1. A teacher selects a random sample of 56 students and records, to the nearest hour, the time spent watching television in a particular week.

Hours	1–10	11–20	21–25	26–30	31–40	41–59
Frequency	6	15	11	13	8	3
Mid-point	5.5	15.5		28		50

(a) Find the mid-points of the 21–25 hour and 31–40 hour groups.

(2)

A histogram was drawn to represent these data. The 11–20 group was represented by a bar of width 4 cm and height 6 cm.

(b) Find the width and height of the 26–30 group.

(3)

(c) Estimate the mean and standard deviation of the time spent watching television by these students.

(5)

(d) Use linear interpolation to estimate the median length of time spent watching television by these students.

(2)

The teacher estimated the lower quartile and the upper quartile of the time spent watching television to be 15.8 and 29.3 respectively.

(e) State, giving a reason, the skewness of these data.

(2)
(Total 14 marks)

2.   The 19 employees of a company take an aptitude test. The scores out of 40 are illustrated in the stem and leaf diagram below.

|	2|6 means a score of 26	
0	7	(1)
1	88	(2)
2	4468	(4)
3	2333459	(7)
4	00000	(5)

Find

(a)   the median score,

(1)

(b)   the interquartile range.

(3)

The company director decides that any employees whose scores are so low that they are outliers will undergo retraining.

An outlier is an observation whose value is less than the lower quartile minus 1.0 times the interquartile range.

(c)   Explain why there is only one employee who will undergo retraining.

(2)

(d)   On the graph paper below, draw a box plot to illustrate the employees' scores.
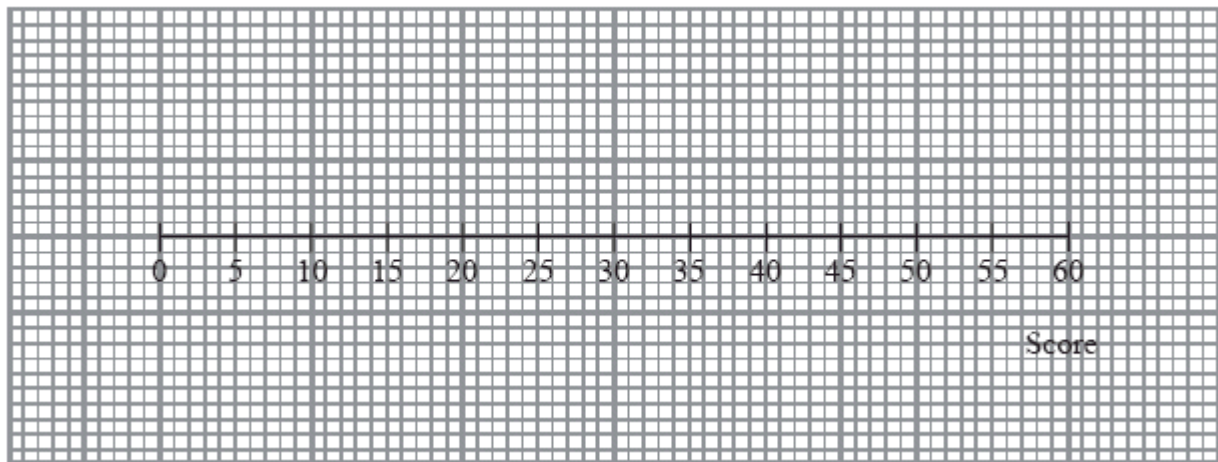
(3)
(Total 9 marks)

3.   The birth weights, in kg, of 1500 babies are summarised in the table below.

Weight (kg)	Midpoint, x kg	Frequency, f
0.0 – 1.0	0.50	1
1.0 – 2.0	1.50	6
2.0 – 2.5	2.25	60
2.5 – 3.0		280
3.0 – 3.5	3.25	820
3.5 – 4.0	3.75	320
4.0 – 5.0	4.50	10
5.0 – 6.0		3

[You may use $\sum \mathrm{f}x = 4841$ and $\sum \mathrm{f}x^2 = 15\ 889.5$]

(a)   Write down the missing midpoints in the table above.

(2)

(b)   Calculate an estimate of the mean birth weight.

(2)

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

(3)

(d)   Use interpolation to estimate the median birth weight.

(2)

(e)   Describe the skewness of the distribution. Give a reason for your answer.

(2)
(Total 11 marks)

4.   There are 180 students at a college following a general course in computing. Students on this course can choose to take up to three extra options.

112 take systems support,
70 take developing software,
81 take networking,
35 take developing software and systems support,
28 take networking and developing software,
40 take systems support and networking,
4 take all three extra options.

(a)   Draw a Venn diagram to represent this information.

(5)

A student from the course is chosen at random.

Find the probability that this student takes

(b)   none of the three extra options,

(1)

(c)   networking only.

(1)

Students who want to become technicians take systems support and networking. Given that a randomly chosen student wants to become a technician,

(d)   find the probability that this student takes all three extra options.

(2)
(Total 9 marks)

5.   The variable x was measured to the nearest whole number. Forty observations are given in the table below.

x	10 – 15	16 – 18	19 –
Frequency	15	9	16

A histogram was drawn and the bar representing the 10 – 15 class has a width of 2 cm and a height of 5 cm. For the 16 – 18 class find

(a)   the width,

(1)

(b)   the height

(2)

of the bar representing this class.

(Total 3 marks)

6.   A researcher measured the foot lengths of a random sample of 120 ten-year-old children. The lengths are summarised in the table below.

Foot length, l, (cm)	Number of children
$10 \leq l < 12$	5
$12 \leq l < 17$	53
$17 \leq l < 19$	29
$19 \leq l < 21$	15
$21 \leq l < 23$	11
$23 \leq l < 25$	7

(a)   Use interpolation to estimate the median of this distribution.

(2)

(b)   Calculate estimates for the mean and the standard deviation of these data.

(6)

One measure of skewness is given by

$$\text{Coefficient of skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

(c)   Evaluate this coefficient and comment on the skewness of these data.

(3)

Greg suggests that a normal distribution is a suitable model for the foot lengths of ten-year-old children.

(d)   Using the value found in part (c), comment on Greg's suggestion, giving a reason for your answer.

(2)
(Total 13 marks)

7.   In a study of how students use their mobile telephones, the phone usage of a random sample of 11 students was examined for a particular week.

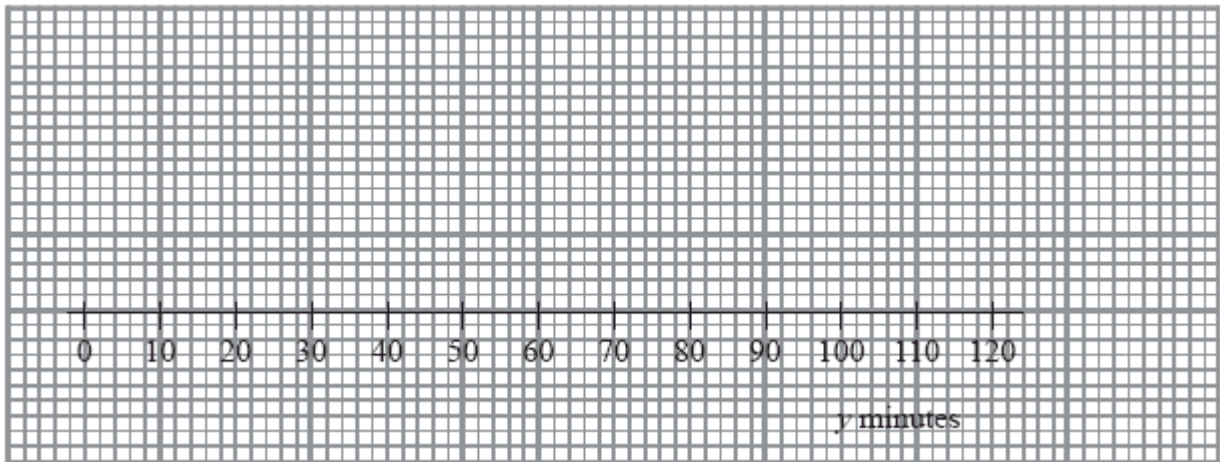The total length of calls, *y* minutes, for the 11 students were

17, 23, 35, 36, 51, 53, 54, 55, 60, 77, 110

(a)   Find the median and quartiles for these data.

(3)

A value that is greater than $Q_3 + 1.5 \times (Q_3 - Q_1)$ or smaller than $Q_1 - 1.5 \times (Q_3 - Q_1)$ is defined as an outlier.

(b)   Show that 110 is the only outlier.

(2)

(c)   Using the graph below draw a box plot for these data indicating clearly the position of the
        outlier.

(3)

The value of 110 is omitted.

(d)   Show that S_{yy} for the remaining 10 students is 2966.9

(3)

These 10 students were each asked how many text messages, x, they sent in the same week.

The values of S_{xx} and S_{xy} for these 10 students are $S_{xx} = 3463.6$ and $S_{xy} = -18.3$.

(e)   Calculate the product moment correlation coefficient between the number of text
        messages sent and the total length of calls for these 10 students.

(2)

A parent believes that a student who sends a large number of text messages will spend fewer
minutes on calls.

(f)   Comment on this belief in the light of your calculation in part (e).

(1)
(Total 14 marks)

8. In a shopping survey a random sample of 104 teenagers were asked how many hours, to the nearest hour, they spent shopping in the last month. The results are summarised in the table below.

Number of hours	Mid-point	Frequency
0 – 5	2.75	20
6 – 7	6.5	16
8 – 10	9	18
11 – 15	13	25
16 – 25	20.5	15
26 – 50	38	10

A histogram was drawn and the group (8 – 10) hours was represented by a rectangle that was 1.5 cm wide and 3 cm high.

(a) Calculate the width and height of the rectangle representing the group (16 – 25) hours.

(3)

(b) Use linear interpolation to estimate the median and interquartile range.

(5)

(c) Estimate the mean and standard deviation of the number of hours spent shopping.

(4)

(d) State, giving a reason, the skewness of these data.

(2)

(e) State, giving a reason, which average and measure of dispersion you would recommend to use to summarise these data.

(2)
(Total 16 marks)

9.   A disease is known to be present in 2% of a population. A test is developed to help determine whether or not someone has the disease.

Given that a person has the disease, the test is positive with probability 0.95

Given that a person does not have the disease, the test is positive with probability 0.03

(a)   Draw a tree diagram to represent this information.

(3)

A person is selected at random from the population and tested for this disease.

(b)   Find the probability that the test is positive.

(3)

A doctor randomly selects a person from the population and tests him for the disease. Given that the test is positive,

(c)   find the probability that he does not have the disease.

(2)

(d)   Comment on the usefulness of this test.

(1)
(Total 9 marks)

10.   The age in years of the residents of two hotels are shown in the back to back stem and leaf diagram below.

Abbey Hotel   8|5|0 means 58 years in Abbey hotel and 50 years in Balmoral hotel Balmoral Hotel

Abbey Hotel			Balmoral Hotel	
(1)	2	0		
(4)	9751	1		
(4)	9831	2	6	(1)
(11)	99997665332	3	447	(3)
(6)	987750	4	005569	(6)
(1)	0	5	000013667	(9)
		6	233457	(6)
		7	015	(3)

For the Balmoral Hotel,

(a)   write down the mode of the age of the residents,

(1)

(b)   find the values of the lower quartile, the median and the upper quartile.

(3)

(c)   (i)   Find the mean, $\bar{x}$, of the age of the residents.

        (ii)   Given that $\sum x^2 = 81213$ find the standard deviation of the age of the residents.

(4)

One measure of skewness is found using

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$$

(d)   Evaluate this measure for the Balmoral Hotel.

(2)

For the Abbey Hotel, the mode is 39, the mean is 33.2, the standard deviation is 12.7 and the measure of skewness is – 0.454

(e)   Compare the two age distributions of the residents of each hotel.

(3)
(Total 13 marks)

11. A person's blood group is determined by whether or not it contains any of 3 substances A, B and C.

A doctor surveyed 300 patients' blood and produced the table below

Blood Contains	No. of Patients
only C	100
A and C but not B	100
only A	30
B and C but not A	25
only B	12
A, B and C	10
A and B but not C	3

(a) Draw a Venn diagram to represent this information.

(4)

(b) Find the probability that a randomly chosen patient's blood contains substance C.

(2)

Harry is one of the patients. Given that his blood contains substance A,

(c) find the probability that his blood contains all 3 substances.

(2)

Patients whose blood contains none of these substances are called universal blood donors.

(d) Find the probability that a randomly chosen patient is a universal blood donor.

(2)
(Total 10 marks)

12. Cotinine is a chemical that is made by the body from nicotine which is found in cigarette smoke. A doctor tested the blood of 12 patients, who claimed to smoke a packet of cigarettes a day, for cotinine. The results, in appropriate units, are shown below.

Patient	A	B	C	D	E	F	G	H	I	J	K	L
Cotinine level, x	160	390	169	175	125	420	171	250	210	258	186	243

[You may use $\sum x^2 = 724961$]

(a)   Find the mean and standard deviation of the level of cotinine in a patient's blood.

(4)

(b)   Find the median, upper and lower quartiles of these data.

(3)

A doctor suspects that some of his patients have been smoking more than a packet of cigarettes per day. He decides to use $Q_3 + 1.5(Q_3 - Q_1)$ to determine if any of the cotinine results are far enough away from the upper quartile to be outliers.

(c)   Identify which patient(s) may have been smoking more than a packet of cigarettes a day. Show your working clearly.
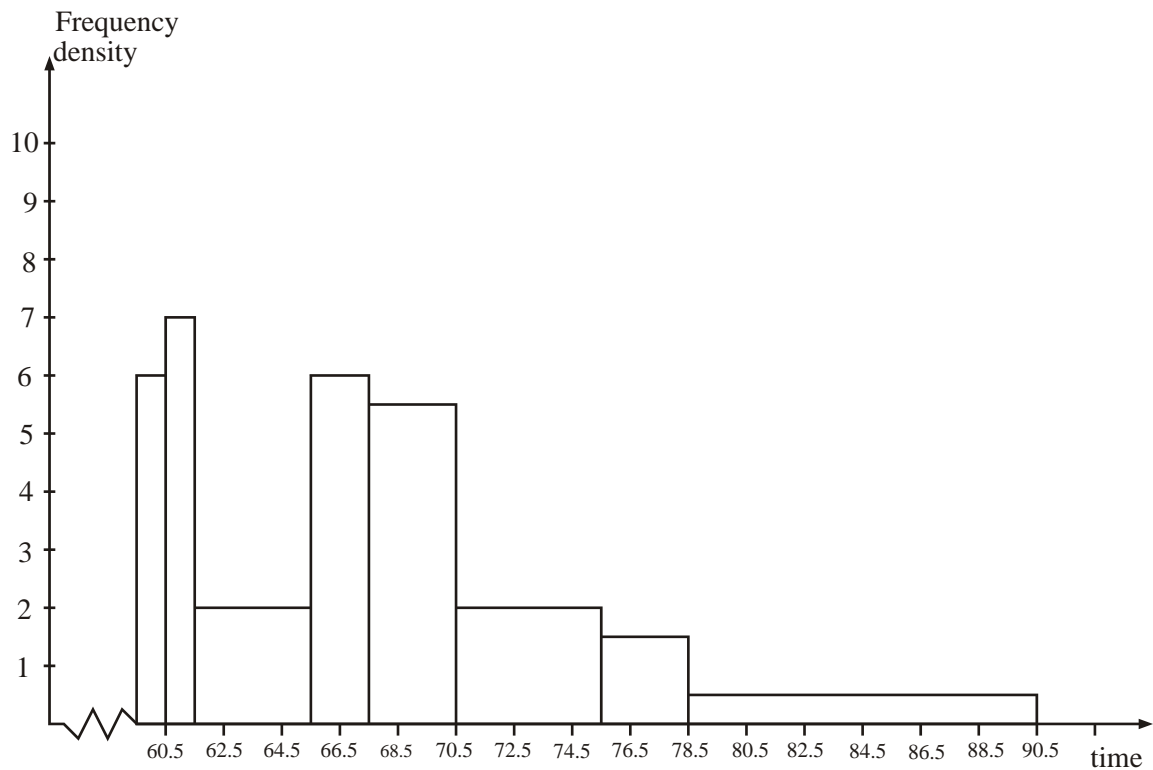
(4)

Research suggests that cotinine levels in the blood form a skewed distribution.

One measure of skewness is found using $\dfrac{(Q_1 - 2Q_2 + Q_3)}{(Q_3 - Q_{1)}}$

(d)   Evaluate this measure and describe the skewness of these data.

(3)
(Total 14 marks)

13.   The histogram below shows the time taken, to the nearest minute, for 140 runners to complete a fun run.

Use the histogram to calculate the number of runners who took between 78.5 and 90.5 minutes to complete the fun run.

(Total 5 marks)

14.   The box plot shown below shows a summary of the weights of the luggage, in kg, for each musician in an orchestra on an overseas tour.
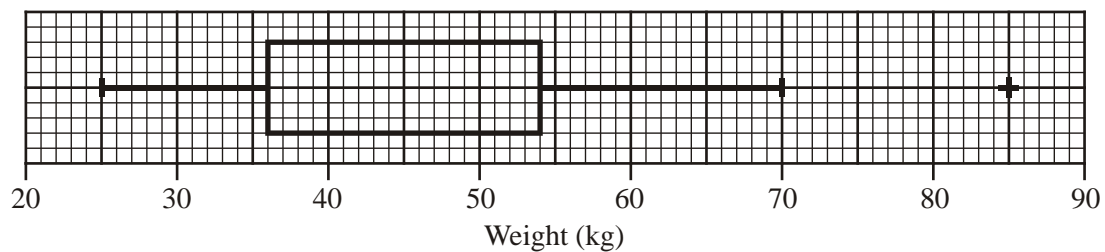
The airline's recommended weight limit for each musician's luggage was 45 kg.
Given that none of the musicians' luggage weighed exactly 45 kg,

(a)   state the proportion of the musicians whose luggage was below the recommended weight limit.

