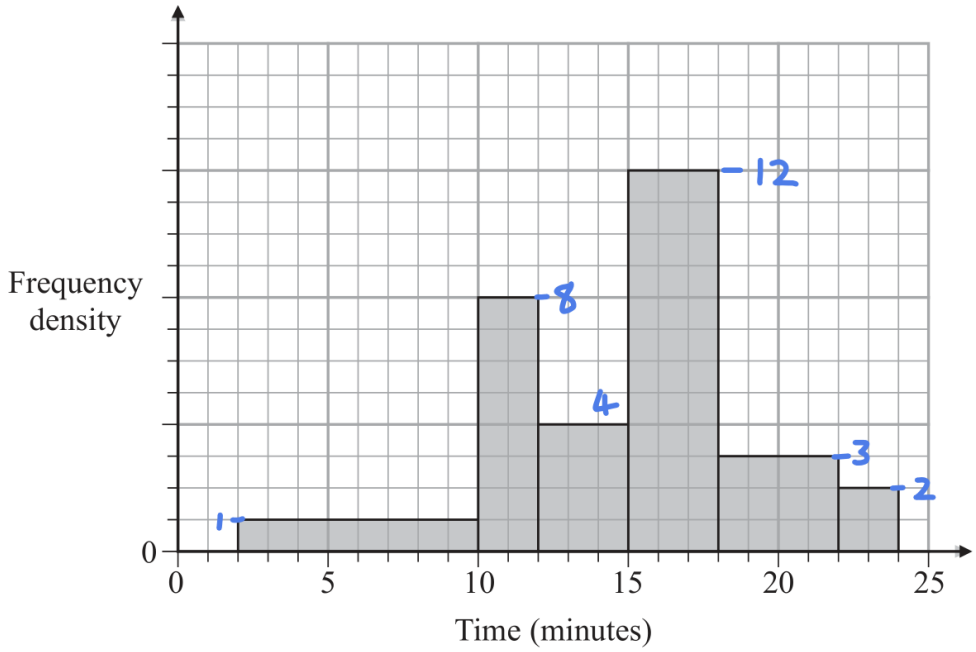


Figure 1

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.

(4)

area above 15 represents 78 students:

$(12\times3+3\times4+2\times2)$ squares $\times k = 78$ students

(k: area scale)

$\therefore$ one square represents $k = \frac{78}{52}$

$= 1.5$

area below 11 $= 8\times1 + 1\times8 = 16$

so $16 \times 1.5 = 24$ students below 11 minutes

BUT we are asked for a percentage so we need the total

Question continued

number of students who did the crossword:

$78 + 24 + 1.5(1 \times 8 + 3 \times 4) = 132$ students

save time - don't recalculate

scale x no. squares

$\Rightarrow$ percentage $= \frac{24}{132} \times 100$

$= 18.181\ldots\%$

$\approx 18\%$

(Total for Question is 4 marks)

7. 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.

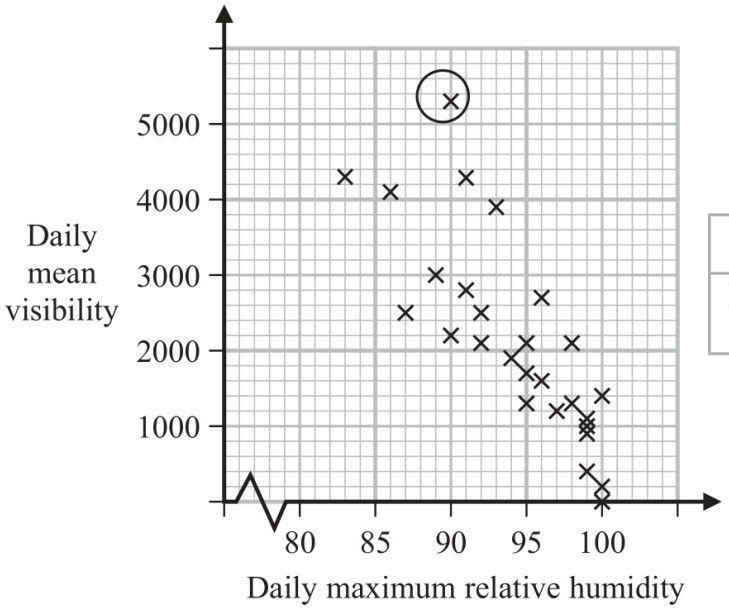


Figure 2

	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.

