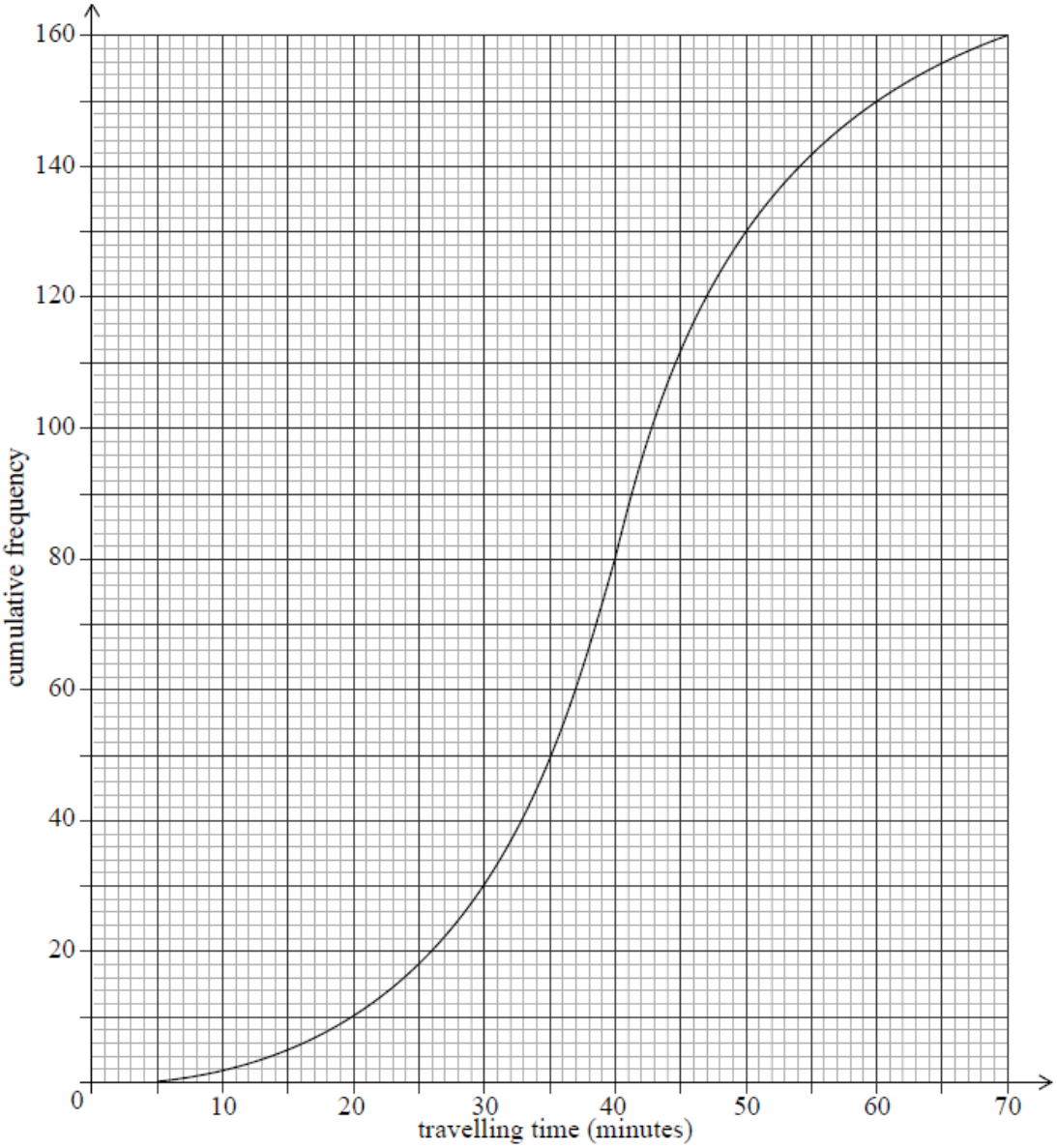
Stats *[88 marks]*

A large company surveyed 160 of its employees to find out how much time they spend traveling to work on a given day. The results of the survey are shown in the following cumulative frequency diagram.



1a. Find the median number of minutes spent traveling to work. *[2 marks]*

Markscheme

evidence of median position ***(M1)***

80th employee

40 minutes ***A1***

[2 marks]

1b. Find the number of employees whose travelling time is within 15 minutes of the median. *[3 marks]*

Markscheme

valid attempt to find interval (25–55) ***(M1)***

18 (employees), 142 (employees) ***A1***

124 ***A1***

[3 marks]

Only 10% of the employees spent more than k minutes traveling to work.

