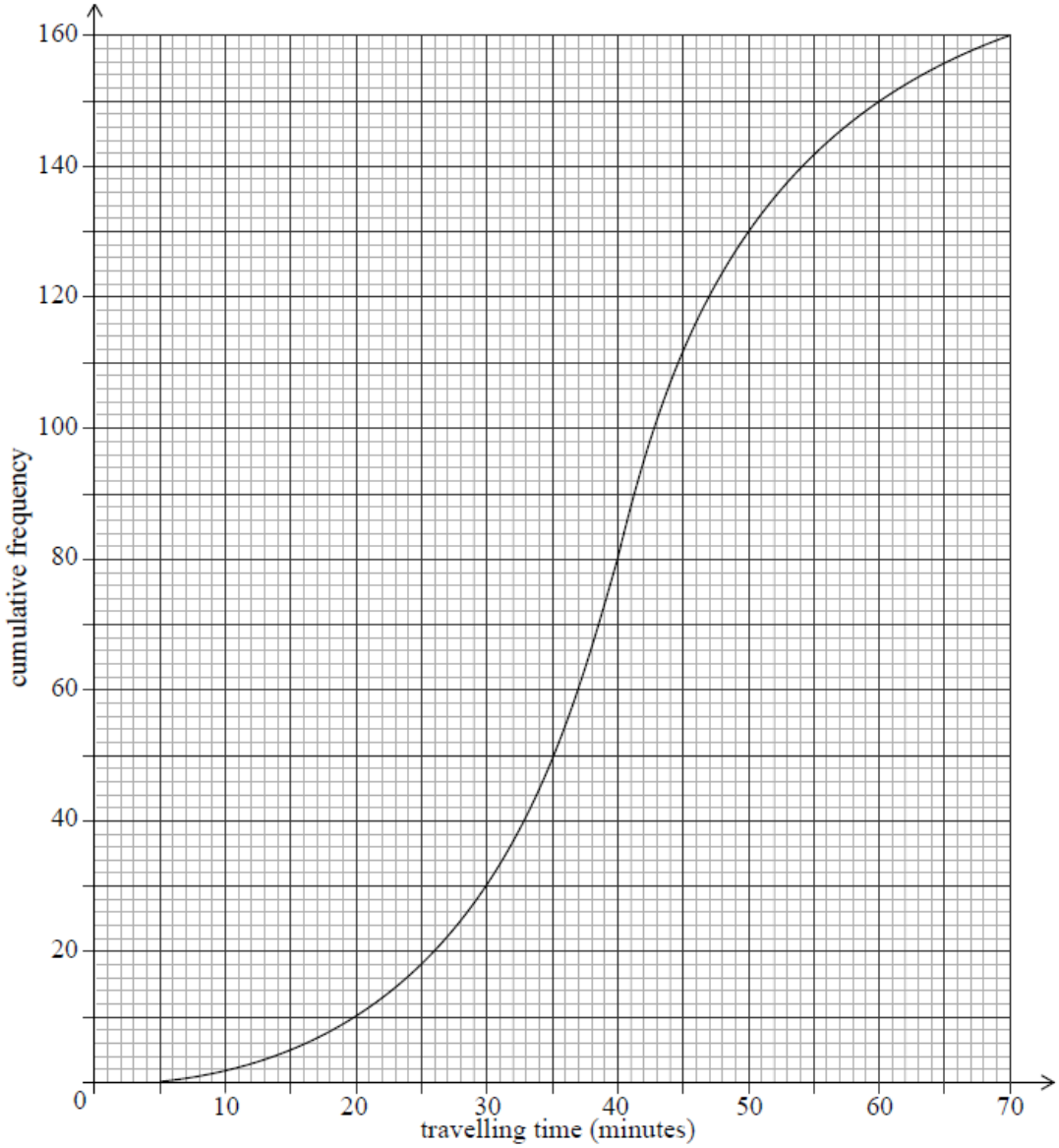
Statistics - Revision [62 marks]

1. [Maximum mark: 15] SPM.1.SL.TZ0.7

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.



(a) Find the median number of minutes spent traveling to work. [2]

Markscheme
evidence of median position ***(M1)***
80th employee

40 minutes **A1**

[2 marks]

(b) Find the number of employees whose travelling time is within 15 minutes of the median. [3]

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.