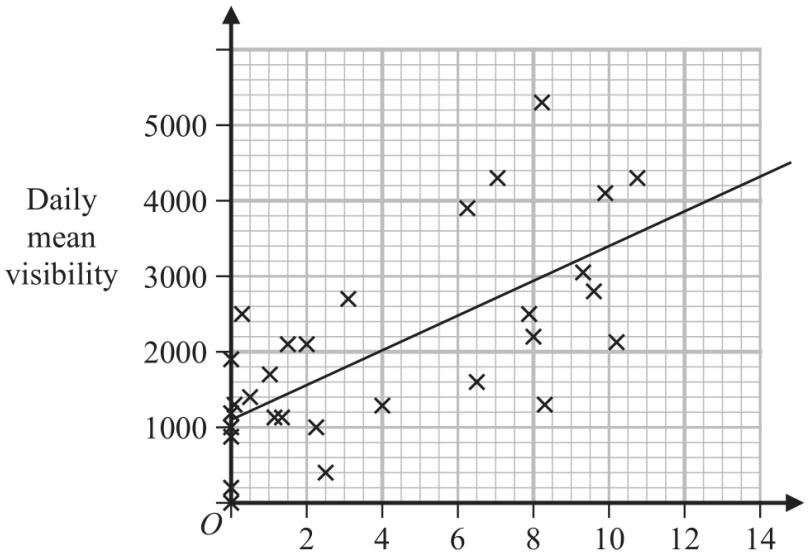


Figure 3

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

familiarise yourself with how the quantities in the set are measured

(1)

a) rounded to nearest hundred decametres

any value 0-50dm rounds down to 0 ⇒ 0 to 500m

b) data point: (90, 5300)

value at which point becomes an outlier:

$Q_3 = Q_1 + IQR = 1100 + 1600 = 2700m$

$Q_3 + 1.5IQR = 2700 + 1.5 \times 1600 = 5100m$

$5300 > 5100$ ⇒ outlier

c) as the humidity increases, the mean visibility decreases

d) we need something that increases visibility

x = hours of sunshine

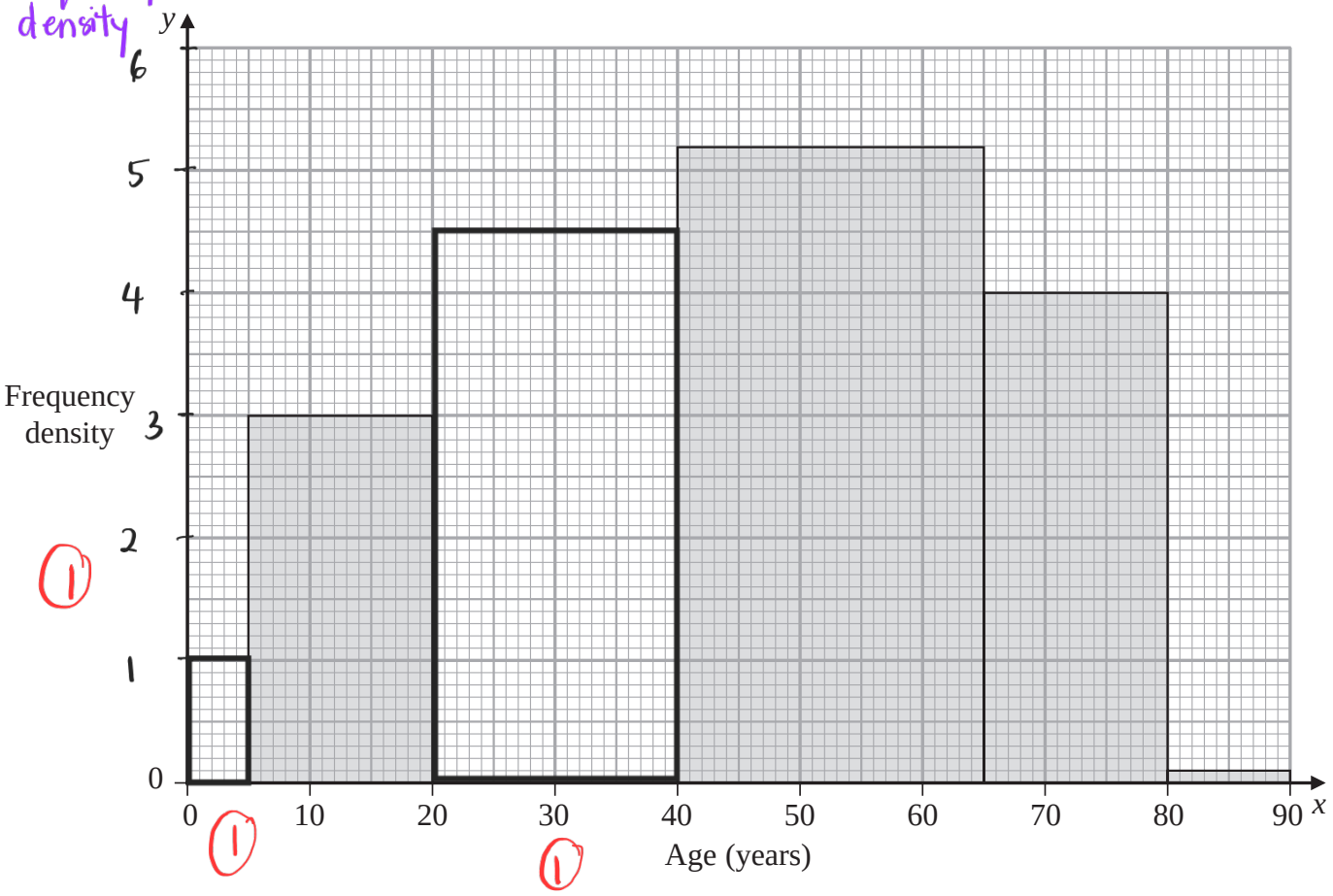
8. The partially completed table and partially completed histogram give information about the ages of passengers on an airline.