1c. Find the value of k. *[3 marks]*

Markscheme

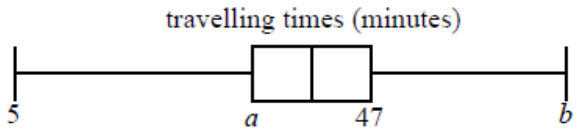
recognising that there are 16 employees in the top 10% ***(M1)***

144 employees travelled more than k minutes ***(A1)***

$k = 56$ ***A1***

[3 marks]

The results of the survey can also be displayed on the following box-and-whisker diagram.



1d. Write down the value of b. *[1 mark]*

Markscheme

$b = 70$ ***A1***

[1 mark]

1e. Find the value of a. *[2 marks]*

Markscheme

recognizing a is first quartile value ***(M1)***

40 employees

$a = 33$ ***A1***

[2 marks]

1f. Hence, find the interquartile range. *[2 marks]*

Markscheme

$47 - 33$ ***(M1)***

IQR = 14 ***A1***

[2 marks]

1g. Travelling times of less than p minutes are considered outliers. *[2 marks]*

Find the value of p.

Markscheme

attempt to find 1.5 × **their** IQR ***(M1)***

33 − 21

12 ***(A1)***

[2 marks]

---

2. A data set consisting of 16 test scores has mean 14.5 . One test score of 9 requires a second marking and is removed from the data set. *[4 marks]*

Find the mean of the remaining 15 test scores.

Markscheme

* This sample question was produced by experienced DP mathematics senior examiners to aid teachers in preparing for external assessment in the new MAA course. There may be minor differences in formatting compared to formal exam papers.

$\frac{\sum_{i=1}^{16} x_i}{16} = 14.5$ **(M1)**

Note: Award **M1** for use of $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

$\Rightarrow \sum_{i=1}^{16} x_i = 232$ **(A1)**

new $\bar{x} = \frac{232-9}{15}$ **(A1)**

$= 14.9 \left(= 14.8\overline{6}, = \frac{223}{15}\right)$ **A1**

Note: Do not accept 15.

[4 marks]

The principal of a high school is concerned about the effect social media use might be having on the self-esteem of her students. She decides to survey a random sample of 9 students to gather some data. She wants the number of students in each grade in the sample to be, as far as possible, in the same proportion as the number of students in each grade in the school.

3a. State the name for this type of sampling technique. *[1 mark]*

Markscheme

Stratified sampling **A1**

[1 mark]

The number of students in each grade in the school is shown in table.

Grade	Number of Students
9	60
10	83
11	33
12	84

3b. Show that 3 students will be selected from grade 12. *[3 marks]*

Markscheme

There are 260 students in total **A1**

$\frac{84}{260} \times 9 = 2.91$ **M1A1**

So 3 students will be selected. **AG**

[3 marks]

3c. Calculate the number of students in each grade in the sample. *[2 marks]*

Markscheme

grade $9 = \frac{60}{260} \times 9 \approx 2$, grade $10 = \frac{83}{260} \times 9 \approx 3$, grade $11 = \frac{33}{260} \times 9 \approx 1$ **A2**

[2 marks]

In order to select the 3 students from grade 12, the principal lists their names in alphabetical order and selects the 28th, 56th and 84th student on the list.

3d. State the name for this type of sampling technique. *[1 mark]*

Markscheme

Systematic sampling **A1**

[1 mark]

Once the principal has obtained the names of the 9 students in the random sample, she surveys each student to find out how long they used social media the previous day and measures their self-esteem using the Rosenberg scale. The Rosenberg scale is a number between 10 and 40, where a high number represents high self-esteem.

Student	A	B	C	D	E	F	G	H	I
Time spent on social media, t (hours)	3	1.2	2.5	4.1	4.7	3.6	2.9	1.7	0.5
Self-Esteem, s (Rosenberg Scale)	25	33	26	20	21	22	23	25	31

3e. Calculate Pearson's product moment correlation coefficient, r. *[2 marks]*

Markscheme

$r = -0.901$ **A2**

[2 marks]

3f. Interpret the meaning of the value of r in the context of the principal's concerns. *[1 mark]*

Markscheme

The negative value of r indicates that more time spent on social media leads to lower self-esteem, supporting the principal's concerns. **R1**

[1 mark]

3g. Explain why the value of r makes it appropriate to find the equation of a regression line. *[1 mark]*

Markscheme

r being close to –1 indicates there is strong correlation, so a regression line is appropriate. ***R1***

[1 mark]

3h. Another student at the school, Jasmine, has a self-esteem value of 29. *[4 marks]*

By finding the equation of an appropriate regression line, estimate the time Jasmine spent on social media the previous day.