(1)

A quarter of the musicians had to pay a charge for taking heavy luggage.

(b)   State the smallest weight for which the charge was made.

(1)

(c)   Explain what you understand by the + on the box plot in the diagram above, and suggest an instrument that the owner of this luggage might play.

(2)

(d)   Describe the skewness of this distribution. Give a reason for your answer.

(2)

One musician of the orchestra suggests that the weights of luggage, in kg, can be modelled by a normal distribution with quartiles as given in the diagram above.

(e)   Find the standard deviation of this normal distribution.

(4)
(Total 10 marks)

15.   A student is investigating the relationship between the price (y pence) of 100g of chocolate and the percentage ($x\%$) of cocoa solids in the chocolate.
The following data is obtained

Chocolate brand

Chocolate brand	A	B	C	D	E	F	G	H
x (% cocoa)	10	20	30	35	40	50	60	70
y(pence)	35	55	40	100	60	90	110	130

(You may use: $\sum x = 315$, $\sum x^2 = 15\,225$, $\sum y = 620$, $\sum y^2 = 56\,550$, $\sum xy = 28\,750$)

(a)   On the graph paper below draw a scatter diagram to represent these data.

y (pence)
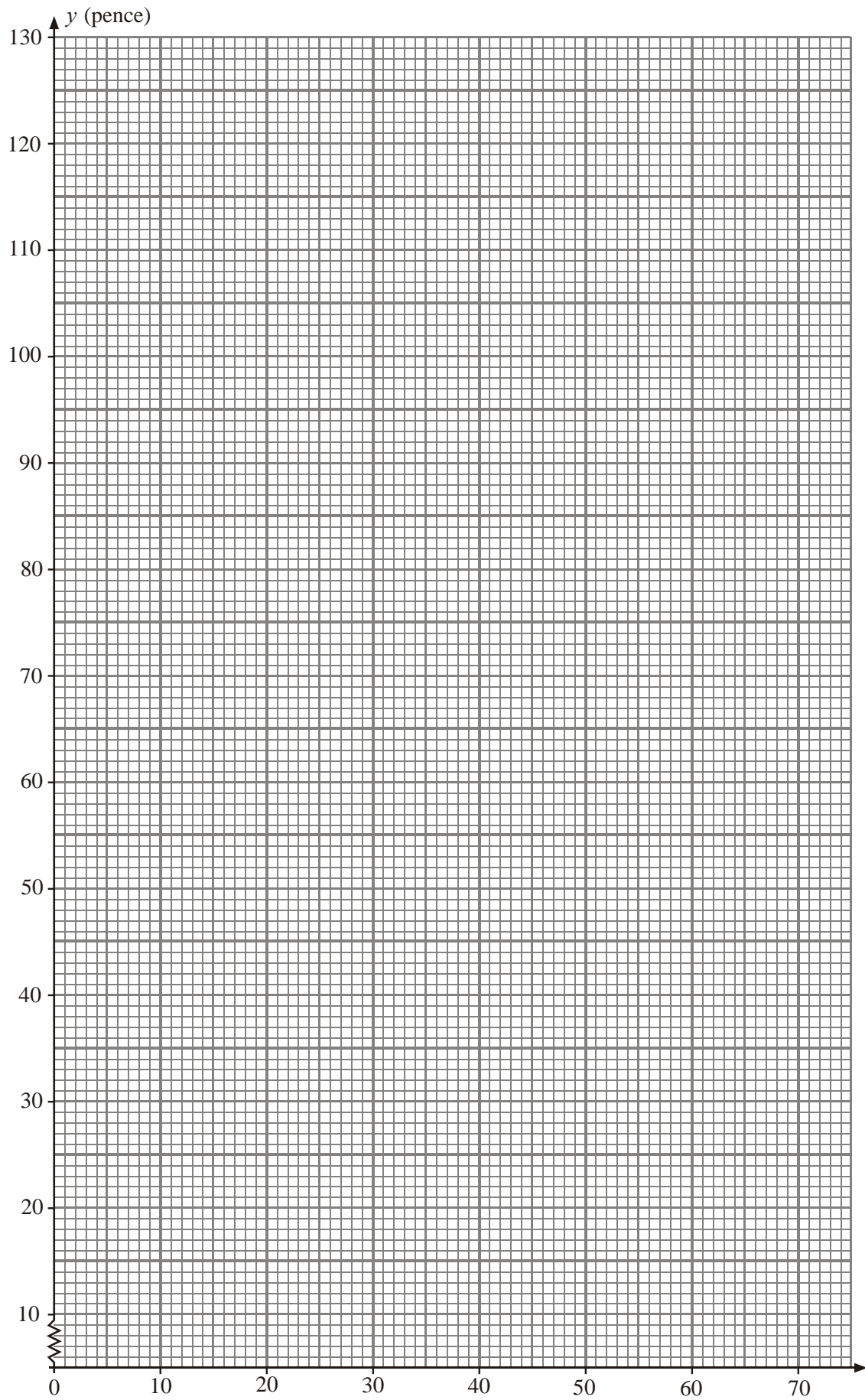
(2)

(b) Show that $S_{xy} = 4337.5$ and find S_{xx}.

(3)

The student believes that a linear relationship of the form $y = a + bx$ could be used to describe these data.

(c) Use linear regression to find the value of a and the value of b, giving your answers to 1 decimal place.

(4)

(d) Draw the regression line on your scatter diagram.

(2)

The student believes that one brand of chocolate is overpriced.

(e) Use the scatter diagram to

(i) state which brand is overpriced,

(ii) suggest a fair price for this brand.

Give reasons for both your answers.

(4)
(Total 15 marks)

16.
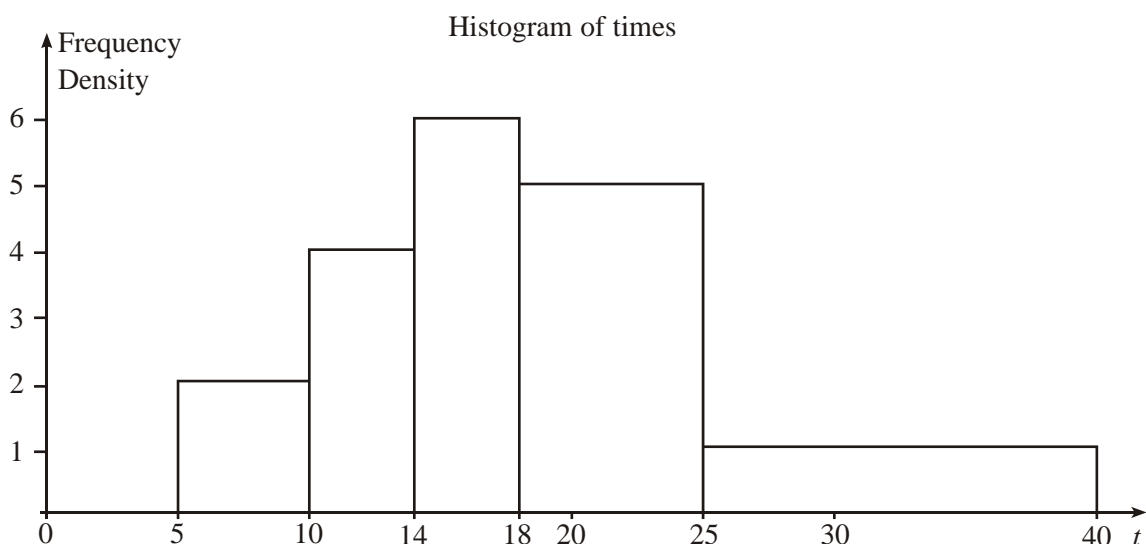
The diagram above shows a histogram for the variable t which represents the time taken, in minutes, by a group of people to swim 500m.

(a)   Complete the frequency table for *t*.

t	5–10	10–14	14–18	18–25	25–40
Frequency	10	16	24		

(2)

(b)   Estimate the number of people who took longer than 20 minutes to swim 500m.

(2)

(c)   Find an estimate of the mean time taken.

(4)

(d)   Find an estimate for the standard deviation of *t*.

(3)

(e)   Find the median and quartiles for *t*.

(4)

One measure of skewness is found using $\dfrac{3(\text{mean} - \text{median})}{\text{standard deviation}}$.

(f)   Evaluate this measure and describe the skewness of these data.

(2)
(Total 17 marks)

17.   Summarised below are the distances, to the nearest mile, travelled to work by a random sample of 120 commuters.

Distance (to the nearest mile)	Number of commuters
0–9	10
10–19	19
20–29	43
30–39	25
40–49	8
50–59	6

60–69	5
70–79	3
80–89	1

For this distribution,

(a)   describe its shape,

(1)

(b)   use linear interpolation to estimate its median.

(2)

The mid-point of each class was represented by x and its corresponding frequency by f giving

$$\Sigma fx = 3550 \text{ and } \Sigma fx^2 = 138020$$

(c)   Estimate the mean and the standard deviation of this distribution.

(3)

One coefficient of skewness is given by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}.$$

(d)   Evaluate this coefficient for this distribution.

(3)

(e)   State whether or not the value of your coefficient is consistent with your description in part (a). Justify your answer.

(2)

(f)   State, with a reason, whether you should use the mean or the median to represent the data in this distribution.

(2)

(g)   State the circumstance under which it would not matter whether you used the mean or the median to represent a set of data.

(1)
(Total 14 marks)

18. A teacher recorded, to the nearest hour, the time spent watching television during a particular week by each child in a random sample. The times were summarised in a grouped frequency table and represented by a histogram.

One of the classes in the grouped frequency distribution was 20–29 and its associated frequency was 9. On the histogram the height of the rectangle representing that class was 3.6 cm and the width was 2 cm.

(a) Give a reason to support the use of a histogram to represent these data.

(1)

(b) Write down the underlying feature associated with each of the bars in a histogram.

(1)

(c) Show that on this histogram each child was represented by 0.8 cm^2.

(3)

The total area under the histogram was 24 cm^2.

(d) Find the total number of children in the group.

(2)
(Total 7 marks)

19. (a) Describe the main features and uses of a box plot.

....

....

....

....

....

(3)

Children from school *A* and *B* took part in a fun run for charity. The times to the nearest minute, taken by the children from school *A* are summarised in the figure below.

School *A*

Time (minutes)
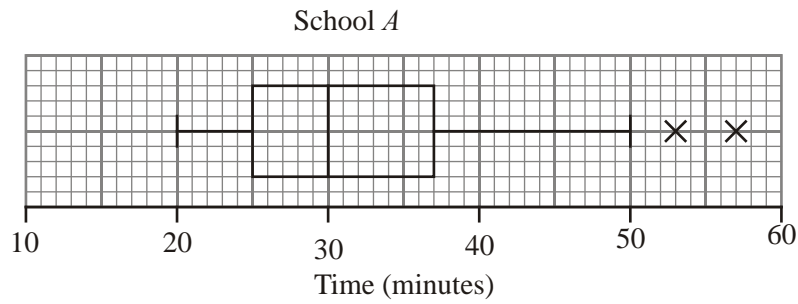
(b) (i) Write down the time by which 75% of the children in school *A* had completed the run.

....

....

(ii) State the name given to this value.

....

(2)

(c) Explain what you understand by the two crosses (✖) on the figure above.

....

....
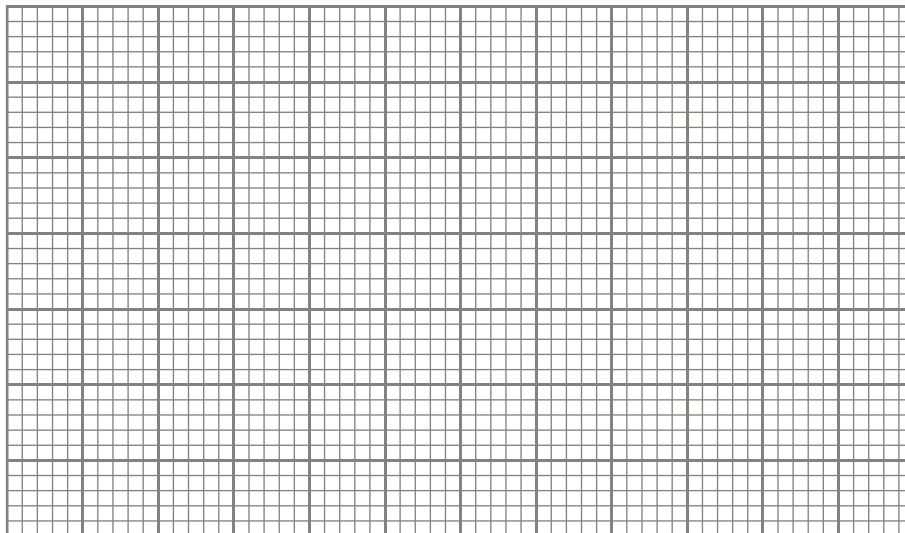
(2)

For school *B* the least time taken by any of the children was 25 minutes and the longest time was 55 minutes. The three quartiles were 30, 37 and 50 respectively.

(d) Draw a box plot to represent the data from school *B*.

(4)

(e)   Compare and contrast these two box plots.

    ....

    ....

    ....

    ....

    ....

(4)
(Total 15 marks)

20.   Sunita and Shelley talk to one another once a week on the telephone. Over many weeks they recorded, to the nearest minute, the number of minutes spent in conversation on each occasion. The following table summarises their results.

Time (to the nearest minute)	Number of Conversations
5–9	2
10–14	9
15–19	20
20–24	13
25–29	8
30–34	3

Two of the conversations were chosen at random.

(a)   Find the probability that both of them were longer than 24.5 minutes.

(2)

The mid-point of each class was represented by x and its corresponding frequency by f, giving $\Sigma fx = 1060$.

(b)   Calculate an estimate of the mean time spent on their conversations.

(2)

During the following 25 weeks they monitored their weekly conversations and found that at the end of the 80 weeks their overall mean length of conversation was 21 minutes.

(c)   Find the mean time spent in conversation during these 25 weeks.

(4)

(d)   Comment on these two mean values.

(2)
(Total 10 marks)

21.   Over a period of time, the number of people x leaving a hotel each morning was recorded. These data are summarised in the stem and leaf diagram below.

Number leaving		3|2 means 32	Totals
2	7 9 9		(3)
3	2 2 3 5 6		(5)
4	0 1 4 8 9		(5)
5	2 3 3 6 6 6 8		(7)
6	0 1 4 5		(4)
7	2 3		(2)
8	1		(1)

For these data,

(a)   write down the mode,

(1)

(b)   find the values of the three quartiles.

(3)

Given that $\Sigma x = 1335$ and $\Sigma x^2 = 71\ 801$ find

(c) the mean and the standard deviation of these data.

(4)

One measure of skewness is found using

$$\frac{\text{mean} - \text{mode}}{\text{standard deviation}}.$$

(d) Evaluate this measure to show that these data are negatively skewed.

(2)

(e) Give two other reasons why these data are negatively skewed.

(4)
(Total 14 marks)

22. The following table summarises the distances, to the nearest km, that 134 examiners travelled to attend a meeting in London.

Distance (km)	Number of examiners
41–45	4
46–50	19
51–60	53
61–70	37
71–90	15
91–150	6

(a) Give a reason to justify the use of a histogram to represent these data.

(1)

(b) Calculate the frequency densities needed to draw a histogram for these data.
     (DO NOT DRAW THE HISTOGRAM)

(2)

(c) Use interpolation to estimate the median Q_2, the lower quartile Q_1, and the upper quartile Q_3 of these data.

(4)

The mid-point of each class is represented by x and the corresponding frequency by f.
Calculations then give the following values

$$\sum fx = 8379.5 \qquad \text{and} \qquad \sum fx^2 = 557489.75$$

(d)   Calculate an estimate of the mean and an estimate of the standard deviation for these data.

(4)

One coefficient of skewness is given by

$$\frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}.$$

(e)   Evaluate this coefficient and comment on the skewness of these data.

(4)

(f)   Give another justification of your comment in part (e).

(1)
(Total 16 marks)

23.   Aeroplanes fly from City *A* to City *B*. Over a long period of time the number of minutes delay in take-off from City *A* was recorded. The minimum delay was 5 minutes and the maximum delay was 63 minutes. A quarter of all delays were at most 12 minutes, half were at most 17 minutes and 75% were at most 28 minutes. Only one of the delays was longer than 45 minutes.