There were no passengers aged 90 or over.

Cumulative frequency : 5 50 140 270 330 331

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	130	60	1

Frequency density 1 3 4.5 5.2 4 0.1



(a) Complete the histogram.

(3)

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

(4)

An outlier is defined as a value greater than $Q_3 + 1.5 \times$ 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)

Question continued

(a) Frequency density (FD) = $\frac{\text{Frequency (F)}}{\text{class width (CW)}}$

FD $0 \leqslant x < 5$: $\frac{5}{5} = 1$

FD $5 \leqslant x < 20$: $\frac{45}{15} = 3$

FD $20 \leqslant x < 40$: $\frac{90}{20} = 4.5$

F $40 \leqslant x < 65$: $5.2 \times 25 = 130$

F $65 \leqslant x < 80$: $4 \times 15 = 60$

(b) Total number of passengers : $5 + 45 + 90 + 130 + 60 + 1$
: 331 (1)

Position of median = $\frac{1}{2} \times 331 = 165.5$

look at the cumulative frequency. which interval does the 165.5^{th} value lie?

median value is in the interval $40 \leqslant x < 65$

$$\text{median} = l + \frac{\left(\frac{n}{2} - cf\right)}{f} h$$

l = lower limit of median class
cf = cumulative frequency of class preceeding the median class
h = median class size
f = frequency of median class
n = total number of passengers

Question continued

$$\text{median} = l + \left(\frac{\frac{n}{2} - cf}{f}\right)h$$

$$= 40 + \left(\frac{\frac{331}{2} - 140}{130}\right) \times 25 \quad (1)$$

$$= 44.9038$$

$$\approx 44.9 \ * \quad (1)$$

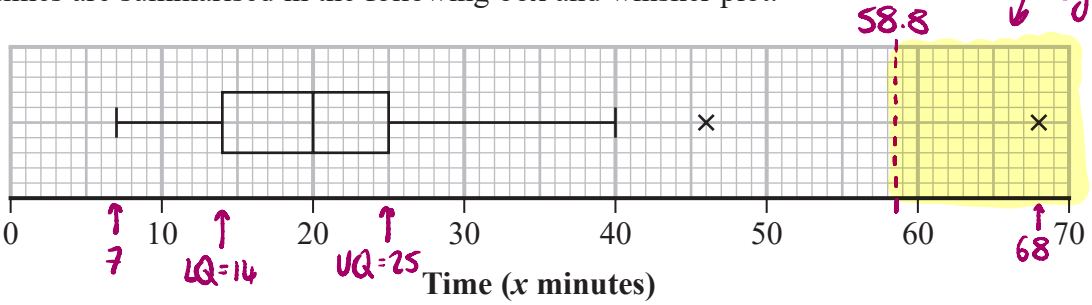
(c) upper outlier limit $= 58.9 + 1.5(58.9 - 27.3) = 106.3$ (1)

Since $106.3 > 90$, the oldest passenger is not an outlier * (1)

9. 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.



Anything in this region classed as an outlier

(a) Find the range of the times.

Range = Highest value − Lowest value $68-7=61$ ① **(1)**

(b) Find the interquartile range of the times.

$IQR = UQ - LQ = 25-14 = 11$ ① **(1)**

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17\,623.25$

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

$\bar{x}$ represents "the mean"

$\bar{x} = \frac{607.5}{27} = 22.5$ ① **(1)**

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

$\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x}^2)} = \sqrt{\frac{17623.25}{27} - (22.5^2)}$ ① $= 12.10218... = 12.1$ (1dp) ① **(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.