Markscheme

Find the regression line of t on s. ***M1***

$t = -0.281s + 9.74$ ***A1***

$t = (-0.2807\ldots)(29) + 9.739\ldots = 1.60$ hours ***M1A1***

[4 marks]

The following table shows the Mathematics test scores (x) and the Science test scores (y) for a group of eight students.

Mathematics scores (x)	64	68	72	75	80	82	85	86
Science scores (y)	67	72	77	76	84	83	89	91

The regression line of y on x for this data can be written in the form $y = ax + b$.

4a. Find the value of a and the value of b. *[2 marks]*

Markscheme

$1.01206\ldots,\ 2.45230\ldots$

$a = 1.01,\ b = 2.45\ (1.01x + 2.45)$ ***A1A1***

[2 marks]

4b. Write down the value of the Pearson's product-moment correlation coefficient, r. *[1 mark]*

Markscheme

$0.981464\ldots$

$r = 0.981$ ***A1***

Note: A common error is to enter the data incorrectly into the GDC, and obtain the answers $a = 1.01700\ldots$, $b = 2.09814\ldots$ and $r = 0.980888\ldots$ Some candidates may write the 3 sf answers, ie. $a = 1.02$, $b = 2.10$ and $r = 0.981$ or 2 sf answers, ie. $a = 1.0$, $b = 2.1$ and $r = 0.98$. In these cases award ***A0A0*** for part (a) and ***A0*** for part (b). Even though some values round to an accepted answer, they come from incorrect working.

[1 mark]

4c. Use the equation of your regression line to predict the Science test score for a student who has a score of 78 on the Mathematics test. Express your answer to the nearest integer. *[2 marks]*

Markscheme

correct substitution of 78 into **their** regression equation ***(M1)***

$81.3930\ldots$ 81.23 from 3 sf answer

81 ***A1***

[2 marks]

The number of hours spent exercising each week by a group of students is shown in the following table.

Exercising time (in hours)	Number of students
2	5
3	1
4	4
5	3
6	x

The median is 4.5 hours.

5a. Find the value of x. *[2 marks]*

Markscheme

EITHER

recognising that half the total frequency is 10 (may be seen in an ordered list or indicated on the frequency table) ***(A1)***

OR

$5 + 1 + 4 = 3 + x$ ***(A1)***

OR

$\sum f = 20$ ***(A1)***

THEN

$x = 7$ ***A1***

[2 marks]

5b. Find the standard deviation. *[2 marks]*

Markscheme

METHOD 1

$1.58429\ldots$

1.58 ***A2***

METHOD 2

EITHER

$\sigma^2 = \frac{5\times(2-4.3)^2+1\times(3-4.3)^2+4\times(4-4.3)^2+3\times(5-4.3)^2+7\times(6-4.3)^2}{20}\ (=2.51)$

(A1)

OR

$\sigma^2 = \frac{5\times2^2+1\times3^2+4\times4^2+3\times5^2+7\times6^2}{20} - 4.3^2\ (=2.51)$ ***(A1)***

THEN

$\sigma = \sqrt{2.51} = 1.58429\ldots$

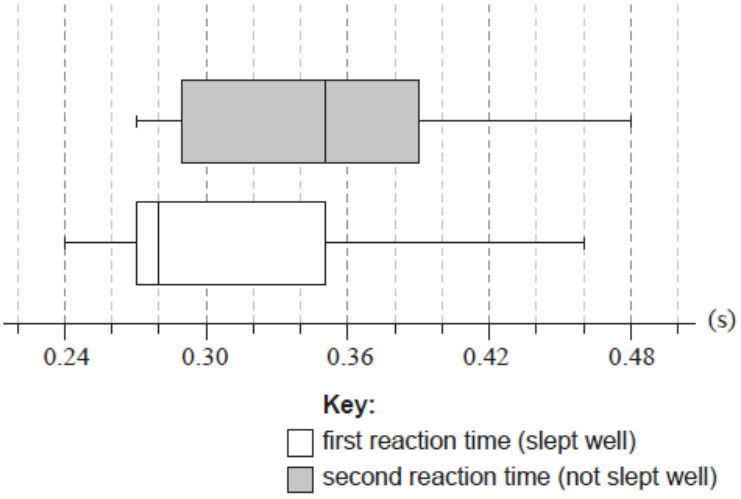
$= 1.58$ ***A1***

[2 marks]

A random sample of nine adults were selected to see whether sleeping well affected their reaction times to a visual stimulus. Each adult's reaction time was measured twice.