An outlier is an observation that falls either 1.5 × (interquartile range) above the upper quartile or 1.5 × (interquartile range) below the lower quartile.
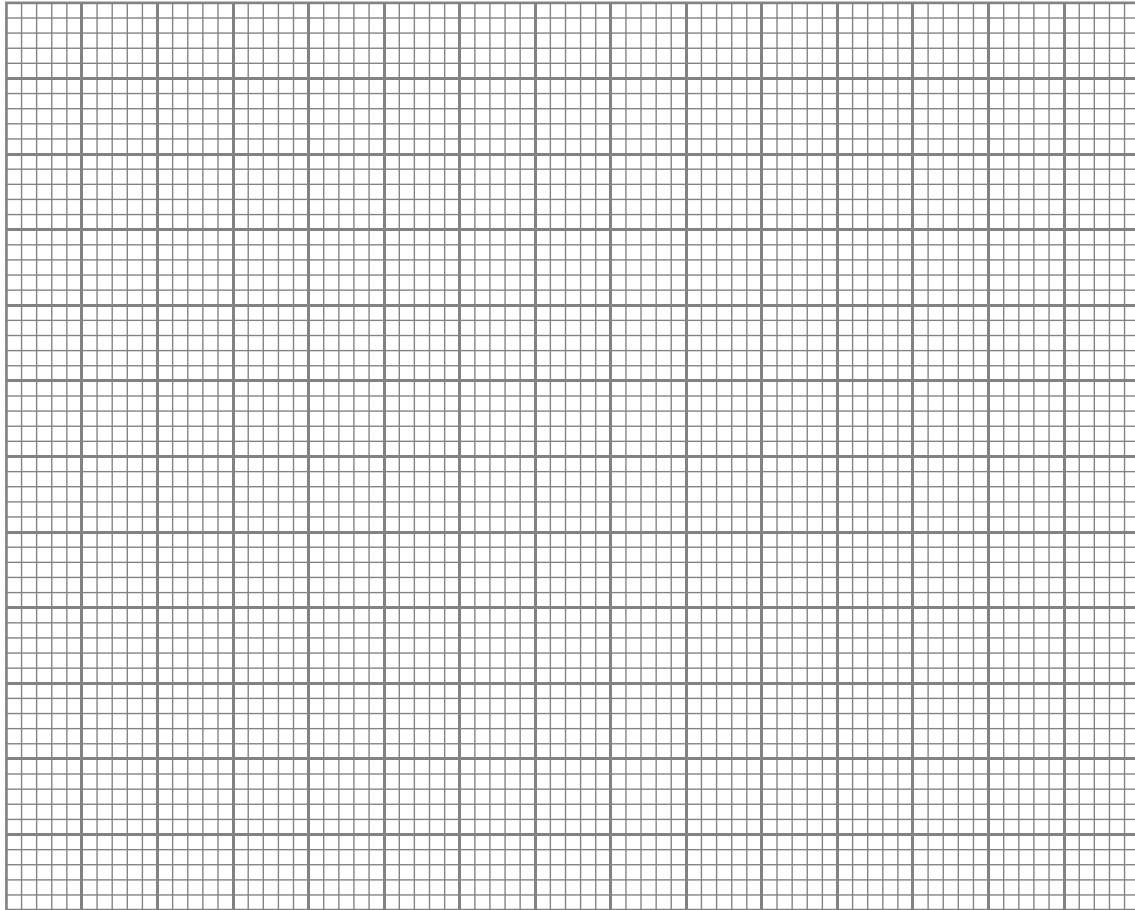
(a)   On graph paper, draw a box plot to represent these data.

(7)

(b)   Comment on the distribution of delays. Justify your answer.

(2)

(c)   Suggest how the distribution might be interpreted by a passenger who frequently flies
        from City *A* to City *B*.

(1)

(Total 10 marks)

24.   The number of caravans on Seaview caravan site on each night in August last year is
        summarised in the following stem and leaf diagram.

```
          Caravans                 1│0 means 10
                          Totals
     1 │            0   5                        (2)
     2 │        1   2   4   8                    (4)
     3 │    0   3   3   3   4   7   8   8        (8)
     4 │  1   1   3   5   8   8   8   9   9      (9)
     5 │        2   3   6   6   7                (5)
     6 │        2   3   4                        (3)
```

(a)   Find the three quartiles of these data.

(3)

During the same month, the least number of caravans on Northcliffe caravan site was 31. The maximum number of caravans on this site on any night that month was 72. The three quartiles for this site were 38, 45 and 52 respectively.

(b)   On graph paper and using the same scale, draw box plots to represent the data for both caravan sites. You may assume that there are no outliers.

(*One sheet of graph paper to be provided*)

(6)

(c)   Compare and contrast these two box plots.

(3)

(d)   Give an interpretation to the upper quartiles of these two distributions.

(2)
(Total 14 marks)

25.   As part of their job, taxi drivers record the number of miles they travel each day. A random sample of the mileages recorded by taxi drivers Keith and Asif are summarised in the back-to-back stem and leaf diagram below.

-

Totals	Keith											Asif								Totals
(9)		8	7	7	4	3	2	1	1	0	18	4	4	5	7					(4)
(11)	9	9	8	7	6	5	4	3	3	1	1	19	5	7	8	9	9			(5)
(6)				8	7	4	2	2	0	20	0	2	2	4	4	8				(6)
(6)				9	4	3	1	0	0	21	2	3	5	6	6	7	9			(7)
(4)					6	4	1	1	22	1	1	2	4	5	5	8				(7)
(2)							2	0	23	1	1	3	4	6	6	7	8			(8)
(2)							7	1	24	2	4	8	9							(4)
(1)								9	25	4										(1)
(2)							9	3	26											(0)

Key: 0 │18│ 4   means 180 for Keith and 184 for Asif

The quartiles for these two distributions are summarised in the table below.

	Keith	Asif
Lower quartile	191	*a*
Median	*b*	218
Upper quartile	221	*c*

(a)   Find the values of *a*, *b* and *c*.

(3)

Outliers are values that lie outside the limits

$$Q_1 - 1.5(Q_3 - Q_1) \quad \text{and} \quad Q_3 + 1.5(Q_3 - Q_1).$$

(b)   On graph paper, and showing your scale clearly, draw a box plot to represent Keith's data.

(8)

(c)   Comment on the skewness of the two distributions.

(3)
(Total 14 marks)

26.   A college organised a 'fun run'. The times, to the nearest minute, of a random sample of 100 students who took part are summarised in the table below.

Time	Number of students
40–44	10
45–47	15
48	23
49–51	21
52–55	16
56–60	15

(a)   Give a reason to support the use of a histogram to represent these data.

(1)

(b)   Write down the upper class boundary and the lower class boundary of the class 40–44.

(1)

(c)   On graph paper, draw a histogram to represent these data.

(4)
(Total 6 marks)

27.   The attendance at college of a group of 18 students was recorded for a 4-week period.

The number of students actually attending each of 16 classes are shown below.

18	18	17	17
16	17	16	18
18	14	17	18
15	17	18	16

(a)   (i)   Calculate the mean and the standard deviation of the number of students attending these classes.

(ii)   Express the mean as a percentage of the 18 students in the group.

(5)

In the same 4-week period, the attendance of a different group of 20, students is shown below.

20	16	18	19
15	14	14	15
18	15	16	17
16	18	15	14

(b)   Construct a back-to-back stem and leaf diagram to represent the attendance in both groups.

(5)

(c)   Find the mode, median and inter-quartile range for each group of students.

(6)

The mean percentage attendance and standard deviation for the second group of students are 81.25 and 1.82 respectively.

(d)   Compare and contrast the attendance of these 2 groups of students.

(3)
(Total 19 marks)

28.   The values of daily sales, to the nearest £, taken at a newsagents last year are summarised in the table below.

Sales	Number of days
1 – 200	166
201 – 400	100
401 – 700	59
701 – 1000	30
1001 – 1500	5

(a)   Draw a histogram to represent these data.

(*One sheet of graph paper to be provided*).

(5)

(b)   Use interpolation to estimate the median and inter-quartile range of daily sales.

(5)

(c)   Estimate the mean and the standard deviation of these data.

(6)

The newsagent wants to compare last year's sales with other years.

(d)   State whether the newsagent should use the median and the inter-quartile range or the mean and the standard deviation to compare daily sales. Give a reason for your answer.

(2)
(Total 18 marks)

29. A travel agent sells holidays from his shop. The price, in £, of 15 holidays sold on a particular day are shown below.

299	1050	2315	999	485
350	169	1015	650	830
99	2100	689	550	475

For these data, find

(a)   the mean and the standard deviation,

(3)

(b)   the median and the inter-quartile range.

(4)

An outlier is an observation that falls either more than $1.5 \times$ (inter-quartile range) above the upper quartile or more than $1.5 \times$ (inter-quartile range) below the lower quartile.

(c)   Determine if any of the prices are outliers.

(3)

The travel agent also sells holidays from a website on the Internet. On the same day, he recorded the price, £x, of each of 20 holidays sold on the website. The cheapest holiday sold was £98, the most expensive was £2400 and the quartiles of these data were £305, £1379 and £1805. There were no outliers.

(d)   On graph paper, and using the same scale, draw box plots for the holidays sold in the shop and the holidays sold on the website.

(4)

(e)   Compare and contrast sales from the shop and sales from the website.

(2)
(Total 16 marks)

30. In a particular week, a dentist treats 100 patients. The length of time, to the nearest minute, for each patient's treatment is summarised in the table below.

Time (minutes)	4 – 7	8	9 – 10	11	12 – 16	17 – 20
Number of patients	12	20	18	22	15	13

Draw a histogram to illustrate these data.

(Total 5 marks)

31. The number of bags of potato crisps sold per day in a bar was recorded over a two-week period. The results are shown below.

20, 15, 10, 30, 33, 40, 5, 11, 13, 20, 25, 42, 31, 17

(a)   Calculate the mean of these data.

(2)

(b)   Draw a stem and leaf diagram to represent these data.

(3)

(c)   Find the median and the quartiles of these data.

(3)

An outlier is an observation that falls either $1.5 \times$ (interquartile range) above the upper quartile or $1.5 \times$ (interquartile range) below the lower quartile.

(d)   Determine whether or not any items of data are outliers.

(3)

(e)   On graph paper draw a box plot to represent these data. Show your scale clearly.

(3)

(f)   Comment on the skewness of the distribution of bags of crisps sold per day. Justify your answer.

(2)
(Total 16 marks)

32. The total amount of time a secretary spent on the telephone in a working day was recorded to the nearest minute. The data collected over 40 days are summarised in the table below.

Time (mins)	90–139	140–149	150–159	160–169	170–179	180–229
No. of days	8	10	10	4	4	4

Draw a histogram to illustrate these data.

(4)

33. A restaurant owner is concerned about the amount of time customers have to wait before being served. He collects data on the waiting times, to the nearest minute, of 20 customers. These data are listed below.

15,   14,   16,   15,   17,   16,   15,   14,   15,   16,
17,   16,   15,   14,   16,   17,   15,   25,   18,   16

(a) Find the median and inter-quartile range of the waiting times.

(5)

An outlier is an observation that falls either 1.5 × (inter-quartile range) above the upper quartile or 1.5 × (inter-quartile range) below the lower quartile.

(b) Draw a boxplot to represent these data, clearly indicating any outliers.

(7)

(c) Find the mean of these data.

(2)

(d) Comment on the skewness of these data.   Justify your answer.

(2)
(Total 16 marks)

34. A botany student counted the number of daisies in each of 42 randomly chosen areas of 1 m by 1 m in a large field. The results are summarised in the following stem and leaf diagram.

Number of daisies								1	1	means 11
1	1	2	2	3	4	4	4			(7)
1	5	5	6	7	8	9	9			(7)
2	0	0	1	3	3	3	3	4		(8)
2	5	5	6	7	9	9	9			(7)
3	0	0	1	2	4	4				(6)
3	6	6	7	8	8					(5)
4	1	3								(2)

(a) Write down the modal value of these data.

(1)

(b) Find the median and the quartiles of these data.

(4)

(c) On graph paper and showing your scale clearly, draw a box plot to represent these data.

(4)

(d) Comment on the skewness of this distribution.

(1)

The student moved to another field and collected similar data from that field.

(e) Comment on how the student might summarise both sets of raw data before drawing box plots.

(1)
(Total 11 marks)

35. Data relating to the lifetimes (to the nearest hour) of a random sample of 200 light bulbs from the production line of a manufacturer were summarised in a group frequency table. The mid-point of each group in the table was represented by x and the corresponding frequency for that group by f. The data were then coded using $y = \dfrac{(x - 755.0)}{2.5}$ and summarised as follows:

$$\Sigma fy = -467, \quad \Sigma fy^2 = 9179.$$

(a)   Calculate estimates of the mean and the standard deviation of the lifetimes of this sample
       of bulbs.

(9)

An estimate of the interquartile range for these data was 27.7 hours.

(b)   Explain, giving a reason, whether you would recommend the manufacturer to use the
       interquartile range or the standard deviation to represent the spread of lifetimes of the
       bulbs from this production line.

(2)
(Total 11 marks)

1.   (a)   23,   35.5 (may be in the table)   B1 B1   2

(b)   Width of 10 units is 4 cm so width of 5 units is **2 cm**   B1

Height = $2.6 \times 4 = $ **10.4 cm**   M1 A1   3

Note

M1 for their width $\times$ their height = 20.8.
Without labels assume width first, height second and award
marks accordingly.

(c) $\sum fx = 1316.5 \Rightarrow \bar{x} = \dfrac{1316.5}{56} = $ awrt **23.5**   M1 A1

$\sum fx^2 = 37378.25$ can be implied   B1

So $\sigma = \sqrt{\dfrac{37378.25}{56} - \bar{x}^2} = $ awrt **10.7**   allow $s = 10.8$ M1 A1   5

Note

1st M1 for reasonable attempt at $\sum x$ and /56

2nd M1 for a method for σ or s, $\sqrt{}$ is required

Typical errors $\sum (fx)^2 = 354806.3$ M0, $\sum f^2 x = 13922.5$ M0

and $\left(\sum fx\right)^2 = 1733172$ M0
Correct answers only, award full marks.

(d) $Q_2 = (20.5) + \dfrac{(28-21)}{11} \times 5 = 23.68\ldots$ **awrt 23.7 or 23.9** M1 A1   2

Note

Use of $\sum f\left(x - \bar{x}\right)^2 = $ awrt 6428.75 for B1

lcb can be 20, 20.5 or 21, width can be 4 or 5 and the fraction part
of the formula correct for M1 – Allow 28.5 in fraction that gives
awrt 23.9 for M1A1

(e) $Q_3 - Q_2 = 5.6$, $Q_2 - Q_1 = 7.9$ (or $\bar{x} < Q_2$)   M1

[7.9 > 5.6 so]   **negative skew**   A1   2

Note

M1for attempting a test for skewness using quartiles or mean and median.
Provided median greater than 22.55 and less than 29.3 award for M1 for
$Q_3 - Q_2 < Q_2 - Q_1$ without values as a valid reason.
SC Accept mean close to median and no skew oe for M1A1

[14]

2.   (a)   Median is 33   B1   1

   (b) $Q_1 = 24, Q_3 = 40$, IQR $= 16$ B1 B1 B1ft   3

   Note

   1st B1   for Q1 = 24 and 2nd B1 for $Q_3 = 40$

   3rd B1ft for their IQR based on their lower and upper
            quartile.

            Calculation of range $(40 - 7 = 33)$ is B0B0B0

   <u>Answer only</u> of IQR = 16 scores 3/3. For any other answer
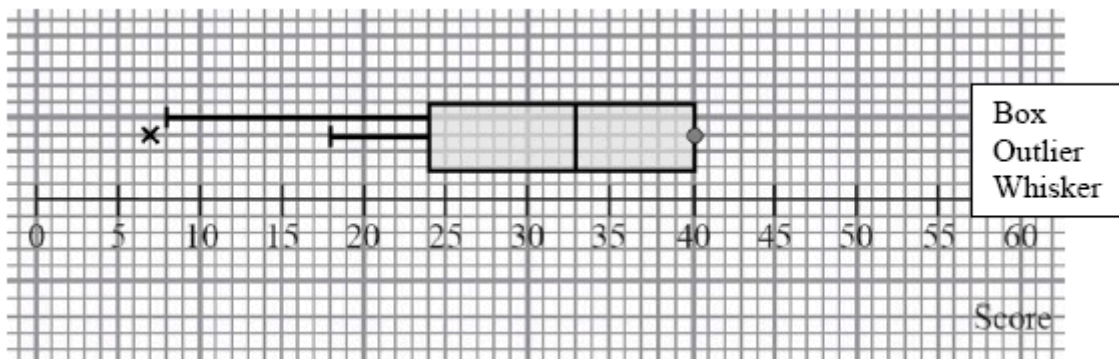   we must see working in (b) or on stem and leaf diagram

   (c) $Q_1 -$ IQR $= 24 - 16 = 8$ M1

         So 7 is only outlier   A1ft   2

   Note

   M1   for evidence that Q_1-IQR has been attempted, their
           "8" (>7) seen or clearly attempted is sufficient

   A1 ft   must have seen their "8" and a suitable comment
           that only one person scored below this.

   (d)

                                                          B1ftB1B1ft   3

   Note

   1st B1ft for a clear box shape and ft their Q_1, Q_2 and Q_3
            readable off the scale.

            Allow this mark for a box shape even if $Q_3 = 40$,
            $Q_1 = 7$ and $Q_2 = 33$ are used

   2nd B1   for only one outlier appropriately marked at 7

   3rd B1ft for either lower whisker. If they choose the whisker
            to their lower limit for outliers then follow through
            their "8".

            (There should be no upper whisker unless their $Q_3 < 40$,
            in which case there should be a whisker to 40)

A typical error in (d) is to draw the lower whisker to 7,
this can only score B1B0B0

[9]

3.   (a)   2.75 or $2\frac{3}{4}$, 5.5 or 5.50 or $5\frac{1}{2}$ B1 B1   2

(b)   Mean birth weight $= \dfrac{4841}{1500} = 3.227\dot{3}$ **awrt 3.23**   M1 A1   2

Note

M1   for a correct expression for mean. Answer only
            scores both.

(c)   Standard deviation $= \sqrt{\dfrac{15889.5}{1500} - \left(\dfrac{4841}{1500}\right)^2} =$

0.421093... or $s = 0.4212337...$ M1 A1ft A1   3

Note

M1   for a correct expression (ft their mean) for sd or
            variance. Condone mis-labelling eg sd=…

            with no square root or no labelling

1$^{\text{st}}$ A1ft   for a correct expression (ft their mean) including
            square root and no mis-labelling

            Allow 1$^{\text{st}}$ A1 for $\sigma^2 = 0.177... \rightarrow \sigma = 0.42...$

2$^{\text{nd}}$ A1   for awrt 0.421. Answer only scores 3/3

(d) $Q_2 = 3.00 + \dfrac{400}{820} \times 0.5 = 3.2457....$ (allow 403.5…..