(c) Find the value of k. [3]

Markscheme

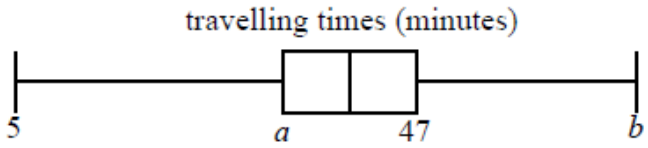
recognising that there are 16 employees in the top 10% ***(M1)***

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.



(d) Write down the value of b. [1]

Markscheme

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

[1 mark]

(e.i) Find the value of a. [2]

Markscheme

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

40 employees

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

[2 marks]

(e.ii) Hence, find the interquartile range. [2]

Markscheme

47 – 33 ***(M1)***

IQR = 14 ***A1***

[2 marks]

(f) Travelling times of less than p minutes are considered outliers.

Find the value of p. [2]

Markscheme

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

33 – 21

12 ***(A1)***

[2 marks]

2. [Maximum mark: 8] EXN.2.SL.TZ0.4

The following table shows the systolic blood pressures, p mmHg, and the ages, t years, of 6 male patients at a medical clinic.

Patient	P1	P2	P3	P4	P5	P6
t (years)	40	72	35	47	21	61
p (mmHg)	105	145	100	130	95	132

(a.i) Determine the value of Pearson's product-moment correlation coefficient, r, for these data. [2]

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.

$r = 0.946$ **A2**

[2 marks]

(a.ii) Interpret, in context, the value of r found in part (a) (i). [1]

Markscheme

the value of r shows a (very) strong positive correlation between age and (systolic) blood pressure **A1**

[1 mark]

The relationship between t and p can be modelled by the regression line of p on t with equation $p = at + b$.

(b) Find the equation of the regression line of p on t. [2]

Markscheme

$p = 1.05t + 69.3$ **A1A1**

Note: Only award marks for an equation. Award **A1** for $a = 1.05$ and **A1** for $b = 69.3$. Award **A1A0** for $y = 1.05x + 69.3$.

[2 marks]

A 50-year-old male patient enters the medical clinic for his appointment.

(c) Use the regression equation from part (b) to predict this patient's systolic blood pressure. [2]

Markscheme

122 (mmHg) **(M1)A1**

[2 marks]

(d) A 16-year-old male patient enters the medical clinic for his appointment.

Explain why the regression equation from part (b) should not be used to predict this patient's systolic blood pressure. [1]

Markscheme

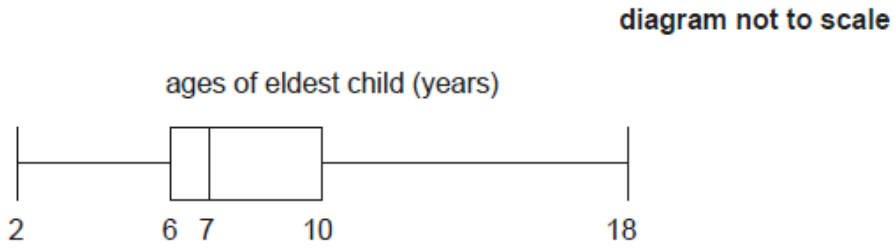
the regression equation should not be used because it involves extrapolation **A1**

[1 mark]

3. [Maximum mark: 7] 22M.1.SL.TZ1.3

A survey at a swimming pool is given to one adult in each family. The age of the adult, a years old, and of their eldest child, c years old, are recorded.

The ages of the eldest child are summarized in the following box and whisker diagram.



(a) Find the largest value of c that would not be considered an outlier. [3]

Markscheme

$\text{IQR} = 10 - 6 (= 4)$ *(A1)*

attempt to find $Q_3 + 1.5 \times \text{IQR}$ *(M1)*

$10 + 6$

16 *A1*

[3 marks]

The regression line of a on c is $a = \frac{7}{4}c + 20$. The regression line of c on a is $c = \frac{1}{2}a - 9$.

(b.i) One of the adults surveyed is 42 years old. Estimate the age of their eldest child. [2]

Markscheme

choosing $c = \frac{1}{2}a - 9$ *(M1)*

$\frac{1}{2} \times 42 - 9$

$= 12$ (years old) *A1*

[2 marks]

(b.ii) Find the mean age of all the adults surveyed. [2]

Markscheme

attempt to solve system by substitution or elimination ***(M1)***

34 (years old) ***A1***

[2 marks]

4. [Maximum mark: 4] 22M.2.SL.TZ1.2

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.