The first measurement for reaction time was taken on a morning after the adult had slept well. The second measurement was taken on a morning after the same adult had not slept well.

The box and whisker diagrams for the reaction times, measured in seconds, are shown below.



Consider the box and whisker diagram representing the reaction times after sleeping well.

6a. State the median reaction time after sleeping well. *[1 mark]*

Markscheme

0.28 (s) ***A1***

[1 mark]

6b. Verify that the measurement of 0.46 seconds is not an outlier. *[3 marks]*

Markscheme

$\text{IQR} = 0.35 - 0.27\ (= 0.08)$ (s) ***(A1)***

substituting **their** IQR into correct expression for upper fence ***(A1)***

$0.35 + 1.5 \times 0.08\ (= 0.47)$ (s)

$0.46 < 0.47$ ***R1***

so 0.46 (s) is not an outlier ***AG***

[3 marks]

6c. State why it appears that the mean reaction time is greater than the median reaction time. *[1 mark]*

Markscheme

EITHER

the median is closer to the lower quartile (positively skewed) ***R1***

OR

The distribution is positively skewed ***R1***

OR

the range of reaction times below the median is smaller than the range of reaction times above the median ***R1***

Note: These are sample answers from a range of acceptable correct answers. Award ***R1*** for any correct statement that explains this.
Do not award ***R1*** if there is also an incorrect statement, even if another statement in the answer is correct. Accept a correctly and clearly labelled diagram.

[1 mark]

6d. Now consider the two box and whisker diagrams. *[1 mark]*

Comment on whether these box and whisker diagrams provide any evidence that might suggest that not sleeping well causes an increase in reaction time.

Markscheme

EITHER

the distribution for ‘not sleeping well’ is centred at a higher reaction time
R1

OR

The median reaction time after not sleeping well is equal to the upper quartile reaction time after sleeping well ***R1***

OR

75% of reaction times are < 0.35 seconds after sleeping well, compared with 50% after not sleeping well ***R1***

OR

the sample size of 9 is too small to draw any conclusions ***R1***

Note: These are sample answers from a range of acceptable correct answers. Accept any relevant correct statement **that relates to the median and/or quartiles shown in the box plots**. **Do not accept** a comparison of means. Do not award ***R1*** if there is also an incorrect statement, even if another statement in the answer is correct.

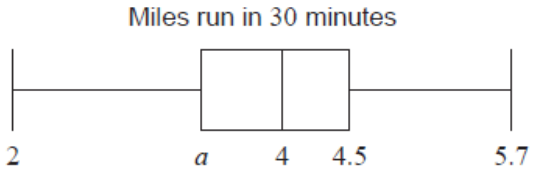
Award ***R0*** to “correlation does not imply causation”.

[1 mark]

Each athlete on a running team recorded the distance (M miles) they ran in 30 minutes.

The median distance is 4 miles and the interquartile range is 1.1 miles.

This information is shown in the following box-and-whisker plot.



7a. Find the value of a. *[2 marks]*

Markscheme

* This question is from an exam for a previous syllabus, and may contain minor differences in marking or structure.

valid approach ***(M1)***

eg $Q_3 - Q_1$, $Q_3 - 1.1$, $4.5 - a = 1.1$

$a = 3.4$ ***A1 N2***

[2 marks]

The distance in miles, M, can be converted to the distance in kilometres, K, using the formula $K = \frac{8}{5}M$.

7b. Write down the value of the median distance in kilometres (km). *[1 mark]*

Markscheme

$\frac{32}{5}$ $(= 6.4)$ (km) ***A1 N1***

[1 mark]

The variance of the distances run by the athletes is $\frac{16}{9}$ km^2.

The standard deviation of the distances is b miles.

7c. Find the value of b. *[4 marks]*

Markscheme

METHOD 1 (standard deviation first)

valid approach ***(M1)***

eg $\text{standard deviation} = \sqrt{\text{variance}}$, $\sqrt{\frac{16}{9}}$

standard deviation$= \frac{4}{3}$ (km) ***(A1)***

valid approach to convert **their** standard deviation ***(M1)***

eg $\frac{4}{3} \times \frac{5}{8}$, $\sqrt{\frac{16}{9}} = \frac{8}{5}M$

$\frac{20}{24}$ (miles) $\left(= \frac{5}{6}\right)$ ***A1 N3***

Note: If no working shown, award ***M1A1M0A0*** for the value $\frac{4}{3}$.
If working shown, and candidate's final answer is $\frac{4}{3}$, award ***M1A1M0A0***.