                                              $\rightarrow$ 3.25)   M1 A1   2

Note

M1   for a correct expression (allow 403.5 i.e. use of
            $n + 1$) but must have 3.00, 820 and 0.5

A1   for awrt 3.25 provided M1 is scored.

            NB 3.25 with no working scores 0/2 as some
            candidates think mode is 3.25.

(e)   Mean(3.23)<Median(3.25)   (or very close)   B1ft

      Negative Skew   (or symmetrical)   dB1ft   2

Note

1$^{\text{st}}$ B1ft   for a comparison of their mean and median
            (may be in a formula but if $\pm$ (mean – median)
            is calculated that's OK. We are not checking

the <u>value</u> but the <u>sign</u> must be consistent.)

Also allow for use of quartiles <u>provided correct values seen</u>: $Q_1 = 3.02, Q_3 = 3.47$

[They should get $(0.22 =)Q_3 - Q_2 < Q_2 - Q_1 (= 0.23)$ and say (slight) negative skew or symmetric]

2nd dB1ft for a compatible comment based on their comparison. Dependent upon a suitable, correct comparison. Mention of "correlation" rather than "skewness" loses this mark.

[11]

4.   (a)
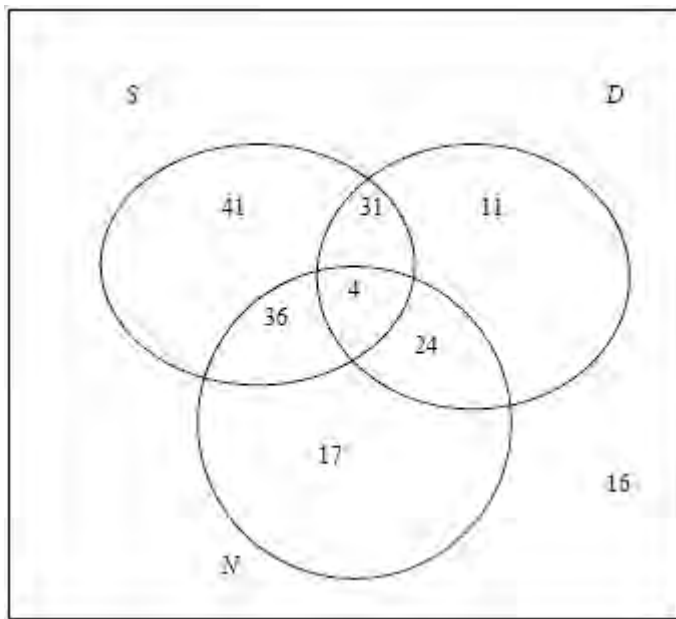
3 closed curves and 4 in centre   M1

Evidence of subtraction   M1

31,36,24   A1

41,17,11   A1

Labels on loops, 16 and box   B1   5

<u>Note</u>

2nd M1   There may be evidence of subtraction in "outer" portions, so with 4 in the centre then 35, 40 28 (instead of 31,36,24) along with 33, 9, 3 can score this mark but A0A0

N.B. This is a common error and their "16" becomes 28 but still scores B0 in part (a)

(b)   P(None of the 3 options) $= \dfrac{16}{180} = \dfrac{4}{45}$ B1ft   1

Note

B1ft   for $\frac{16}{180}$ or any exact equivalent. Can ft their "16" from their box. If there is no value for their "16" in the box only allow this mark if they have <u>shown</u> some working.

(c)   P(Networking only) = $\dfrac{17}{180}$ B1ft   1

Note

B1ft   ft their "17". Accept any exact equivalent

(d)   P(All 3 options/technician) = $\dfrac{4}{40} = \dfrac{1}{10}$ M1 A1   2

Note

If a probability greater than 1 is found in part
(d) score M0A0

M1   for clear sight of $\dfrac{P(S \cap D \cap N)}{P(S \cap N)}$ and an attempt
at one of the probabilities, ft their values.

Allow P(all 3 | $S \cap N$) = $\dfrac{4}{36}$ or $\dfrac{1}{9}$ to score M1 A0.

Allow a correct ft from their diagram to score M1A0
e.g. in 33,3,9 case in (a): $\frac{4}{44}$ or $\frac{1}{11}$ is M1A0 A ratio of
probabilities with a <u>product</u> of probabilities on top is
M0, even with a correct formula.

A1   for $\dfrac{4}{40}$ or $\dfrac{1}{10}$ or an exact equivalent

Allow $\dfrac{4}{40}$ or $\dfrac{1}{10}$ to score both marks if this follows
from their diagram, otherwise some explanation
(method) is required.

[9]

5.   (a)   1(cm)   B1

cao

(b)   10 cm^2 represents 15

10/15 cm^2 represents 1   or 1cm^2 represents 1.5

Therefore frequency of 9 is

$\dfrac{10}{15} \times 9$ or $\dfrac{9}{1.5}$ Require $\times \dfrac{2}{3}$ or $\div 1.5$ M1

height = 6(cm)   A1

Note

If 3(a) and 3(b) incorrect, but their
(a) $\times$ their (b)=6 then award B0M1A0

Alternative method:

f/cw=15/6=2.5 represented by 5 so factor x2 award M1

So f/cw=9/3=3 represented by 3x2=6. Award A1.

[3]

6.   (a) $Q_2 = 17 + \left(\dfrac{60 - 58}{29} \right) \times 2$ M1

= 17.1 (17.2 if use 60.5)   awrt 17.1 (or17.2)   A1   2

Note

Statement of 17+ $\dfrac{\text{freq into class}}{\text{class freq}} \times cw$

and attempt to sub or $\dfrac{m - 17}{19 - 17} = \dfrac{60(.5) - 58}{87 - 58}$
or equivalent award M1

cw = 2 or 3 required for M1.

17.2 from cw = 3 award A0.

(b) $\sum fx = 2055.5$ $\sum fx^2 = 36500.25$ Exact answers can
                                                              be seen below or
                                                              implied by correct
                                                              answers.   B1 B1

                                Evidence of attempt to use midpoints with
                                at least one correct   M1

Mean = 17.129…   awrt 17.1   B1

$\sigma = \sqrt{\dfrac{36500.25}{120} - \left(\dfrac{2055.5}{120} \right)^2}$ M1

= 3.28 (*s*= 3.294)   awrt 3.3   A1   6

Note

Correct $\sum fx$ and $\sum fx^2$ can be seen in working

for both B1s

Midpoints seen in table and used in calculation award M1

Require complete correct formula including use of square root and attempt to sub for M1. No formula stated then numbers as above or follow from (b) for M1

$\left(\sum fx\right)^{2}, \sum \left(fx\right)^{2}$ *or* $\sum f^{2}x$ used instead of $\sum fx^{2}$
in sd award M0

Correct answers only with no working award 2/2 and 6/6

(c) $\dfrac{3(17.129 - 17.1379...)}{3.28} = -0.00802$ Accept 0 or

                                                                        awrt 0.0   M1 A1

         No skew/ slight skew   B1   3

Note

Sub in their values into given formula for M1

(d)   The skewness is very small. Possible.   B1 B1dep   2

Note

No skew / slight skew / 'Distribution is almost symmetrical' / 'Mean approximately equal to median' or equivalent award first B1. Don't award second B1 if this is not the case. Second statement should imply 'Greg's suggestion that a normal distribution is suitable is possible' for second B1 dep.

If B0 awarded for comment in (c).and (d) incorrect, allow follow through from the **comment** in (c).

                                                                                        [13]
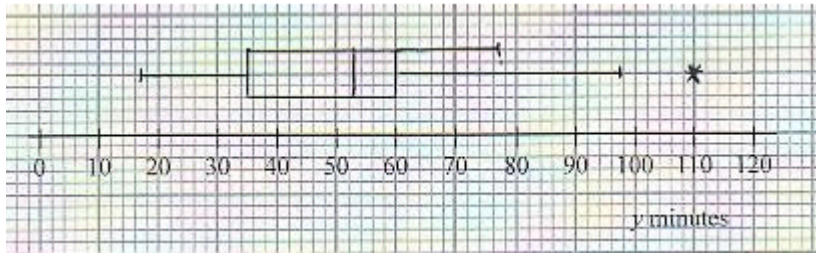
7. (a) $Q_2 = 53, \quad Q_1 = 35, \quad Q_3 = 60$ B1, B1, B1   3

Note

1st B1   for median
2nd B1   for lower quartile
3rd B1   for upper quartile

(b) $Q_3 - Q_1 = 25 \Rightarrow Q_1 - 1.5 \times 25 = -2.5$ (no outlier)   M1

            $Q_3 + 1.5 \times 25 = 97.5$ (so 110 is an outlier)   A1   2

Note

M1   for attempt to find one limit
A1   for both limits found and correct. No explicit comment about outliers needed.

(c)

M1
A1ft
A1ft   3

Note

M1   for a box and two whiskers

1st A1ft   for correct position of box, median and quartiles. Follow through
          their values.

2nd A1ft   for 17 and 77 or "their" 97.5 and * . If 110 is not an outlier then
          score A0 here. Penalise no gap between end of whisker and outlier.
          Must label outlier, needn't be with * .

Accuracy should be within the correct square so 97 or 98 will do for 97.5

(d) $\sum y = 461, \sum y^2 = 24\,219 \therefore S_{yy} = 24219 - \dfrac{461^2}{10}, = 2966.9(*)$ B1, B1,

                                                            B1cso   3

Note

1st B1   for $\sum y$ N.B. $\left(\sum y\right)^2 = 212521$ and can imply this mark

2nd B1   for $\sum y^2$ or at least three correct terms of $\sum (y - \bar{y})^2$ seen.

3rd B1   for complete correct expression seen leading to 2966.9. So all 10
          terms of $\sum (y - \bar{y})^2$

(e) $r = \dfrac{-18.3}{\sqrt{3463.6 \times 2966.9}}$ or $\dfrac{-18.3}{3205.64...} = -0.0057$

       AWRT -0.006 or -6×10^{-3} M1 A1   2

Note

M1   for attempt at correct expression for *r*. Can ft their S$_{yy}$ for M1.

(f) *r* suggests correlation is close to zero so parent's claim is not justified   B1   1

Note

B1   for comment <u>rejecting</u> parent's claim on basis of <u>weak or zero</u> correlation
Typical error is "negative correlation so comment is true" which scores B0
Weak negative or weak positive correlation is OK as the basis for their rejection.

[14]

8.   (a)   8 – 10 hours: width = 10.5 – 7.5 = 3 represented by 1.5cm
16 – 25 hours: width = 25.5 – 15.5 = 10 so represented by <u>5 cm</u>   B1
8 – 10 hours: height = fd = 18/3 = 6 represented by 3 cm   M1
16 – 25 hours: height = fd = 15/10 = 1.5 represented by <u>0.75 cm</u>   A1   3

Note

M1   For attempting both frequency densities
$\frac{18}{3}(=6)$ and $\frac{15}{10}$, <u>and</u> $\frac{15}{10} \times SF$, where $SF \neq 1$ NB Wrong class
widths (2 and 9) gives $\frac{h}{1.66...} = \frac{3}{9} \rightarrow h = \frac{5}{9}$ or 0.55...
and scores M1A0

(b) $Q_2 = 7.5 + \dfrac{(52-36)}{18} \times 3 = 10.2$

$Q_1 = 5.5 + \dfrac{(26-20)}{16} \times 2[= 6.25 \text{ or } 6.3]$ or $5.5 + \dfrac{(26.25-20)}{16} \times 2[= 6.3]$ A1

$Q_3 = 10.5 + \dfrac{(78-54)}{25} \times 5[=15.3]$ or $10.5 + \dfrac{(78.75-54)}{25} \times 5[=15.45 \setminus 15.5]$ A1

IQR = (15.3 – 6.3) = <u>9</u>   A1ft   5

Note

M1   for identifying correct interval and a correct fraction e.g.
$\dfrac{\frac{1}{2}(104)-36}{18}$. Condone 52.5 or 53

1st A1   for 10.2 for median. Using (*n* + 1) allow awrt 10.3
<u>NB:</u>

2nd A1   for a correct expression for either Q_1 or Q_3
(allow 26.25 and 78.75)   <u>Must see</u>

3rd A1   for correct expressions for both Q_1 and Q_3 <u>some</u>

4th A1ft for IQR, ft their quartiles. Using (n + 1) gives 6.28
and 15.45   <u>method</u>

(c) $\sum fx = 1333.5 \Rightarrow \overline{x} = \dfrac{1333.5}{104} =$ AWRT <u>12.8</u>   M1 A1

$\sum fx^2 = 27254 \Rightarrow \sigma_x = \sqrt{\dfrac{27254}{104} - \overline{x}^2} = \sqrt{262.05 - \overline{x}^2}$ AWRT <u>9.88</u>   M1 A1   4

<u>Note</u>

1st M1   for attempting $\sum fx$ and $\overline{x}$

2nd M1   for attempting $\sum fx^2$ and $\sigma_x , \sqrt{}$ is needed for M1.
             Allow s = awrt 9.93

(d) $Q_3 - Q_2$ [=5.1] $> Q_2 - Q_1$ [=3.9] or $Q_2 < \overline{x}$ B1ft
                                                                                  dB1   2

<u>Note</u>

1st B1ft   for suitable test, values need not be seen but statement must
             be compatible with values used. Follow through their values
2nd dB1   Dependent upon their test showing positive and for stating
             positive skew If their test shows negative skew they can score
             1st B1 but lose the second

(e)   So data is positively skew

      Use median and IQR,   B1
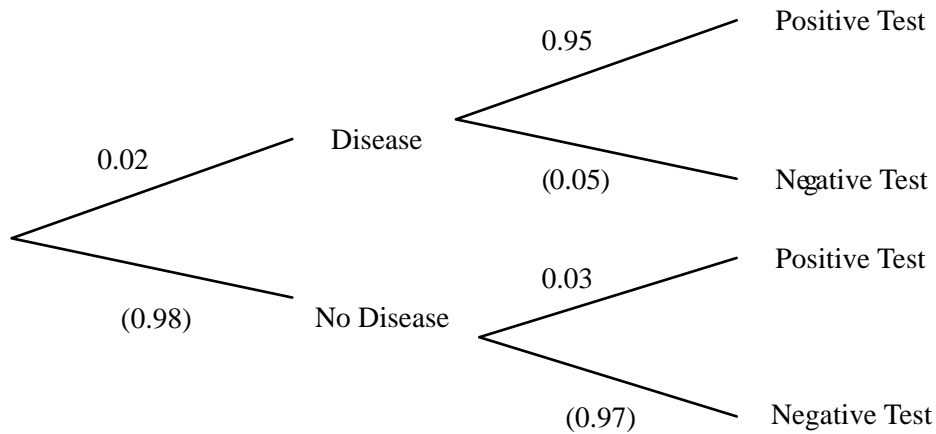   since data is skewed <u>or</u> not affected by extreme values or outliers   B1   2

<u>Note</u>

1st B1   for choosing median and IQR. Must mention <u>both</u>.   }   <u>Award independently</u>

2nd B1   for suitable reason   }
             e.g. "use median because data is skewed" scores B0B1 since IQR
             is not mentioned

                                                                                  [16]

9.   (a)

Disease — 0.95 — Positive Test

0.02 (Disease)

(0.05) — Negative Test

No Disease — 0.03 — Positive Test

(0.98)

(0.97) — Negative Test

Tree without probabilities or labels	M1	
0.02 (Disease), 0.95 (Positive) on correct branche	A1	
0.03 (Positive) on correct branch	A1	3

M1: All 6 branches.

Bracketed probabilities not required.

(b)   P(Positive Test) = $0.02 \times 0.95 + 0.98 \times 0.03$ M1A1ft

      = 0.0484   A1   3

M1 for sum of two products, at least one correct from their diagram

A1ft follows from the probabilities on their tree

A1 for correct answer only or $\dfrac{121}{2500}$

(c)   P(Do not have disease | Postive test) = $\dfrac{0.98 \times 0.03}{0.0484}$ M1

            = 0.607438...   awrt 0.607   A1   2

M1 for condirtional probability with numerator following from their tree and denominator their answer to part (b).

A1 also for $\dfrac{147}{242}$.

(d)   Test not very useful OR
       High probability of not having the disease for a person with a
       positive test   B1   1

                                                                        [9]

10.   (a)   50   B1   1

(b) $Q_1 = 45$ B1
        $Q_2 = 50.5$ ONLY   B1
        $Q_3 = 63$ B1   3

(c)   Mean = $\dfrac{1469}{28}$ = 52.464286..   awrt 52.5   M1A1

        Sd = $\sqrt{\dfrac{81213}{28} - \left(\dfrac{1469}{28}\right)^2}$ M1

        = 12.164…. or 12.387216… for divisor $n - 1$ awrt 12.2 or 12.4   A1   4

        M1 for their 1469 between 1300 and 1600, divided by 28,

        A1 for awrt 52.5 ..

        Please note this is B1B1 on Epen

        M1 use of correct formula including sq root

        A1 awrt 12.2 or 12.4

        Correct answers with no working award full marks.