(a) Find the value of x. [2]

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]

(b) Find the standard deviation. [2]

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**

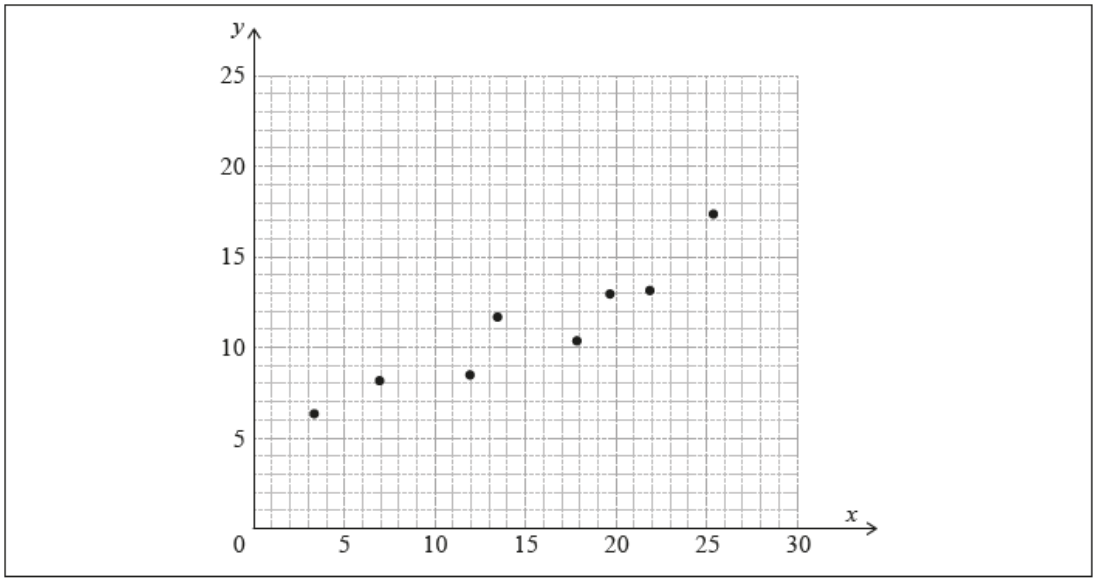
[2 marks]

5. [Maximum mark: 7] 21M.2.SL.TZ1.2

The following table shows the data collected from an experiment.

x	3.3	6.9	11.9	13.4	17.8	19.6	21.8	25.3
y	6.3	8.1	8.4	11.6	10.3	12.9	13.1	17.3

The data is also represented on the following scatter diagram.



The relationship between x and y can be modelled by the regression line of y on x with equation $y = ax + b$, where $a,\ b \in \mathbb{R}$.

(a) Write down the value of a and the value of b. [2]

Markscheme

$a = 0.433156\ldots,\ b = 4.50265\ldots$

$a = 0.433,\ b = 4.50$ **A1A1**

[2 marks]

(b) Use this model to predict the value of y when $x = 18$. [2]

Markscheme

attempt to substitute $x = 18$ into their equation *(M1)*

$y = 0.433 \times 18 + 4.50$

$= 12.2994\ldots$

$= 12.3$ *A1*

[2 marks]

(c) Write down the value of $\overline{x}$ and the value of $\overline{y}$. [1]

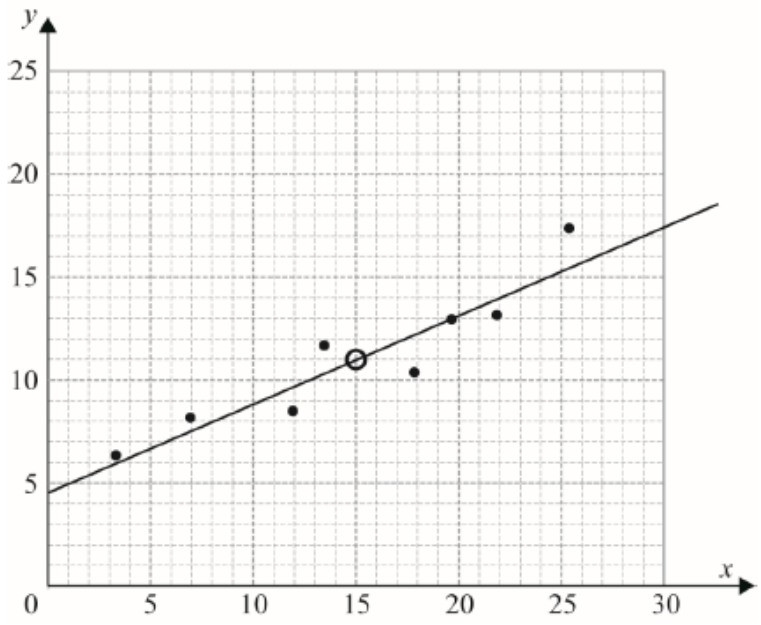
Markscheme

$\overline{x} = 15,\ \overline{y} = 11$ *A1*

[1 mark]