METHOD 2 (variance first)

valid approach to convert variance ***(M1)***

eg $\left(\frac{5}{8}\right)^2$, $\frac{64}{25}$, $\frac{16}{9} \times \left(\frac{5}{8}\right)^2$

variance $= \frac{25}{36}$ ***(A1)***

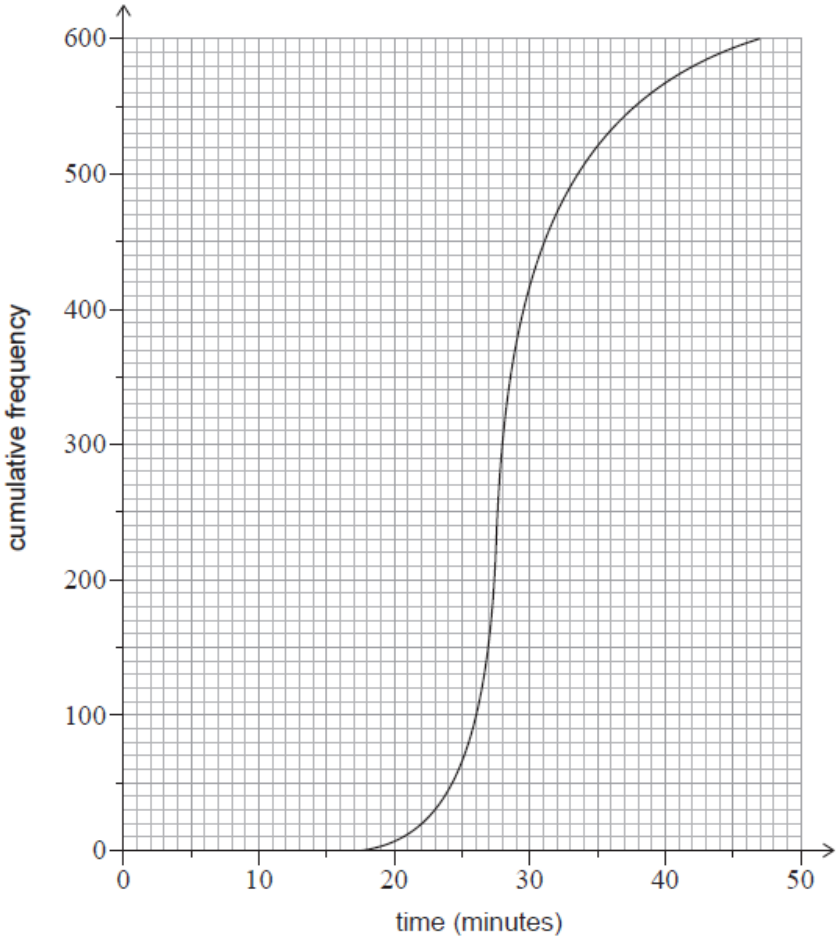
valid approach ***(M1)***

eg $\text{standard deviation} = \sqrt{\text{variance}}$, $\sqrt{\frac{25}{36}}$, $\sqrt{\frac{16}{9} \times \left(\frac{5}{8}\right)^2}$

$\frac{20}{24}$ (miles) $\left(= \frac{5}{6}\right)$ ***A1 N3***

[4 marks]

A total of 600 athletes from different teams compete in a $5\ \text{km}$ race. The times the 600 athletes took to run the $5\ \text{km}$ race are shown in the following cumulative frequency graph.



There were 400 athletes who took between 22 and m minutes to complete the $5\ \text{km}$ race.

7d. Find m. *[3 marks]*

Markscheme

correct frequency for 22 minutes ***(A1)***

eg 20

adding **their** frequency (do not accept $22 + 400$) ***(M1)***

eg $20 + 400$, 420 athletes

$m = 30$ (minutes) ***A1 N3***

[3 marks]

7e. The first 150 athletes that completed the race won a prize. *[5 marks]*

Given that an athlete took between 22 and m minutes to complete the $5\ \text{km}$ race, calculate the probability that they won a prize.