(d) $\dfrac{52.46.. - 50}{sd}$ = awrt 0.20 or 0.21   M1A1   2

        M1 for their values correctly substituted

        A1 Accept 0.2 as a special case of awrt 0.20 with 0 missing

(e)   1.   mode/median/mean Balmoral>mode/median/mean Abbey

        2.   Balmoral sd < Abbey sd or similar sd or correct comment
               from their values,
               Balmoral range<Abbey range,
               Balmoral IQR>Abbey IQR or similar IQR

        3.   Balmoral positive skew or almost symmetrical AND
               Abbey negative skew, Balmoral is less skew than Abbey
               or correct comment from their value in (d)

        4.   Balmoral residents generally older than Abbey residents
               or equivalent.

               Only one comment of each type max 3 marks   B1B1B1   3

Technical terms required in correct context in lines 1 to 3

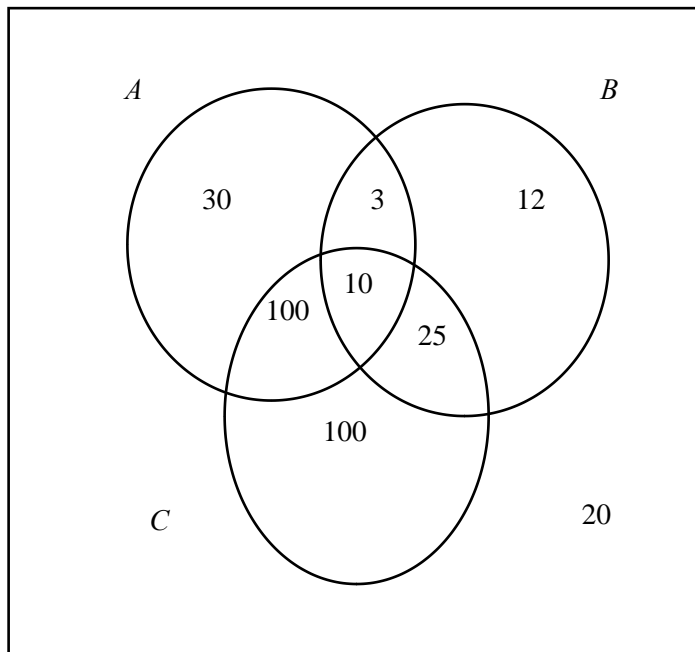e.g. 'average' and 'spread' B0

1 correct comment B1B0B0

2 correct comments B1B1B0

3 correct comments B1B1B1

[13]

11. (a)

3 closed intersecting curves with labels	M1	
100 100, 30	A1	
12, 10, 3, 25	A1	
Box	B1	4

20 not required.
Fractions and exact equivalent decimals or percentages.

(b) P(Substance C) = $\dfrac{100+100+10+25}{300} = \dfrac{235}{300} = \dfrac{47}{60}$ or exact equivalent M1A1ft 2

M1 For adding their positive values in C and finding a probability
A1ft for correct answer or answer from their working

(c) P(All 3 $|A$) = $\dfrac{10}{30+3+10+100} = \dfrac{10}{143}$ or exact equivalent M1A1ft 2

M1 their 10 divided by their sum of values in A
A1ft for correct answer or answer from their working

(d)   P(Universal donor) = $\dfrac{20}{300} = \dfrac{1}{15}$ or exact equivalent   M1A1cao   2

M1 for 'their 20' divided by 300
A1 correct answer only

[10]

12.   (a)   mean is $\dfrac{2757}{12}$, $= 229.75$ AWRT 230   M1, A1

sd is $\sqrt{\dfrac{724961}{12} - (229.75)^2}$, $= 87.34045$ AWRT 87.3   M1, A1   4

[Accept s = AWRT 91.2]

1st M1 for using $\dfrac{\sum x}{n}$ with a credible numerator and $n = 12$.

2nd M1 for using a correct formula, root required but can ft their mean

Use of $s = \sqrt{8321.84...} = 91.22...$ is OK for M1 A1 here.
Answers only from a calculator in (a) can score full marks

(b)   Ordered list is:
125, 160, 169, 171, 175, 186, 210, 243, 250, 258, 390, 420

$Q_2 = \dfrac{1}{2}(186 + 210) = 198$ B1

$Q_1 = \dfrac{1}{2}(169 + 171) = 170$ B1

$Q_3 = \dfrac{1}{2}(250 + 258) = 254$ B1   3

1st B1 for median = 198 only, 2nd B1 for lower quartile 3rd B1 for upper quartile

S.C.   If all Q_1 and Q_3 are incorrect but an ordered list (with ≥ 6 correctly placed) is seen and used then award B0B1 as a special case for these last two marks.

(c) $Q_3 + 1.5(Q_3 - Q_1) = 254 + 1.5(254 - 170), = 380$

                                            Accept AWRT (370-392)   M1, A1

Patients F (420) and B (390) are outliers.   B1ftB1ft   4

M1   for a clear attempt using their quartiles in given formula,

A1   for any value in the range 370 – 392

1st B1ft for any one correct decision about B or F – ft their
         limit in range (258, 420)

2nd B1ft for correct decision about both F and B – ft their
         limit in range (258, 420)
         If more points are given score B0 here for the second B mark.
         (Can score M0A0B1B1 here)

(d) $\dfrac{Q_1 - 2Q_2 + Q_3}{Q_3 - Q_1} = \dfrac{170 - 2 \times 198 + 254}{254 - 170}, = 0.\dot{3}$ AWRT 0.33   M1, A1

Positive skew.   A1ft   3

M1   for an attempt to use their figures in the correct formula
         – must be seen
         (≥ 2 correct substitutions)

1st A1   for AWRT 0.33

2nd A1ft for positive skew. Follow through their value/sign of skewness.
         Ignore any further calculations.
         "positive correlation" scores A0

                                                                    [14]

13.   M1

Width	1	1	4	2	3	5	3	12
Freq. Density	6	7	2	6	5.5	2	1.5	0.5

0.5×12 or 6   A1

Total area is $(1 \times 6) + (1 \times 7) + (4 \times 2) + ..., = 70$

$(90.5 - 78.5) \times \dfrac{1}{2} \times \dfrac{140}{\text{their } 70}$ M1

                                    "70 seen anywhere"   B1

Number of runners is 12   A1   5

1st M1   for attempt at width of the correct bar (90.5 – 78.5)
[Maybe on histogram or in table]

1st A1   for 0.5 × 12 or 6 (may be seen on the histogram).
Must be related to the area of the bar above 78.5 – 90.5.

2nd M1   for attempting area of correct bar × $\dfrac{140}{\text{their } 70}$

B1   for 70 seen anywhere in their working

2nd A1   for correct answer of 12.

Minimum working required is 2 × 0.5 × 12 where the 2 should come from $\dfrac{140}{70}$

Beware 90.5 – 78.5 = 12 (this scores M1A0M0B0A0)

Common answer is 0.5×12= 6 (this scores M1A1M0B0A0)

If unsure send to review e.g. 2 × 0.5 × 12 = 12 without 70 being seen

[5]

14.   (a) $\dfrac{1}{2}$ B1   1

            Accept 50% or half or 0.5>
            Units not required.

   (b)   54   B1   1

            Correct answers only.
            Units not required.

   (c)   + is an 'oulier' or 'extreme value'   B1
            Any heavy musical instrument or a statement that the instrument
            is heavy   B1   2

            'Anomaly' only award B0
            Accept '85 kg was heaviest instrument on the trip' or equivalent
            for second B1.
            Examples of common acceptable instruments; double bass, cello,
            harp, piano, drums, tuba
            Examples of common unacceptable instruments: violin, viola,
            trombone, trumpet, French horn, guitar

(d) $Q_3 - Q_2 = Q_2 - Q_1$ B1

so symmetrical or no skew   Dependent – only award if B1 above   B1   2

'Quartiles equidistant from median' or equivalent award B1 then
symmetrical or no skew for B1
Alternative:
'Positive tail is longer than negative tail' or 'median closer to
lowest value' or equivalent so slight positive skew.
B0 for 'evenly' etc. instead or 'symmetrical'
B0 for 'normal' only

(e)   P($W < 54$) = 0.75 (or p($W > 54$) = 0.25) or correctly labelled and   M1
shaded diagram

$\dfrac{54 - 45}{\sigma} = 0.67$ M1B1

$\sigma = 13.43.....$ A1   4

Please note that B mark appears first on ePEN
First line might be missing so first M1 can be implied by second.
Second M1 for standardising with sigma and equating to z value
NB Using 0.7734 should not be awrded second M1
Anything which rounds to 0.67 for B1.
Accept 0.675 if to 3sf obtained by interpolation
Anything that rounds to 13.3. – 13.4 for A1.

[10]

15.   (a)   Use overlay   B2   2

Points B2, within 1 small square of correct point, subtract 1 mark
each error minimum 0.

(b) $S_{xy} = 28750 - \dfrac{315 \times 620}{8} = 4337.5$ **answer given** so award

for method   M1

$S_{xx} = 15225 - \dfrac{315^2}{8} = 2821.875$ M1A1   3

Anything that rounds to 2820 for A1

(c) $b = \dfrac{4377.5}{S_{xx}}, = 1.537... = 1.5$ M1, A1

$a = \overline{y} - b\overline{x} = \dfrac{620}{8} - b\dfrac{315}{8} = 16.97... = 17.0$ M1, A1   4

Anything that rounds to 1.5 and 17.0 (accept 17)

(d)   Use overlay   B1ft
                                                                               B1ft   2

Follow through for the intercept for first B1.
Correct slope of straight line for second B1.

(e)   Brand D.   B1
since a long way above / from the line (dependent upon 'Brand D' above) B1
Using line: $y = 17 + 35 \times 1.5 = 69.5$ M1A1   4

Anything that rounds to 69p – 71p for final A1.
Reading from graph is acceptable for M1A1.
If value read from graph at $x = 35$ is answer given but out of range,
then award M1A0.

                                                                                         [15]

16. (a)   18-25 group, area = $7 \times 5 = 35$ B1
25-40 group, area = $15 \times 1 = 15$ B1   2

(b) $(25 - 20) \times 5 + (40 - 25) \times 1 = 40$ M1A1   2

5×5 is enough evidence of method for M1.
Condone 19.5, 20.5 instead of 20 etc.
Award 2 if 40 seen.

(c)   Mid points are 7.5, 12, 16, 21.5, 32.5   M1
$\Sigma f = 100$ B1

$$\frac{\sum ft}{\sum f} = \frac{1891}{100} = 18.91$$ M1A1   4

Look for working for this question in part (d) too.
Use of some mid-points, at least 3 correct for M1. These may be
tabulated in (d).

Their $\dfrac{\sum ft}{\sum f}$ for M1 and anything that rounds to 18.9 for A1.

(d) $\sigma_t = \sqrt{\dfrac{41033}{100} - \bar{t}^2}$ $\sqrt{\dfrac{n}{n-1}\left(\dfrac{41033}{100} - \bar{t}^2\right)}$ alternative OK   M1

                                                                                                                    M1

$\sigma_t = \sqrt{52.74...} = 7.26$ A1   3

Clear attempt at $\dfrac{41033}{100} - \bar{t}^2$ or $\dfrac{n}{n-1}\left(\dfrac{41033}{100} - \bar{t}^2\right)$ alternative

for first M1.
They may use their $\bar{t}$ and gain the method mark.
Square root of above for second M1
Anything that rounds to 7.3 for A1.

(e) $Q_2 = 18$ or 18.1 if $(n + 1)$ used   B1

$Q_1 = 10 + \dfrac{15}{16} \times 4 = 13.75$ or 15.25 numerator gives 13.8125   M1A1

$Q_3 = 18 + \dfrac{25}{35} \times 7 = 23$ or 25.75 numerator gives 23.15   A1   4

Clear attempt at either quartile for M1
These will take the form 'their lower limit' + correct fraction
× 'their class width'.
Anything that rounds to 13.8 for lower quartile.
23 or anything that rounds to 23.2 dependent upon method used.

(f) 0.376...   B1
     Positive skew   B1ft   2

Anything that rounds to 0.38 for B1 or 0.33 for B1 if $(n + 1)$ used.
Correct answer or correct statement that follows from their
value for B1.

                                                                                                            [17]

17. (a) Positive skew   (both bits)   B1   1

(b) $19.5 + \dfrac{(60-29)}{43} \times 10, = 26.7093....$ awrt **<u>26.7</u>**   M1, A1   2
     (N.B. Use of 60.5 gives 26.825… so allow awrt 26.8)

M1 for (19.5 or 20) + $\dfrac{(60-29)}{43} \times 10$ or better.

Allow 60.5 giving awrt 26.8 for M1A1

Allow their $0.5n$ [or $0.5(n + 1)$] instead of 60 [or 60.5] for M1.

(c) $\mu = \dfrac{3550}{120} = 29.5833...\text{ or } 29\dfrac{7}{12}$ awrt **29.6** B1

$\sigma^2 = \dfrac{138020}{120} - \mu^2 \text{ or } \sigma = \sqrt{\dfrac{138020}{120} - \mu^2}$ M1

$\sigma = 16.5829...\text{ or }(s = 16.652...)$ awrt **16.6** (or $s = 16.7$) A1 3

M1 for a correct expression for σ, σ^2, s or s^2.
NB $\sigma^2 = 274.99$ and $s^2 = 277.30$
Condone poor notation if answer is awrt16.6 (or 16.7 for s)

(d) $\dfrac{3(29.6 - 26.7)}{16.6}$ M1A1ft

$= 0.52...$ awrt **0.520** (or with s awrt 0.518) A1 3
(N.B. 60.5 in (b) ...awrt 0.499 [or with s awrt 0.497])

M1 for attempt to use this formula using their values to any accuracy. Condone missing 3.

1st A1ft for using their values to at least 3sf Must have the 3.

2nd A1 for using accurate enough values to get awrt 0.520
(or 0.518 if using s)
NB Using only 3 sf gives 0.524 and scores M1A1A0

(e) $0.520 > 0$ correct statement about their (d) being > 0 or < 0 B1ft
So it is consistent with (a) ft their (d) dB1ft 2

1st B1 for saying or implying correct sign for their (d).
B1g and B1ft. Ignore "correlation" if seen.

2nd B1 for a comment about consistency with their (d) and (a)
being positive skew, ft their (d) only
This is dependent on 1st B1: so if (d) > 0, they say yes,
if (d) < 0 they say no.

(f) Use <u>Median</u> B1
Since the data is skewed <u>or</u> less affected by outliers/extreme values dB1 2

2nd B1 is dependent upon choosing median.

(g) If the data are <u>symmetrical</u> or <u>skewness is zero</u> or
<u>normal/uniform distribution</u> B1 1
("mean =median" or "no outliers" or "evenly distributed" all score B0)

[14]

18. (a) Time is a <u>continuous</u> variable <u>or</u> data is in a <u>grouped</u> frequency table B1 1

(b)   Area is proportional to frequency <u>or</u> $A \propto f$ or $A = kf$ B1   1

      1st B1   for one of these correct statements.
                 "Area proportional to frequency density" or
                 "Area = frequency" is B0

(c) $3.6 \times 2 = 0.8 \times 9$ M1
                                            dM1
      1 child represented by 0.8   A1cso   3

      1st M1   for a correct combination of any 2 of the 4 numbers:
                 3.6, 2, 0.8 and 9

                 e.g. 3.6×2 or $\dfrac{3.6}{0.8}$ or $\dfrac{0.8}{2}$ etc   BUT e.g. $\dfrac{3.6}{2}$ is M0

      2nd M1   dependent on 1st M1 and for a correct combination of
                 3 numbers leading to 4th.
                 May be in separate stages but must see all 4 numbers

      A1cso   for fully correct solution.
                 Both Ms scored, no false working seen and <u>comment required.</u>

(d)   (Total) = $\dfrac{24}{0.8}$, = **30**   M1, A1   2

      M1   for $\dfrac{24}{0.8}$ seen or implied.

                                                           [7]

19.   (a)   Indicates max / median / min / upper quartile / lower quartile (2 or more) B1

          Indicates outliers (or equivalent description)   B1

          Illustrates skewness (or equivalent description e.g. shape)   B1   3

          Allows comparisons

          Indicates range / IQR / spread
               Any 3 rows

      (b)   (i)   37 (minutes)   B1

           (ii)   Upper quartile or Q_3 or third quartile or 75th percentage or P_{75} B1   2

(c) outliers

How to calculate correctly

'Observation that are very different from the other observations

and need to be treated with caution'   B1

These two children probably walked / took a lot longer   B1   2
    Any 2

(d)

Time (School B)
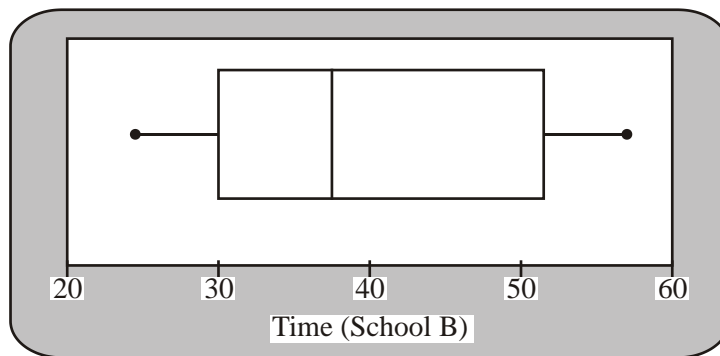
Box & median & whiskers	M1
Sensible scale	B1
30, 37, 50	B1
25, 55	B1 4

(e) Children from school A generally took less time   B1

50% of B ≤ 37 mins, 75% of A < 37 mins (similarly for 30)   B1

Median / Q1 / Q3 / of A < median / Q1 / Q3 / of (1 or more)   B1

A has outliers, (B does not)   B1   4

Both positive skew

IQR of A < IQR of B, range of A > range of B
    *Any **correct** 4 lines*

[15]

20. (a) P (both longer than 24.5) = $\frac{11}{55} \times \frac{10}{54} = \frac{1}{27}$ or $0.\dot{0}3\dot{7}$ or 0.037

2 fracs × w / o rep   M1 A1   2
    awrt 0.037

(b)   Estimate of mean time spent on their conversation is

$$\bar{x} = \frac{1060}{55} = 19\frac{3}{11} \text{ or } 19.\dot{2}\dot{7} \text{ or } 19.3 \qquad \text{M1 A1} \qquad 2$$