(d) Draw the line of best fit on the scatter diagram. [2]

Markscheme



A1A1

Note: Award marks as follows:

A1 for a straight line going through $(15, 11)$

A1 for intercepting the y-axis between their $b \pm 1.5$ (when their line is extended), which includes all the data for $3.3 \leq x \leq 25.3$.

If the candidate does not use a ruler, award ***A0A1*** where appropriate.

[2 marks]

6. [Maximum mark: 6] 20N.1.SL.TZ0.T_3

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

(a) Write down the modal group for these data. [1]

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]

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

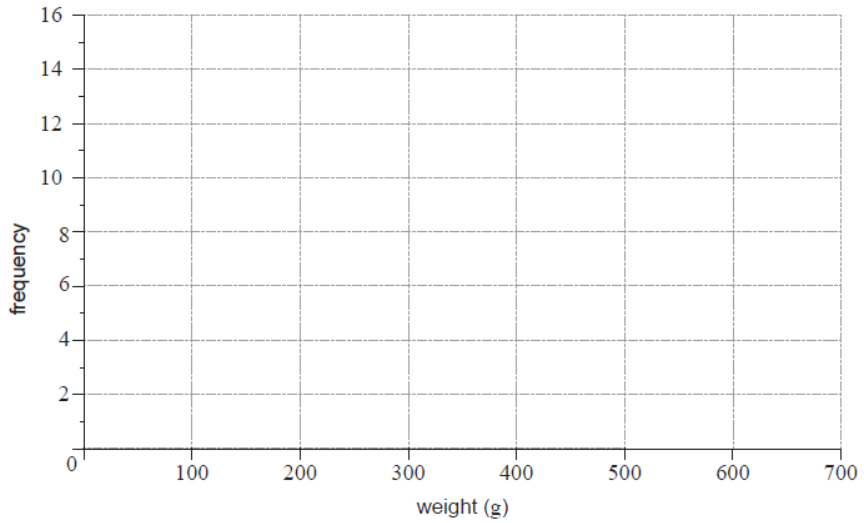
Markscheme

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

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

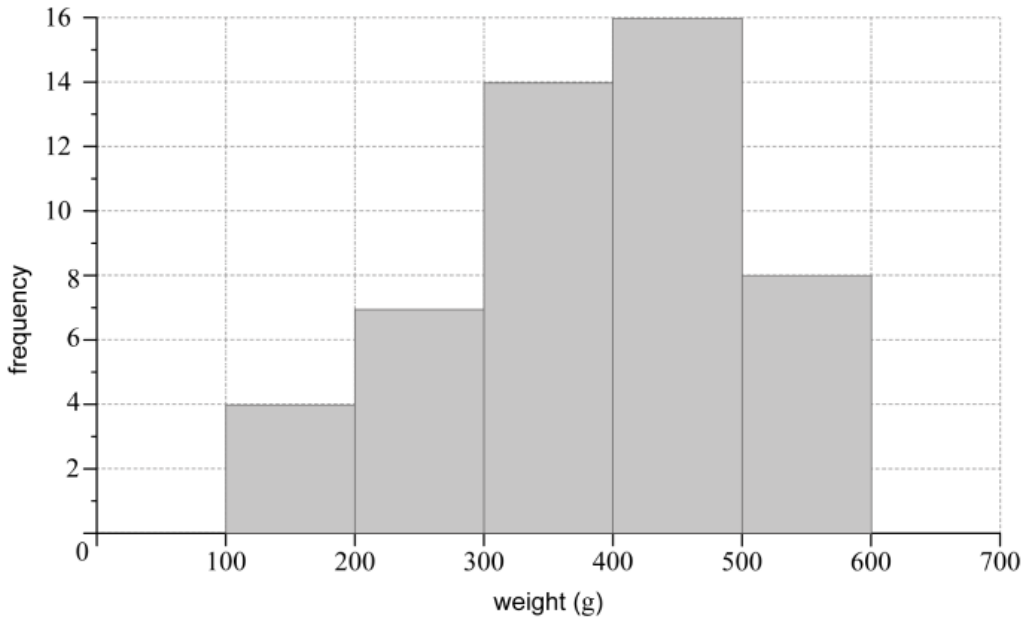
[2 marks]

(c) On the grid below, draw a histogram for the data in the table.