Markscheme

27 (minutes) ***(A1)***

correct working ***(A1)***

eg 130 athletes between 22 and 27 minutes, $\text{P}(22 < t < 27) = \frac{150-20}{600}$, $\frac{13}{60}$

evidence of conditional probability or reduced sample space ***(M1)***

eg $\text{P}(A \mid B)$, $\text{P}(t < 27 \mid 22 < t < 30)$, $\frac{\text{P}(22<t<27)}{\text{P}(22<t<m)}$, $\frac{150}{400}$

correct working ***(A1)***

eg $\frac{\frac{130}{600}}{\frac{400}{600}}$, $\frac{150-20}{400}$

$\frac{130}{400}$ $\left(\frac{13}{40} = \frac{78000}{240000} = \frac{390}{1200} = 0.325\right)$ ***A1*** ***N5***

Note: If no other working is shown, award ***A0A0M1A0A0*** for answer of $\frac{150}{400}$.
Award ***N0*** for answer of $\frac{3}{8}$ with no other working shown.

[5 marks]

Hafizah harvested 49 mangoes from her farm. The weights of the mangoes, w, in grams, are shown in the following grouped frequency table.

Weight (g)	$100 \leq w < 200$	$200 \leq w < 300$	$300 \leq w < 400$	$400 \leq w < 500$	$500 \leq w < 600$
Frequency	4	7	14	16	8

8a. Write down the modal group for these data. *[1 mark]*

Markscheme

* This question is from an exam for a previous syllabus, and may contain minor differences in marking or structure. It appeared in a paper that permitted the use of a calculator, and so might not be suitable for all forms of practice.

$400 \leq w < 500$ ***(A1)*** ***(C1)***

Note: Accept alternative notation $[400,\ 500)$ or $[400,\ 500[$.
Do not accept "400-500".

[1 mark]

8b. Use your graphic display calculator to find an estimate of the standard deviation of the weights of mangoes from this harvest. *[2 marks]*