1060 / total, awrt 19.3 or 19 mins 16s

(c) $$\frac{1060 + \sum fy}{80} = 21 \qquad \text{B1}$$

21 × 80 = 1680

$$\sum fy = 620 \qquad \text{M1}$$

Subtracting 'their 1060'

$$\therefore \bar{y} = \frac{620}{25} = 24.8 \qquad \text{M1 A1} \qquad 4$$

Dividing their 620 by 25

(d)   Increase in mean value   B1

Length of conversation increased considerably

During 25 weeks relative to 55 weeks   B1ft   2
*Context- ft only from **comment** above*

[10]

21. (a)   Mode is 56   B1   1

(b) $Q_1 = 35$, $Q_2 = 52$, $Q_3 = 60$ B1, B1, B1   3

(c) $\bar{x} = \dfrac{1335}{27} = 49.4 \text{ or } 49\dfrac{4}{9}$ exact or awrt 49.4   B1

$$\sigma^2 = \frac{71801}{27} - \left(\frac{1335}{27}\right)^2 = 214.5432.... \qquad \text{M1 A1ft}$$

$\sigma = 14.6$ or 14.9 awrt 14.6(5) or 14.9   A1   4

(d) $\dfrac{49.4 - 56}{14.6} = -0.448$ awrt   range –0.44 to –0.46   M1A1   2

(e)   For negative skew;   M1
       Mean<median<mode   2 compared correctly   A1
       (49.4<52<56 not required)   3 compared correctly   M1
       $Q_3 - Q_2 < Q_2 - Q_1$ A1 ft   4

       8 and 17

       Accept other valid reason eg. 3(mean–median)/sd as alt for M1 A1

                                                                                      [14]

22. (a)   Distance is a continuous.   B1   1
                        continuous

        (b)   F.D = freq/class width $\Rightarrow$ 0.8, 3.8, 5.3, 3.7, 0.75, 0.1
                        or the same multiple of   M1 A1   2

        (c) $Q_2 = 50.5 + \dfrac{(67-23)}{53} \times 10 = 58.8$ M1 A1
                        awrt 58.8/58.9

               $Q_1 = 52.48; Q_3 = 67.12$ A1 A1   4
                        Special case: no working B1 B1 B1 ($\equiv$A's on the epen)

        (d) $\overline{x} = \dfrac{8379.5}{134} = 62.5335\ldots$ B1
                        awrt 62.5

               $s = \sqrt{\dfrac{557489.75}{134} - \left(\dfrac{8379.5}{134}\right)^2}$ M1 A1ft

               $s = 15.8089\ldots (S_{n-1} = 15.86825\ldots)$ A1   4
                        awrt 15.8 (15.9)
                        Special case: answer only B1 B1 ($\equiv$A's on the epen)

(e) $\dfrac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \dfrac{67.12 - 2 \times 58.8 + 52.48}{67.12 - 52.48}$ M1 A1ft

        subst their Q_1, Q_2 & Q_3 need to show working for A1ft
        and have reasonable values for quartiles

$= 0.1366 \Rightarrow$; +ve skew   A1; B1   4
      awrt 0.14

(f) For +ve skew Mean > Median & 62.53 > 58.80
<u>or</u> $Q_3 - Q_2$ (8.32) > $Q_2 - Q_1$ (6.32)
Therefore +ve skew   B1   1

                                                                      [16]

23. (a) $1.5 (Q_3 - Q_1) = 1.5(28 - 12) = 24$ B1
           may be implied

$Q_3 + 24 = 52 \Rightarrow 63$ is outlier
       Att Q3 + ... or Q1 − ...,   M1
       52 and −12 or   0 or evidence of no lower outliers   A1

$Q_1 - 24 < 0 \Rightarrow$ no outliers   A1
       63 is an outlier

                                                    M1
                                                       A1
                                                       A1   7

(b) Distribution is +ve skew; $Q_2 - Q_1$ (5) < $Q_3 - Q_2$ (11)   B1; B1   2

(c) Many delays are small so passengers should find these acceptable   B1   1
or sensible comment in the context of the question.

                                                                      [10]

24. (a) $Q_1 = 33$, $Q_2 = 41$, $Q_3 = 52$ B1B1B1   3

(b)

Scale & labels   B1
Box plot   M1
SV - Q_{11} Q_{31} Q_8 A1ft
          10, 34   A1
NC - 38, 45, 52   A1
          31, 42   A1   ⑥
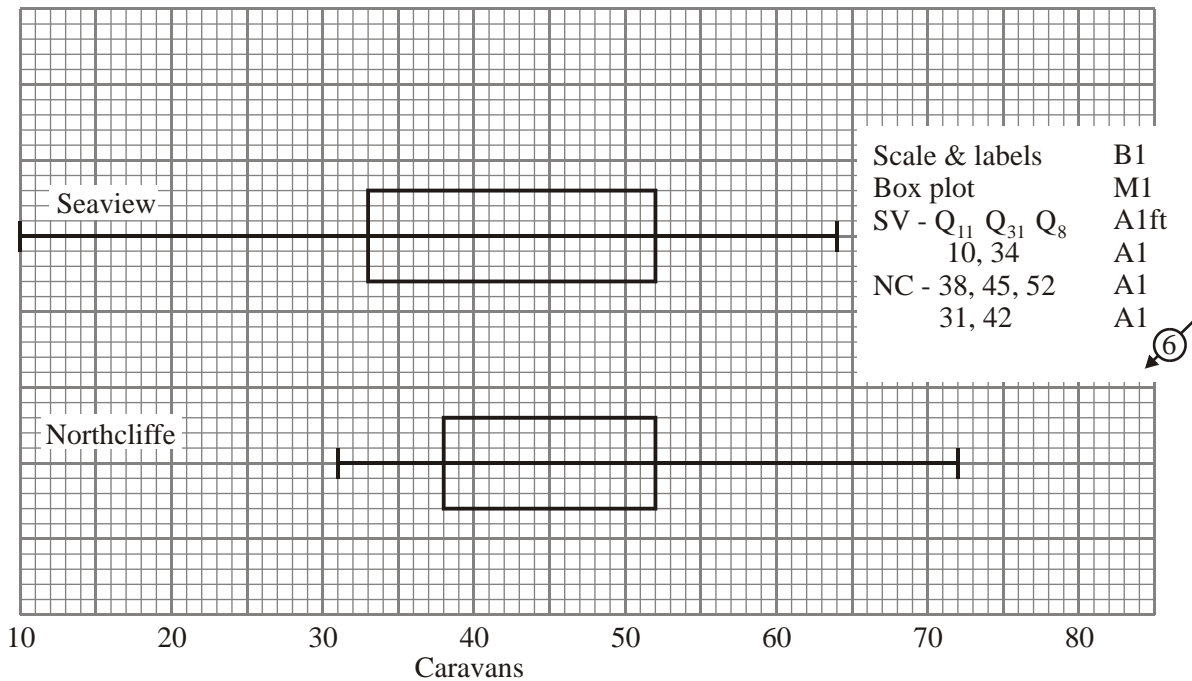
(c) Median of Northcliffe is greater than median of Seaview.   B1B1B1   3
Upper quartiles are the same
IQR of Northcliffe is   less than IQR of Seaview
Northcliffe positive skew, Seaview negative skew
Northcliffe symmetrical, Seaview positive skew (quartiles)
Range of Seaview greater than range of Northcliffe

any 3 acceptable comments

(d) On 75% of the nights that month   B1
both had no more than 52 caravans on site.   B1   2

[14]

25. (a) $a = 202$, $b = 202$, $c = 233$ B1,B1,B1 3

(b) $Q_1 - 1.5(Q_3 - Q_1) = 191 - 1.5(221 - 191) = 146$,
$Q_3 = 1.5(Q_3 - Q_1) = 221 + 1.5(221 - 191) = 266$
attempt at one calculation, 146, 266 M1A1A1
$\Rightarrow$ 269 is an outlier 269 A1dep
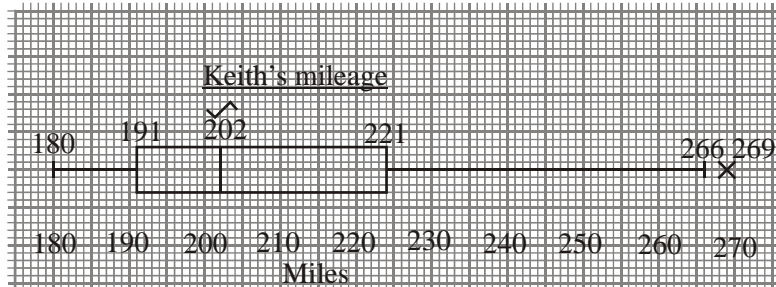
Scale and 'miles' B1
Box with two whiskers M1
191, their median, 221 A1ft
180,266 or 263,269 A1 8

(c) Keith: $Q_2 - Q_1 = 11$, $Q_3 - Q_2 = 19 \Rightarrow$ positive skew one calc, +ve skew M1,A1
Asif: $Q_2 - Q_1 = 16$, $Q_3 - Q_2 = 15 \Rightarrow$ almost symm or slight –ve skew A1 3

[14]

26. (a) Time data is a continuous variable B1 1

(b) 39.5, 44.5 both B1 1

(c)
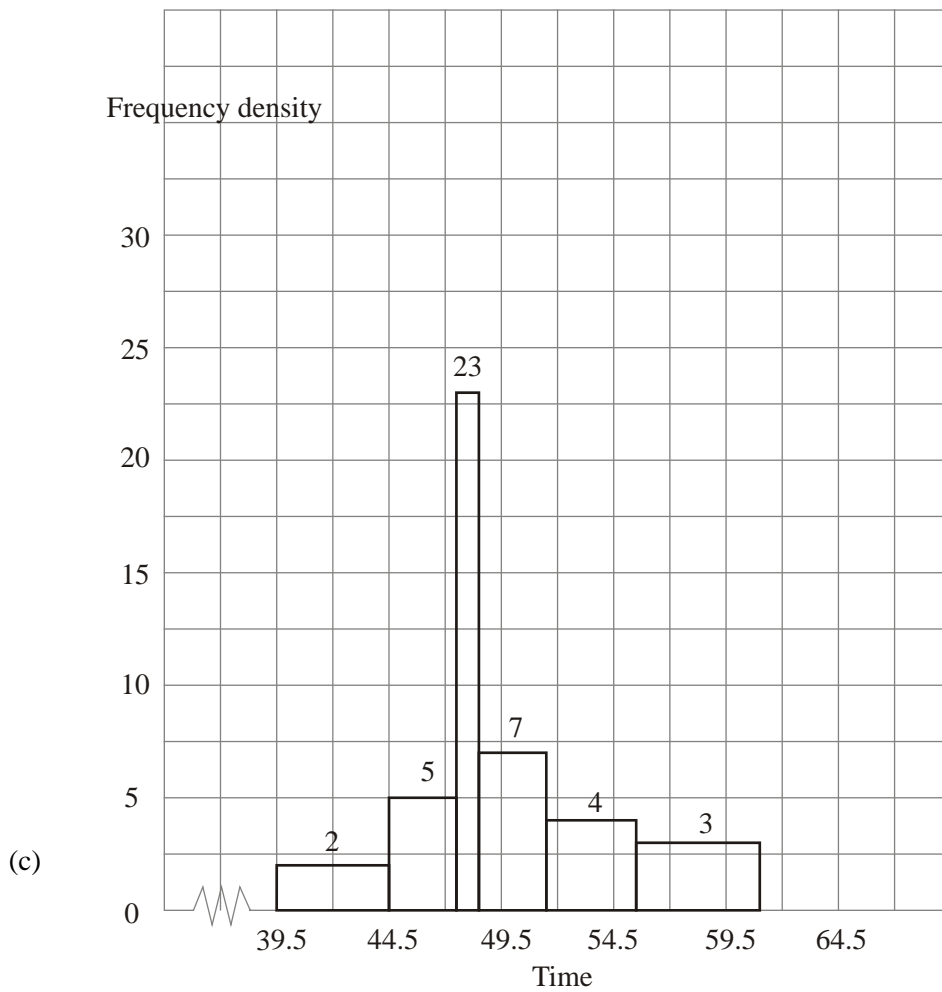
Freq / class width (implied)	M1
Scales and labels	B1
Histogram, no gaps & their fd	M1
All correct	A1 4

[6]

27. (a) (i) $\bar{x} = \dfrac{270}{16} = \underline{16.875}$ B1

16.875, 16⅞ ; 16.9; 16.88

s.d. $= \sqrt{\dfrac{4578}{16} - 16.875^2}$

$\dfrac{\sum x^2}{16} - \bar{x}^2$ & $\surd$ M1

All correct A1 ft

$= \underline{1.16592\ldots.}$

AWRT 1.17 *A1*

SR: No working B1 only

(ii) Mean % attendance $= \dfrac{16.875}{18} \times 100 \ (= 93.75)$ B1 ft 5

cao

(b)

	First 4|1 means 14		Second 1|8 means 18	
(1)	4	1	4 4 4	(3)
(1)	5	1	5 5 5 5	(4)
(3)	6 6 6	1	6 6 6	(3)
(5)	7 7 7 7 7	1	7	(1)
(6)	8 8 8 8 8 8	1	8 8 8	(3)
(0)		1	9	(1)
(0)		2	0	(1)

Both Labels and 1 key B1
Back-to-back S and L (ignore totals) M1
Sensible splits of 1 dep. M1
First-correct A1
Second - correct A1 5

(c)

	Mode	Median	IQR		
First (F)	18	17	2	B1 B1 B1	
Second (S)	15	16	3	B1 B1 B1	6

(d) $Median_S < Median_F$; $Mode_F > Mode_S$;   B1 B1 B1   3
      $IQR_s > IQR_F$; Only 1 student attends all
      classes in second; $Mean\%_F > Mean\%_S$

      Any THREE sensible comments

                                                                            [19]

28. (a)

Sales	No. of days	Class width	Frequency density
1-200	166	200	0.830
201-400	100	200	0.500
401-700	59	300	0.197
701-1000	30	300	0.100
1001-1500	5	500	0.010

      Frequency   M1
      densities   A1   5

      Graph
                  ③

NB Frequency densities can be scored on graph

Scales and Labels   B1
Bases   B1
Heights   B1ft

(3)
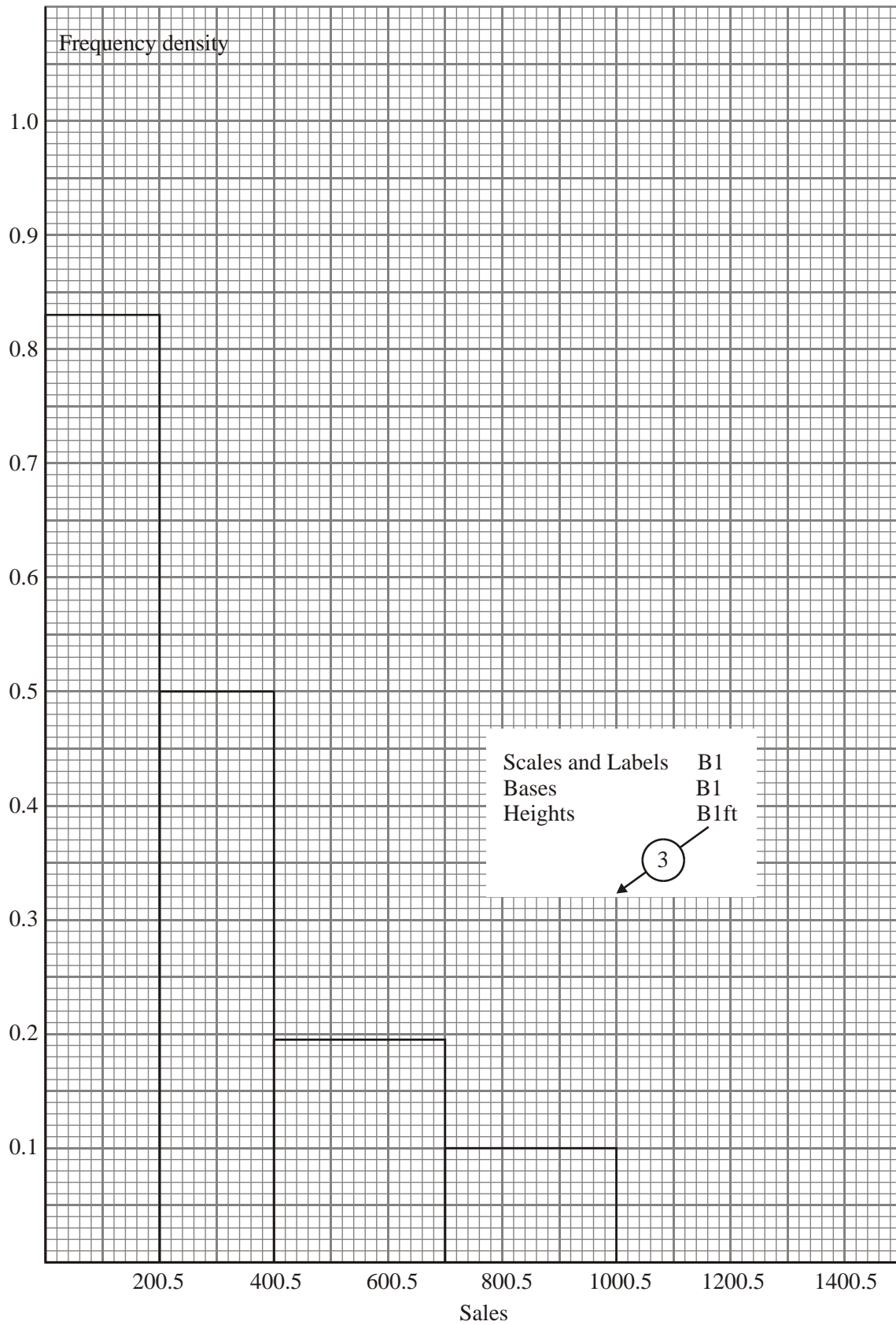
(b) $Q_2 = 200.5 + \dfrac{(180 - 166)}{100} \times 200 = \underline{228.5}$ 228/229/230   M1 A1

     $Q_1 = 0.5 + \dfrac{90}{166} \times 200 = \underline{108.933}\ldots$ 109 AWRT   A1

     $Q_3 = 400.5 + \dfrac{(270 - 266)}{59} \times 300 = \underline{420.838}$ AWRT 421/425   A1

     $(n = 270.75 \Rightarrow Q_3 = 424.6525)$

     IQR $= 420.830\ldots - 108.933\ldots = \underline{311.905}$ B1ft   5

(c) $\Sigma fx = 110980$; $\Sigma fx^2 = 58105890$ M1

     Attempt at Σfx or Σfy

     $\Sigma fy = 748$; $\Sigma fy^2 = 3943.5$ where $y = \dfrac{x - 100.5}{100}$ M1

     Attempt at Σfx^2 or Σfy^2

     $\mu = 308.277\dot{7}$ M1 A1   6

       308 AWRT

     $\sigma = 257.6238$

       258 AWRT

     No working shown: SR B1 B1 only for μ, σ.