[3]

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.

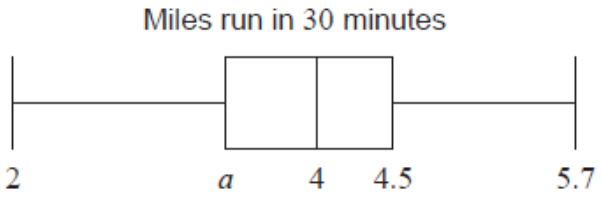
[3 marks]

7. [Maximum mark: 15] 20N.1.SL.TZ0.S_8

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.



(a) Find the value of a. [2]

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 $K = \frac{8}{5}M$.

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

Markscheme

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

[1 mark]

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

The standard deviation of the distances is b miles.

(c) Find the value of b. [4]

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)***

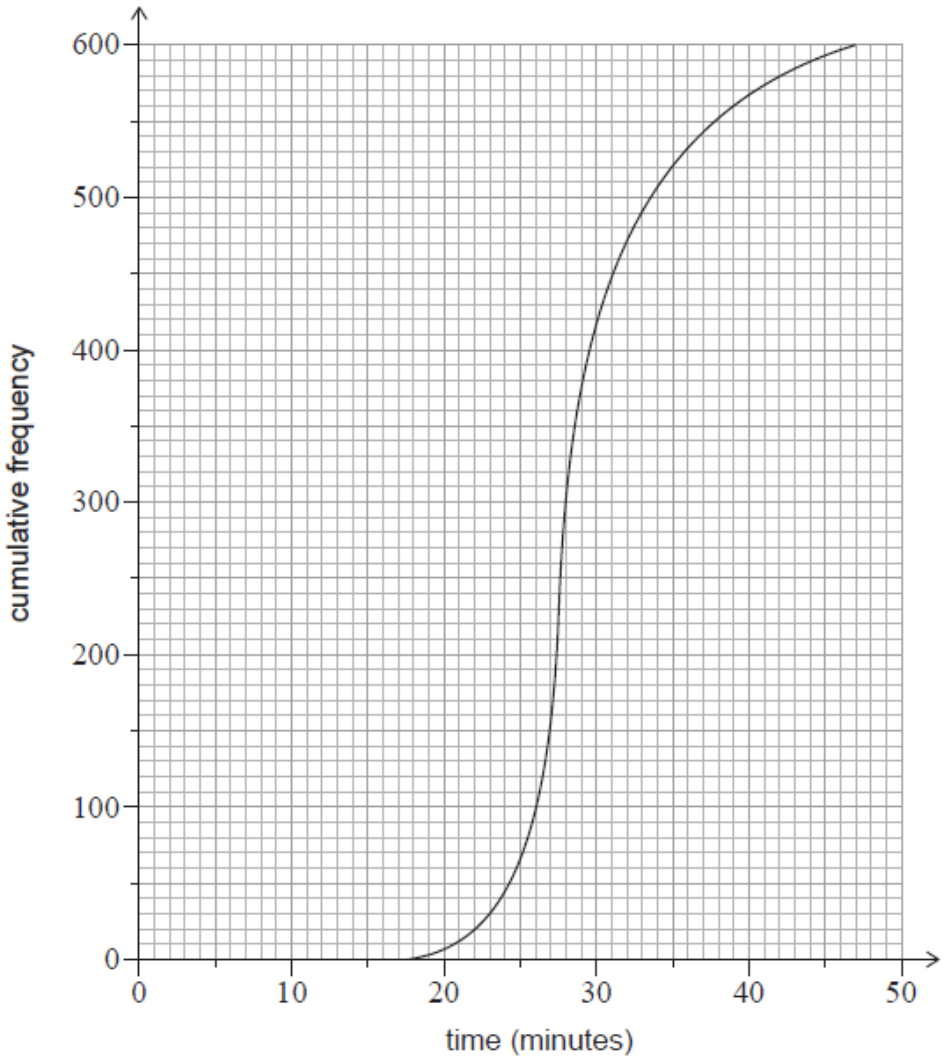
valid approach ***(M1)***

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.

(d) Find m. [3]

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]

(e) The first 150 athletes that completed the race won a prize.

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. [5]

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

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]