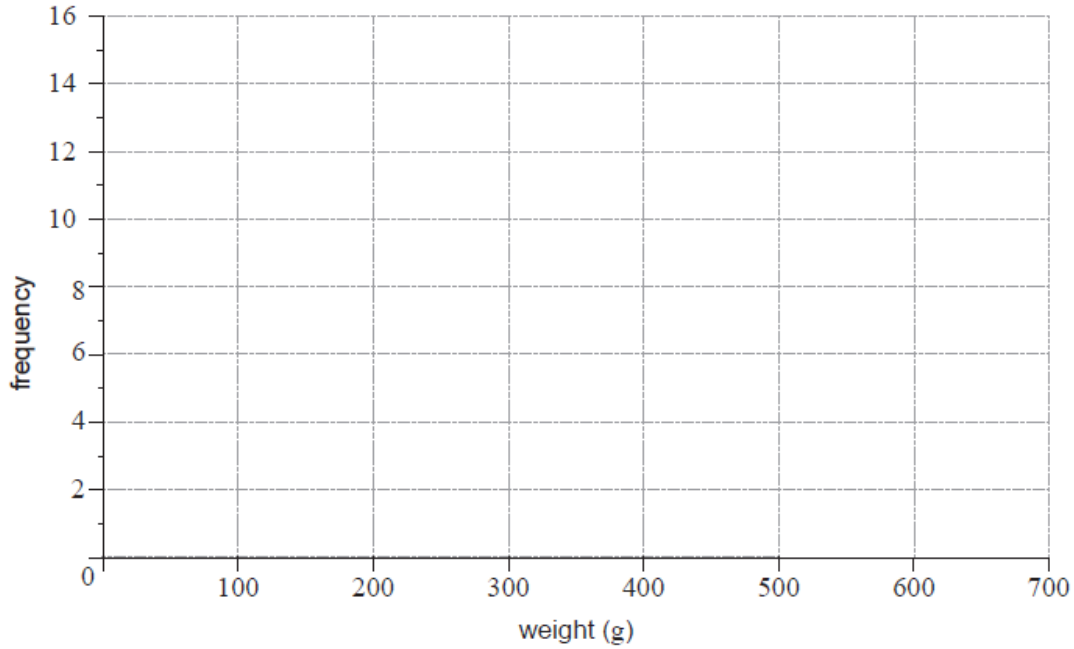
Markscheme

115 $(115.265\ldots\ (g))$ ***(A2)*** ***(C2)***

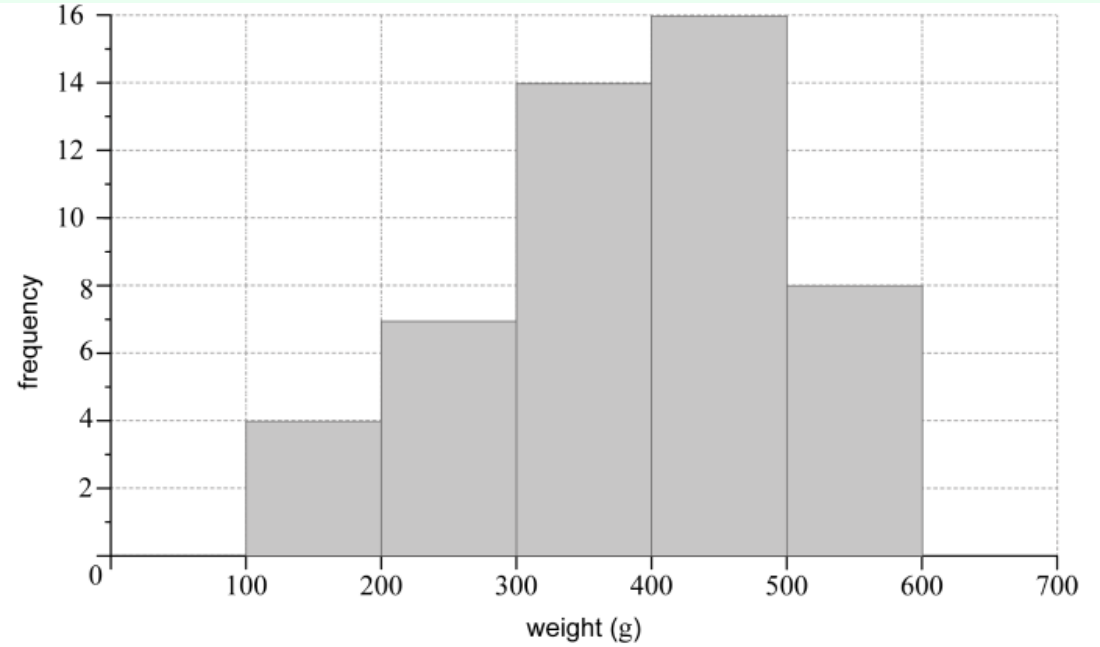
Note: Award ***(A1)(A0)*** for an answer of 116 $(116.459\ldots)$.

[2 marks]

8c. On the grid below, draw a histogram for the data in the table. *[3 marks]*



Markscheme



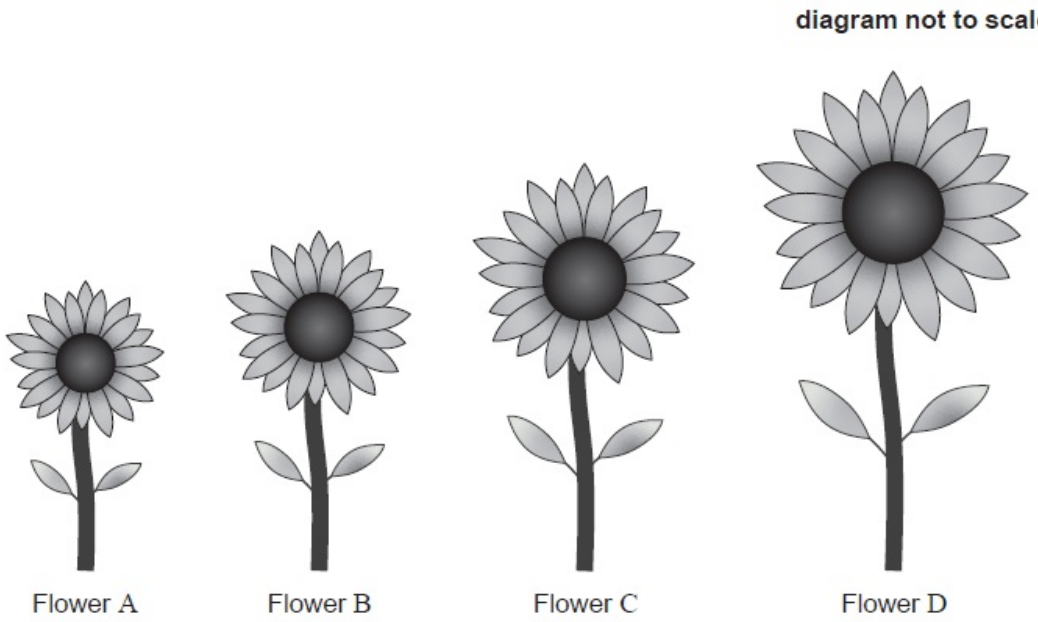
(A2)

(A1) ***(C3)***

Note: Award ***(A2)*** for all correct heights of bars or ***(A1)*** for three or four correct heights of bars.
Award ***(A1)*** for rectangular bars all with correct left and right end points ($100,\ 200,\ 300,\ 400,\ 500$ and 600) and for no gaps; the bars do **not** have to be shaded.
Award at most ***(A2)(A0)*** if a ruler is not used for all lines.