(d) Median & IQR   B1

     Sensible reason e.g. Assuming other years are skewed.   B1 dep   2

                                                 [18]

29. (a) $\Sigma x = 12075$; $\Sigma x^2 = 15\,499\,685$

     $\therefore \bar{x} = \dfrac{12075}{15} = \underline{805}$ B1

       cao

     sd $= \sqrt{\dfrac{15499685}{15} - 805^2} = 620.71491$ M1

       $\sqrt{}$ *& correct method*

     3 s.f.   621   A1   3

     (NB Using $n - 1$ gives $642.50125\ldots$)   (643)

(b) 99, 169, 299, 350, 475, 485, 550, 650, 689, 830,   M1
     999, 1015, 1050, 2100, 2315

      Attempt to order

$\therefore Q_2 = \underline{650}$ A1

      cao 650

$\therefore IQR = Q_3 - Q_1 = 1015 - 350 = \underline{665}$

      Attempt at $Q_3 - Q_1$ M1
      cao 665   A1   4

(c) $Q_3 + 1.5(Q_3 - Q_1) = 1015 + 1.5 \times 665 = 2012.5$ M1

      Use of given outlier formula

$Q_1 - 1.5(Q_3 - Q_1) = 350 - 1.5 \times 665 < 0$ M1

      Evidence both ends considered

$\therefore$ 2100 and 2315 are outliers   A1   3

(d)
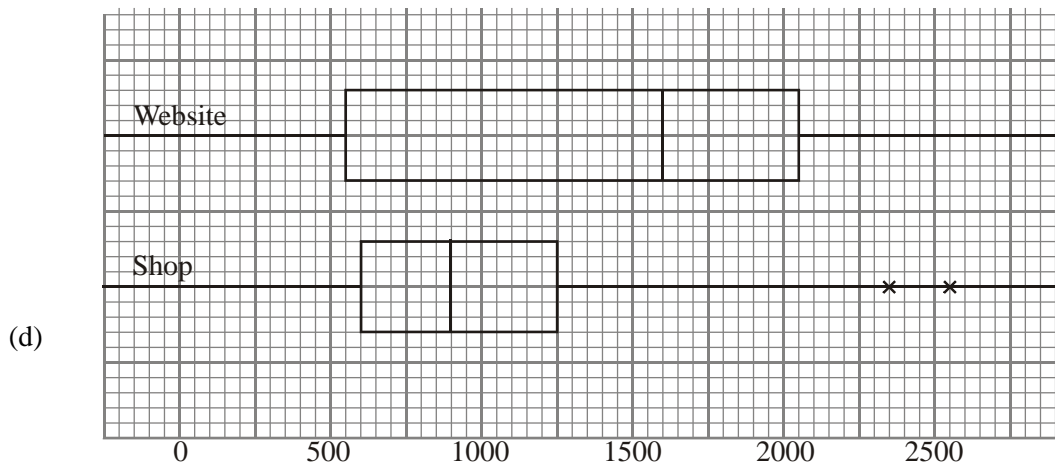
Two boxplots   B1
same scale
both labelled

Website   B1

Shop Box-plot   B1

Both outliers   B1   4

NB: For shop, right band whisker drawn to 2012.5 is acceptable.

(e)   Median website > median shop

Website negative skew; shop approx symmetrical
                Ignoring outliers

Ranges approximately equal
Shop Q_3 < Website Q_3 $\Rightarrow$ shop sales low value

Website sales more variable in value

$IQR_W \geq IQR_S$

Any two sensible comments   B1 B1   2
                                                                    [16]

30. Frequency densities: 3.0, 20.0, 9.0, 22.0, 3.0, 3.25

                Can be implied   M1

                from graph   A1

Scales and labels	B1
Bases	B1
Heights	B1 ft
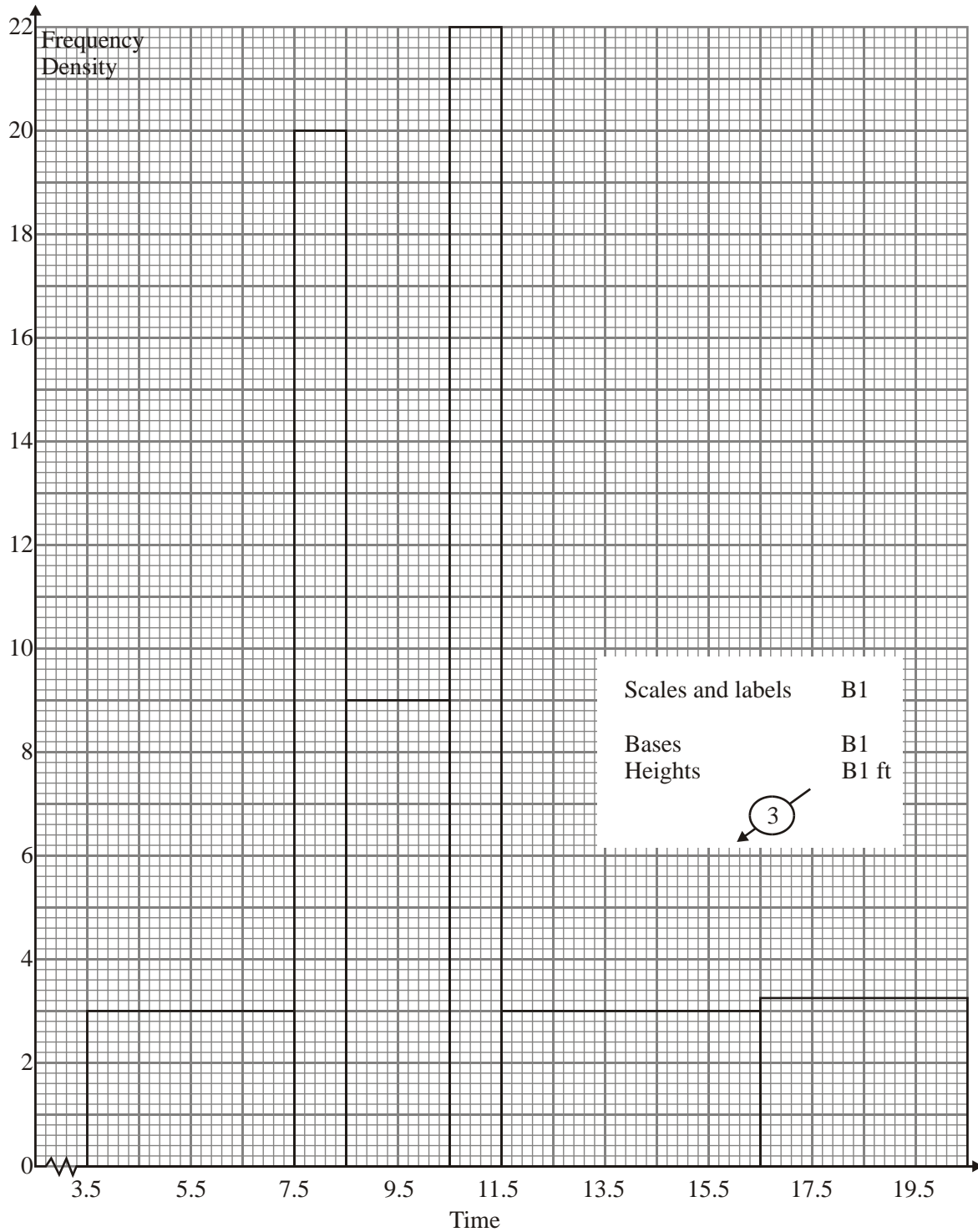
[5]

31.   (a) $\bar{x} = \dfrac{20 + 15 + \ldots + 17}{14} = \dfrac{312}{14} = 22.2857\ldots$ (awrt 22.3)   M1 A1   2

   (b)

Bags of crisps	1/0 means 10	Total
0	5	(1)
1	0 1 3 5 7	(5)
2	0 0 5	(3)
3	0 1 3	(3)
4	0 2	(2)

                Label & key   B1
                2 correct rows   B1
                All correct   B1   3

   (c) $Q_2 = 20$; $Q_1 = 13$; $Q_3 = 31$ B1; B1; B1   3

   (d) $1.5 \times \text{IQR} = 1.5 \times (31 - 13) = 27$ (can be implied)   B1
   $31 + 27 = 58$; $13 - 27 = -14$ (both)   M1
   No outliers   A1   3

   (e)

Box plot & Scale & Label   B1

$Q_1 Q_2 Q_3$ B1ft

5, 42   B1
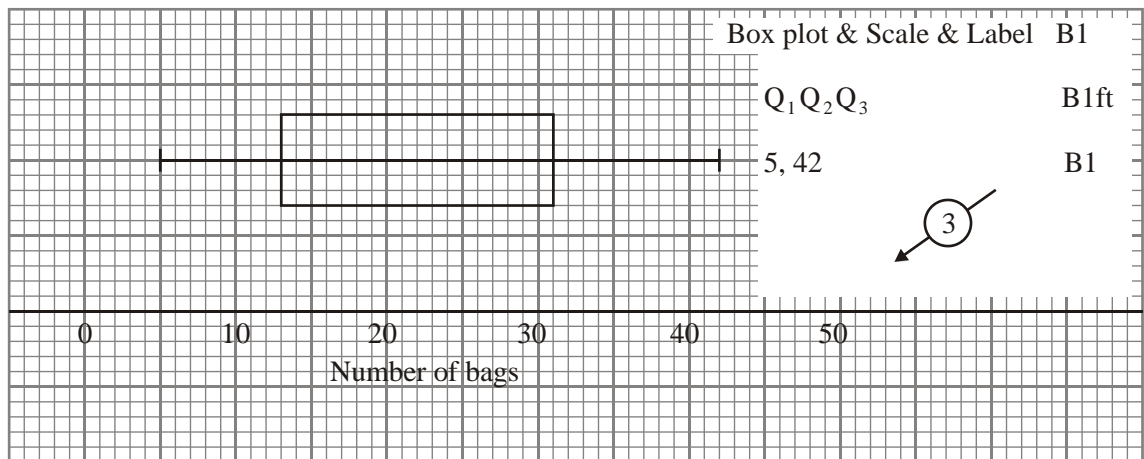
   (f) $Q_2 - Q_1 = 7$; $Q_3 - Q_2 = 11$; $Q_3 - Q_2 > Q_2 - Q_1$ M1
   Positive skew   A1   2

          [13]

32. Frequency densities: 0.16, 1.0, 1.0, 0.4, 0.4, 0.08   M1, A1

     Histogram: Scale and labels   B1

     Correct histogram   B1

     [4]

33. (a) $Q_2 = \dfrac{16+16}{2} = 16$; $Q_1 = 15$; $Q_3 = 16.5$; IQR = 1.5   M1 A1; B1; B1; B1   5

     (b) $1.5 \times$ IQR $= 1.5 \times 1.5 = 2.25$ M1 A1

         $Q_1 - 1.5 \times$ IQR $= 12.75 \Rightarrow$ no outliers below Q_1 A1

         $Q_3 + 1.5 \times$ IQR $= 18.75 \Rightarrow 25$ is an outlier   A1

         Boxplot, label scale   M1

         14, 15, 16, 16.5, 18.75 (18)   A1

         Outlier   A1   7

     (c) $\bar{x} = \dfrac{322}{20} = 16.1$ M1 A1   2

     (d)   Almost symmetrical/slight negative skew   B1

         Mean (16.1) $\approx$ Median (16) and $Q_3 - Q_2$ (0.5) $\approx Q_2 - Q_1$ (1.0)   B1   2

     [16]

34. (a)   Mode = 23   B1   1

         For Q_1: $\dfrac{n}{4} = 10.5 \Rightarrow$ 11th observation $\therefore Q_1 = 17$ B1

         For Q_2: $\dfrac{n}{2} = 21 \Rightarrow = \frac{1}{2}$ (21st & 22nd) observations

         $\therefore Q_2 = \dfrac{23+24}{2} = 23.5$ M1 A1

         For Q_3: $\dfrac{3n}{4} = 31.5 \Rightarrow$ 32nd observation $\therefore Q_3 = 31$ B1   4

(c)

Number of Daisies
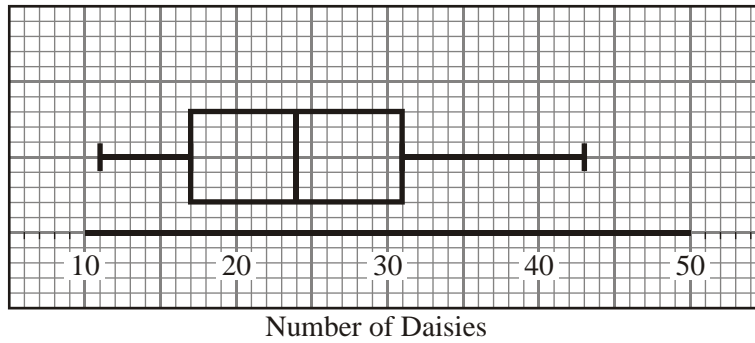
Box plot	M1
Scale & label	M1
Q_1, Q_2, Q_3	A1
11, 43	A1 4

(d)   From box plot or   M1

$Q_2 - Q_1 = 23.5 - 17 = 6.5$

$Q_3 - Q_2 = 31 - 23.5 = 7.5$(slight) positive skew   B1   1

(e)   Back-to-back stem and leaf diagram   B1   1

[11]

35.   (a) $\bar{y} = \dfrac{-467}{200}$ (can be implied)   B1

$\therefore \ \bar{x} = 2.5\bar{y} + 755.0$ M1

$= 2.5\left(\dfrac{-467}{200}\right) + 755.0$ A1

$= 749.1625$

(accept awrt 749)   A1

$S_y = \sqrt{\dfrac{9179}{200} - \left(\dfrac{-467}{200}\right)^2}$ M1 A1

$= 6.35946$ A1

$\therefore \ S_x = 2.5 \times 6.35946$ M1

$= 15.89865$ (accept awrt 15.9)   A1   9

(b)   Standard deviation $< \frac{2}{3}$ (interquartile range)   B1

        Suggest using standard deviation since it shows less variation
        in the lifetimes   B1   2

[11]

1.   Finding the midpoints of the given groups was predominantly carried out correctly with very few errors seen. In contrast, attempts at finding the width and height of the 26 –30 group were extremely varied, with most candidates finding this particularly challenging, especially in finding the height. In the majority of cases, candidates obtained the wrong width and height, mostly with no clear strategy, although these did multiply together to make 20.8 in some cases. Calculation of the mean was carried out successfully on the whole, although there were some apparent misconceptions, with quite a few candidates merely summing the midpoints (without multiplying by the frequency) and dividing this by 56.

The standard deviation proved to be more problematic, with frequent mistakes in both the formula and in their calculations. Some candidates used the sum of the f ^2x's and others the sum of the (f$x)^2$ or the sum of the fx's all squared in their formula. Very few candidates calculated s. Most candidates were able to use the correct interpolation technique to obtain the median, although many lost the accuracy mark through their use of 21 as the lower class boundary (which was relatively common) and /or 4 as the class width. Quite a few candidates worked with 28.5. A few candidates tried to apply the correct formula to the wrong class interval, however. Some candidates appeared to have a limited understanding of the class boundaries and failed to recognise the continuous nature of the data.

The majority of candidates were able to carry out a suitable test to determine the skewness of the data correctly. This mostly involved comparing $Q_3 - Q_2$ to $Q_2 - Q_1$ (with or without explicit substitutions), although the wrong conclusion was often drawn, either following on from a previous error in evaluating the median or from a lack of understanding of what their result was showing. A few students evaluated 3(mean – median)/standard deviation. Quite often the result of their test was described in words not figures, for example Q_2 is closer to Q_3 than Q_1. Some candidates merely attempted to describe the skewness without carrying out any test.

2.   Parts (a) and (b) were answered very well although a few candidates gave the upper quartile as 39 or 39.5 (usually as a result of incorrectly rounding $\dfrac{3n}{4}$) however the follow through marks meant that no further penalty need occur. A few found the upper and lower quartiles but failed to give the interquartile range. Most found the limit for an outlier using the given definition, although a few used $1.5 \times$ IQR, and went on to make a suitable comment about the one employee who needed retraining. There were some excellent box plots seen with all the correct features clearly present but a number failed to plot the outlier appropriately and simply drew their lower whisker to 7. A not insignificant minority were confused by the absence of an upper whisker and felt the need to add one usually at $Q_3 +$ IQR.