$\bar{x} + 3\sigma = 22.5 + 3(12.10218...) = 58.8$ (1dp) ∴ only one outlier ① **(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)

g) Median increases implies that both values must be > 20 ①

Since current median is 20

let y = new total sum of times

$\frac{y}{29} = 22.5$ ← because we know mean doesn't change

$y = 652.5$

new total sum of times = 652.5

652.5 − 607.5 = 45 ∴ a + b = 45 (1)

"old sum"

a + b = 45
a > 20, b > 20
a > b

Possible values could be: a = 24 b = 21 (1)

g) $\bar{x} = 22.5$ $\sigma = 12.1$

22.5 + 12.1 = 34.6
22.5 − 12.1 = 10.4

because of conditions on a, b impossible for either to be outside
10.4 < a < 34.6 and 10.4 < b < 34.6
(so both values less than 1 standard deviation from mean) ← so values "less spread out" so smaller standard deviation

Both values will be less than 1 standard deviation from the mean and so the standard deviation of all values will be smaller (1)

10. 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.

$n = 30$

(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	NE	E	W

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)

a) Hectopascal (OR hPa) (1)

b) $\bar{x} = \bar{y} + 1010$

$= \frac{214}{30} + 1010$ (1)

$= 1017.133..$

$= 1017$ (4 s.f) (1)

$y = x - 1010 \xrightarrow{+1010} x = y + 1010$

$\Sigma y = 214$ so mean is $\frac{\Sigma y}{n} = \frac{214}{30}$

Question continued.

c) $\sigma_x = \sigma_y$ (1) as σ is not affected by this coding

$\sigma_y = \sqrt{\frac{5912}{30} - 7.13^2}$ (1) using formula $\sigma = \sqrt{\frac{\Sigma y^2}{n} - (\bar{y})^2}$

$= \sqrt{146.182..}$

$= 12.0905.$

$= 12.1$ (3.s f) (1)

d) High pressure, so clockwise. (1)

Locations are (North → South): Leuchars, Heathrow, Hurn

Wind direction is direction wind blows from, so:

Heathrow = NE (1) (for answers in table).

Hurn = E

Leuchars = W

(Total for Question is 8 marks)

11. 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.

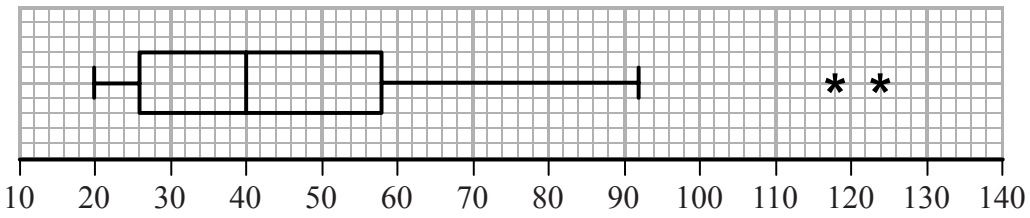
(2)

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.



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

(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)

a) Opportunity sampling. - (1) (or convenience sampling)

b) Quota sampling - Charlie could ask 20 men and 20 women how long their journey was. - (1)

↳ Could've also given 'take 4 people every 10 minutes' for the 2nd mark as another example. (1)

c) A census - (1)

d) IQR = UQ - LQ

IQR = 58 - 26

= 32 - (1)

e) $\bar{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{4133}{95}$

$\bar{x} = 43.5$ (3s.f.) - (1)

$\sigma_x = \sqrt{\left(\frac{\Sigma x^2}{n}\right) - \left(\frac{\Sigma x}{n}\right)^2}$

$\sigma_x = \sqrt{\frac{202294}{95} - \left(\frac{4133}{95}\right)^2}$ - (1)

$\sigma_x = 15.4$ (3s.f.) - (1)

f) Due to outliers in the data, the median and interquartile range should be used - as outliers affect the mean and standard deviation. - (2)

g) The median and upper quartile will change, (1) as there are now more values less than 40 (median) (1) and so the median will decrease with this, as will the upper quartile - (1)

→ For first mark, could have said that the value at 20, the lower quartile at 26, and the outliers will not change.

12. 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)

a) it isn't possible to have a sampling frame

(we don't know the exact number of carp in the lake &

Question continued

don't have a list of all of them)

b) use quota sampling.

estimated total: 1600 → 10x sample size

divide estimated numbers of each type by 10:

catch 85 common, 45 mirror, & 30 leather carp.

ignore any fish caught once the quota for that type is full.

c) $\sigma = \sqrt{\frac{\Sigma fm^2}{n} - \left(\frac{\Sigma fm}{n}\right)^2}$

$= \sqrt{\frac{3053}{160} - \left(\frac{692}{160}\right)^2}$

$= 0.6129...$

we use the frequency in each class.

d) i. the data point stays in the same class, so this would not change the standard deviation.

ii. 116 ~~this~~ is outside the available classes, so does change the mean by a small amount. $6.4 - 4.6 = 1.8 \simeq 3\sigma$ so the estimate of σ will increase.

$\frac{692}{160} = 4.3...$ so 4.6 is close to the mean, 6.4 is far from it