[3 marks]

Anne-Marie planted four sunflowers in order of height, from shortest to tallest.



Flower C is $32\ \mathrm{cm}$ tall.

The median height of the flowers is $24\ \mathrm{cm}$.

9a. Find the height of Flower null. *[2 marks]*

Markscheme

* This question is from an exam for a previous syllabus, and may contain minor differences in marking or structure. It appeared in a paper that permitted the use of a calculator, and so might not be suitable for all forms of practice.

$24-8$ **OR** $24-(32-24)$ **OR** $24=\frac{32+h}{2}$ ***(M1)***

Note: Award ***(M1)*** for subtracting 8 from the median, or equivalent.

$16\ (\mathrm{cm})$ ***(A1)*** ***(C2)***

[2 marks]

The range of the heights is $50\ \mathrm{cm}$. The height of Flower A is $p\ \mathrm{cm}$ and the height of Flower D is $q\ \mathrm{cm}$.

9b. Using this information, write down an equation in p and q. *[1 mark]*

Markscheme

$q - p = 50$ (or equivalent) ***(A1)*** ***(C1)***

[1 mark]

The mean height of the flowers is $27\ \text{cm}$.

9c. Write down a second equation in p and q. *[1 mark]*

Markscheme

$\frac{p+16+32+q}{4} = 27$ **OR** $p + q = 60$ (or equvalent) ***(A1)*****(ft)** ***(C1)***

Note: Follow through from part (a).

[1 mark]

9d. Using your answers to **parts (b)** and **(c)**, find the height of Flower A. *[1 mark]*

Markscheme

$5\,(\text{cm})$ ***(A1)*****(ft)** ***(C1)***

Note: Follow through from parts (b) and (c).

[1 mark]

9e. Using your answers to **parts (b)** and **(c)**, find the height of Flower D. *[1 mark]*

Markscheme

$55\,(\text{cm})$ ***(A1)*****(ft)** ***(C1)***

Note: Follow through from parts (b) and (c).

[1 mark]

Lucy sells hot chocolate drinks at her snack bar and has noticed that she sells more hot chocolates on cooler days. On six different days, she records the maximum daily temperature, T, measured in degrees centigrade, and the number of hot chocolates sold, H. The results are shown in the following table.

Maximum temperature (T)	14	8	4	18	13	11
Number of hot chocolates (H)	79	143	191	58	84	105

The relationship between H and T can be modelled by the regression line with equation $H = aT + b$.

10a. Find the value of a and of b. *[3 marks]*

Markscheme

valid approach ***(M1)***

eg correct value for a or b (or for r or $r^2 = 0.962839$ seen in (ii))

$a = -9.84636,\ b = 221.592$

$a = -9.85,\ b = 222$ ***A1A1 N3***

[3 marks]

10b. Write down the correlation coefficient. *[1 mark]*

Markscheme

-0.981244

$r = -0.981$ ***A1 N1***

[1 mark]

10c. Using the regression equation, estimate the number of hot chocolates that Lucy will sell on a day when the maximum temperature is $12\,^{\circ}\mathrm{C}$. *[2 marks]*

Markscheme

correct substitution into their equation ***(A1)***

eg $-9.85 \times 12 + 222$

103.435 (103.8 from 3 sf)

103 (hot chocolates) ***A1 N2***

[2 marks]

The number of messages, M, that six randomly selected teenagers sent during the month of October is shown in the following table. The table also shows the time, T, that they spent talking on their phone during the same month.

Time spent talking on their phone (T minutes)	50	55	105	128	155	200
Number of messages (M)	358	340	740	731	800	992

The relationship between the variables can be modelled by the regression equation $M = aT + b$.

11a. Write down the value of a and of b. *[3 marks]*

Markscheme

evidence of set up ***(M1)***

eg correct value for a or b (accept $r = 0.966856$)

$4.30161, 163.330$

$a = 4.30$, $b = 163$ (accept $y = 4.30x + 163$) ***A1A1 N3***

[3 marks]

11b. Use your regression equation to predict the number of messages sent by a teenager that spent 154 minutes talking on their phone in October. *[3 marks]*

Markscheme

valid approach ***(M1)***

eg $4.30\,(154) + 163$

eg 825.778 (825.2 from 3 sf values) ***(A1)***

number of messages $= 826$ (must be an integer) ***A1 N3***

[3 marks]