3.   Part (a) gave most candidates two easy marks but the rest of the question proved more demanding. The calculation of the mean in part (b) was usually answered well but there were still some dividing by 8 and a few using $\dfrac{\sum fx^2}{\sum fx}$. The calculation of the standard deviation was better than on previous occasions with many reaching 0.421 but there is still some confusion over the formula with $\sqrt{\left(15889.5 - \overline{x}^2\right)}$, a hybrid of the correct formula and S$_{xx}$, being quite common. Candidates should be aware that the formula for standard deviation is very sensitive to rounding errors and an accurate value for the mean (stored on their calculator) should be used rather than a rounded answer. A number of candidates failed to use the given values for Σfx and Σfx^2 and lost marks because of numerical slips. The attempts at interpolation in part (d) were

much improved with the correct fraction often being added to a lower class boundary. Unfortunately many used 2.95 or 2.5 as their class boundary and lost the marks. In part (e) the better candidates used their values for the mean and median and made an appropriate comment. Some spent the next page calculating Q_1 and Q_3 , often correctly, in order to use the quartiles to justify their description of the skewness.

4. There were many good answers to this question. The Venn diagram was often totally correct although a number failed to subtract for the intersections and obtained value of 35, 40 and 28 instead of 31, 36 and 24 for the numbers taking two options. Parts (b) and (c) were answered very well with only a minority of candidates failing to give probabilities. Part (d) proved straightforward for those who knew what was required but some attempted complicated calculations, often involving a product of probabilities, whilst others simply gave their answer as 4/180.

5. Although there were more correct solutions than in previous papers for this type of question the process required to answer this question was not applied successfully by a large number of candidates. The most common error in part (a) was to give an answer of 0.8. In tandem with this was an answer to part (b) of 7.5 where candidates recognised that the answer to part (a) times the answer to part (b) should be 6. Many candidates divided 9 by 3 in part (b) but failed to multiply by 2. Other candidates however produced two correct answers but nothing else. The variety of approaches may suggest some logical thinking rather than a taught approach to this type of problem.

6. Very few candidates got full marks for this question, being unable to perform the calculations for grouped data, although the mean caused the least problems. Those candidates with good presentation particularly those who tabulated their workings tended to fare better. In spite of the well defined groups many candidates subtracted or added 0.5 to the endpoints or adjusted the midpoints to be 0.5 less than the true value with the majority getting part (a) incorrect as a result. As usual all possible errors were seen for the calculation of Σfx^2 i.e. $(\Sigma fx)^2$, $\Sigma (fx)^2$, Σf^2x and Σx^2. Use of 17.1 for the mean in the calculation of the standard deviation led to the loss the accuracy mark. Candidates are once again reminded not to use rounded answers in subsequent calculations even though they usually gain full marks for the early answer. The comment in part (c) was often forgotten perhaps indicating that candidates are able to work out the figures but do not know what they mean, although many did appreciate in part (d) that there is no skew in a normal distribution. As opposed to question 1, correlation was often mentioned instead of skewness although again this is becoming less common.

7. This question was usually answered well. In part (b) some did not realise that they needed to check the lower limit as well in order to be sure that 110 was the only outlier. Part (c) was answered very well although some lost the last mark because there was no gap between the end of their whisker and the outlier. Part (d) was answered very well and most gave the correct values for $\sum y$ and $\sum y^2$ in the appropriate formula. A few tried to use the $\sum (y - \bar{y})^2$ approach but this requires all 10 terms to be seen for a complete "show that" and this was rare.

Part (e) was answered well although some gave the answer as –5.7 having forgotten the 10^{-3}, or failed to interpret their calculator correctly. Many candidates gave comments about the

correlation being small or negative in part (f) but they did not give a clear reason for rejecting the parent's belief. Once again the interpretation of a calculated statistic caused difficulties.

8.   Part (a) was not answered well. Many candidates attempted to calculate frequency densities but they often forgot to deal with the scale factor and the widths of the classes were frequently incorrect. There are a variety of different routes to a successful answer here but few candidates gave any explanation to accompany their working and it was therefore difficult for the examiners to give them much credit. The linear interpolation in part (b) was tackled with more success but a number missed the request for the Inter Quartile Range. Whilst the examiners did allow the use of $(n + 1)$ here, candidates should remember that the data is being treated as continuous and it is therefore not appropriate to "round" up or down their point on the cumulative frequency axis. Although the mean was often found correctly the usual problems arose in part (c) with the standard deviation. Apart from those who rounded prematurely, some

forgot the square root and others used $\sum f^2 x$ or $\left(\sum fx\right)^2$ or $\dfrac{\sum fx^2}{\sum fx}$ instead of the correct first

term in their expression and there was the usual crop of candidates who used $n = 6$ instead of 104. The majority were able to propose and utilise a correct test for the skewness in part (d) with most preferring the quartiles rather than the mean and median. Few scored both marks in part (e) as, even if they chose the median, they missed mentioning the Inter Quartile Range. A number of candidates gave the mean and standard deviation without considering the implications of their previous result.

9.   Most candidates were able to draw a six-branched tree diagram correctly, although a number of candidates had incorrect or missing labels. From a correct diagram most gained full marks in part (b). The conditional probability in part (c) once again caused difficulty for many of the candidates. Many of the responses in part (d) were, incorrectly, referring to the importance of testing people for a disease rather than referring to the probability in part (c).

10.   Parts (a) to (d) represented a chance for all students who had an average grasp of statistics to score highly. The median in part (b) was incorrectly identified by a significant number of candidates, but the standard deviation was often correct.

Part (e) was done surprisingly well with students appearing to have a much greater understanding of what is required for a comparison than in previous years. Often numbers were stated without an actual comparison. Confusion was evident in some responses as skewness was often referred to as correlation. A small minority of candidates had failed to take note of the 'For the Balmoral Hotel' and had done some correct statistics for all 55 students.

11.   A lot of fully correct Venn Diagrams were seen in part (a) although it was surprising the number who resorted to decimals rather than just using straightforward fractions; this often led to loss of many accuracy marks. A significant minority had negative numbers in their Venn diagram and saw nothing wrong in this when converting them to probabilities later in the question. Fewer candidates forgot the box this time. Part (c) proved to be the only difficult part, as many candidates struggled with the concept of conditional probability, and many denominators of 300 were seen.

12.   The mean was calculated accurately by the majority of the candidates but the standard deviation calculation still caused problems for many. There were few summation errors but missing square roots or failing to square the mean were some of the more common errors.

   Part (b) was poorly answered. The examiners were disappointed that such a sizeable minority failed to order the list and worked quite merrily with a median larger than their upper quartile. Some worked from the total of 2757 to get quartiles of 689.5, 1378.5 and 2067.5 whilst others used cumulative totals and obtained quartiles in the thousands but still failed to see the nonsensical nature of these figures. Those who did order the list used a variety of methods to try and establish the quartiles. Whilst the examiners showed some tolerance here any acceptable method should give a median of 198 but many candidates used 186. Those who knew the rules usually scored full marks here and in parts (c) and (d).

   The examiners followed through a wide range of answers in part (c) and most candidates were able to secure some marks for correctly identifying patients *B* and *F* and in part (d) for describing their skewness correctly.

13.   The common error here was to assume that frequency equals the area under a bar, rather than using the relationship that the frequency is <u>proportional</u> to the area under the bar. Many candidates therefore ignored the statement in line 1 of the question about the histogram representing 140 runners and simply gave an answer of $12 \times 0.5 = 6$. A few candidates calculated the areas of the first 7 bars and subtracted this from 140, sadly they didn't think to look at the histogram and see if their answer seemed reasonable. Those who did find that the total area was 70 usually went on to score full marks. A small number of candidates had difficulty reading the scales on the graph and the examiners will endeavor to ensure that in any future questions of this type such difficulties are avoided. A small number of candidates had difficulty reading the scales on the graph and the examiners will endeavor to ensure that in any future questions of this type such difficulties are avoided.

14.   Parts (a), (b) and (c) were generally well done, although in part (c) there were many with strange ideas of heavy instruments. In part (d) the majority of candidates were able to make a credible attempt at this with most giving one of the two possible solutions with a reason. The majority used the median and quartiles to find that the distribution was symmetrical. The use of the words 'symmetrical skew', similar to 'fair bias', is all too often seen but was accepted. Equal, even or normal skew were also often seen and were given no credit.
   Part (e) was attempted successfully by a minority of candidates. A large number of candidates did not understand the distinction between z–values and probabilities. A lot gave 0.68 as z-value leading to the loss of the accuracy mark. Others tried to put various values into standard deviation formulae.

15.   This was generally well answered. The majority of the errors occurred in part (c) by rounding too early and getting 18.4 for *a*. The regression line was often inaccurately plotted. In part (e) many used chocolate content to justify answer and often did not use the regression line to get a suitable price. Some misunderstood the question and attempted to find the best value in the second part of part (e)

16.   Many candidates started well with this question, but a large number of inaccurate answers were seen for the latter parts. Part (a) was usually correct and part (b) was generally done well. In part (c) there were a lot of mistakes in finding midpoints and also $\sum f$. Most knew the correct method

for finding the mean, but rather fewer knew how to find the standard deviation in part (d) although most remembered to take the square root. Part (e) was very badly answered, with the majority unable to interpolate correctly which was often due to wrong class boundaries and / or class widths. In part (f), although the majority got an incorrect numerical value, most picked up the mark for interpreting their value correctly.

17. This question caused problems for many candidates. Part (a) did not always generate a comment about the skewness of the data and many who did eventually mention skewness thought it was negative. The calculation of the median in part (b) often caused difficulties. An endpoint of 19.5 was often used, but some thought the width was 9 not 10 and many simply opted for the midpoint of 24.5. The calculation of the mean in part (c) was sometimes the only mark scored by the weakest candidates and the examiners were disappointed at how many candidates were unable to find the standard deviation. Aside from the usual error of missing the square root or failing to square the mean, a number were using formulae such as $\sqrt{\dfrac{\sum fx^2}{\sum fx}}$. Most scored some marks in part (d) for attempting to use their values in the given formula, but the final mark required an answer accurate to 3 sf and this was rarely seen. In part (e) many failed to comment on the sign of their coefficient and there was often a discussion of correlation here rather than skewness. Of those who attempted the last two parts, part (g) was often successful, but in part (f), candidates often chose the mean because it used all the data rather than the median, which wouldn't be affected by the extreme values.

18. Parts (a) and (b) were not answered well. Few mentioned the type of variable in part (a) and in part (b) many simple stated that the frequency equals the area rather than stating that it was proportional to the area.

    Many were able to give a correct calculation in part (c) but they sometimes failed to state that the 0.8 related to each individual child; the question was a "show that" and a final comment was required. The calculation in part (d) was usually correct.

19. Part (a) often scored full marks although some still mention 'mean' instead of 'median'. Part (d) was very straightforward for the vast majority of candidates. Those candidates who used a scale of 4cm to 10 units were sometimes prone to placing the median inaccurately. Part (e) was also quite well done but some only listed the 5 important values with little or no mention of IQR, range, outliers or skewness. There was occasional confusion thinking the bigger numbers meant school B had done better.

20. In part (a) there were very few correct solutions. It was rare for a candidate to appreciate that the selection was without replacement. The rest of this question was well answered by many, although a surprising number averaged the two means in part (c).

21. This question was a good source of marks with most candidates able to find the correct values for the mode and median, but too many getting the upper quartile wrong. A surprising number of candidates had problems with finding the standard deviation. In quite a few cases the square

root was omitted but more often marks were lost due to the misinterpretation and misuse of standard formulae. This was not helped by some candidates ignoring given totals and calculating their own. In part (e) a large number of responses gave one reason rather than two.

22.   It was very disappointing that so few candidates could carry out a simple analysis of a set of data. Few scored well.

(a)   Relatively few candidates were able to state, "distance is a continuous variable." The most common wrong answer in this part referred to the unequal class widths.

(b)   In general frequency densities were well done. The most common mistake was to calculate the incorrect class width, taking the first class width as 4. Other mistakes were class width divided by frequency or frequency multiplied by class width although these were less prevalent than in previous examinations.

(c)   Interpolation was not familiar to many candidates. Those pupils who did attempt to interpolate to find the median and quartiles were on the whole successful, common errors being the use of 50 instead of 50.5 or the wrong class interval. Many used the mid-point of the class for the quartiles or more frequently used 134/2 or (134 + 1)/2 as their responses for an estimate of the median.

(d)   The mean was calculated successfully by the vast majority of candidates with only occasional error through using the sum of fx^2 as opposed to the sum of fx. The standard deviation proved more difficult – where students used the wrong formula, omitted the square root or lost accuracy marks through using the rounded value of the mean. Some candidates wasted time by recalculating the values given.

(e)   Those candidates with sensible values for their quartiles managed to substitute successfully to calculate the coefficient although it was surprising how many could not get 0.14 from a correct expression. On the whole they drew the correct conclusion about the data being positively skewed, although a small number of candidates managed a correct calculation and then concluded negative skewness.

(f)   Although a fair number of students could give a reason to confirm that the skewness was positive, most lost this mark by not justifying their comment using numerical values.

23.   (a)   The vast majority of candidates were able to make an attempt at drawing a box plot though labels were not always added and the upper whisker often extended to 63. For many candidates this was the only mark they obtained for the question. Few candidates bothered, or were able, to use the information regarding 1.5 IQR in order to identify the limits of acceptable data. Of those candidates who did show some working more often than not, they did not do so in enough detail. The number 24 was usually implied but candidates often ignored showing working for the lower end. The numbers of 52 and -12 were visible on a number of papers, but the conclusion about which numbers were outliers was often omitted.

(b) The majority of candidates recognized positive skewness, but many did not justify their answer numerically. Some candidates did not understand the request about the "distribution of delays" and gave an interpretation more suited to part (c).

(c) Most managed a comment on the distribution that was relevant, but few wrote in terms of whether passengers would be bothered by the delays – the majority of students used technical statistical terms, referring to quartiles and percentages of the data, rather than simply interpreting the data in non-technical language.

24. A lack of detailed labelling in the box plot was common. Candidates should realise that 3 marks for parts where they are comparing etc. requires them to find three relevant points. Many only had one or 2 points and seemed to think that if they wrote enough about one point they could get the 3 marks. The last part was not well interpreted by many. They were likely to just say that the 2 values for Q_3 were the same. Most candidates can find quartiles and know how to display the information in box plots. There are still some candidates who do not draw a clearly labelled axis for their scale. Candidates need to remember that the purpose is to compare data so the scale needs to be the same for both sets of data. Some candidates can give good comparisons referring to range, IQR, median and quartiles, but many give vague descriptions concerning 'spread' and 'average' which gain no marks. They should be encouraged to be specific in their descriptions. Very few can interpret the upper quartile in context.

25. Many candidates were able to calculate the median and both quartiles accurately; the most common error was to give Keith's median as 201. Some candidates ignored or failed to show the calculation of outliers. However, the vast majority of candidates were able to draw a reasonable box plot although the scale was often unlabelled. Other common errors were to extend the left hand whisker to 146 and the right hand whisker to 269 with 266 marked on as a bound for outliers.

In part (c), a failure to show working was all too common. Asif's distribution was often said to be "negatively skewed", with only a minority qualifying it as weak or almost symmetrical. Some candidates confused negative skew with positive skew.

26. Part (a) produced a poor response, with very few candidates realising when a histogram should be used. Part (b) was answered correctly by a very large majority of candidates. Most candidates appreciated the need for frequency densities and they were usually calculated accurately. The chosen scale was often good, but unsuitable scales are still seen too frequently at this level. Candidates generally labelled their axes. The heights if the rectangles were usually correctly plotted, but unsuitable scales sometimes proved a hindrance to candidates.

27. Candidates knew how to tackle this question but too many of them did not pay sufficient attention to detail. They often calculated the variance and not the standard deviation and their arithmetic was not always as accurate as it should have been. Some poor computational methods were seen when calculating the standard deviation. The stem and leaf diagram caused few problems for the candidates but few of them presented it in the most appropriate form. The mode and quartiles were often correct and it was pleasing to see many of them making a good attempt to compare and contrast the attendance data.

28. This was a long question that needed an understanding of several different but frequently linked concepts and many of the candidates did not have the stamina for such a question. Common errors included the drawing of a bar chart; fd = class width/frequency; poorly drawn histograms; quartiles calculated without using a lower class boundary; IQR omitted; a variety of incorrect mid-points; poor arithmetic and no appreciation of the skewness of the data. For a routine type of question the overall response was very disappointing.

29. This question posed few conceptual problems for the candidates but few of them gained full marks. The majority of lost marks were the result of candidates not paying sufficient attention to detail. The standard deviation was rarely given to an appropriate degree of accuracy and some candidates did not look for outliers at both ends of the data set. Although choosing a scale that would fit on the grid supplied for the candidates was not easy, most managed to do so but then forgot to label the axis or the two box plots. Others ignored the outliers even though they had identified them. Whilst candidates tried to find two sensible comparisons very few expressed them clearly.

30. There was still some uncertainty about how to calculate frequency densities. There were examples of candidates using (class width)/frequency or mid-point × frequency. Wrongly labelled axes, poor scales and wrong bases for the histogram bars lost marks for many candidates. A completely correct histogram was not very common.

31. Most candidates could make a reasonable attempt at this question. The stem and leaf diagram rarely had a label and although almost all candidates gave the correct value for the median, far too many did not give correct values for the other quartiles. Showing that there were no outliers was not always well attempted since many could not calculate the IQR correctly. The box plot was often spoiled by having no label, a poor or no proper scale and inaccurate plotting. Positive skewness was usually recognised and a correct justification was often given.

32. This question was generally well answered but some candidates could not work out frequency densities correctly. Many histograms were poorly labelled and many candidates gained marks on this question only because examiners followed through their frequency densities. There were still too many candidates who drew bar charts instead of histograms.

33. The upper quartile caused problems for many candidates but they should have seen an example with 20 observations and been able to deal with the quartiles. The values obtained by the candidates were followed through and this allowed many of them to score most of the marks. Working for the outliers was often omitted, with a loss of marks as stated in the rubric, and the label on the box plot was often missing. The mean was usually correct and some attempt to comment on the skewness was nearly always made.

34. No Report available for this question.

35. No Report available for this question.