SL Paper 2

The manager of a folder factory recorded the number of folders produced by the factory (in thousands) and the production costs (in thousand Euros), for six consecutive months.

	January	February	March	April	May	June
Folders produced, x (thousands)	10	15	22	30	28	21
Production costs, C (thousand Euros)	35	40	60	72	68	55

Every month the factory sells all the folders produced. Each folder is sold for 2.99 Euros.

a. Draw a scatter diagram for this data. Use a scale of 2 cm for 5000 folders on the horizontal axis and 2 cm for 10 000 Euros on the vertical axis. [4]

b.i. Write down, for this set of data the mean number of folders produced, $\bar{x}$; [1]

b.ii. Write down, for this set of data the mean production cost, $\bar{C}$. [1]

c. Label the point $\mathrm{M}(\bar{x},\ \bar{C})$ on the scatter diagram. [1]

d. Use your graphic display calculator to find the Pearson's product–moment correlation coefficient, r. [2]

e. State a reason why the regression line C on x is appropriate to model the relationship between these variables. [1]

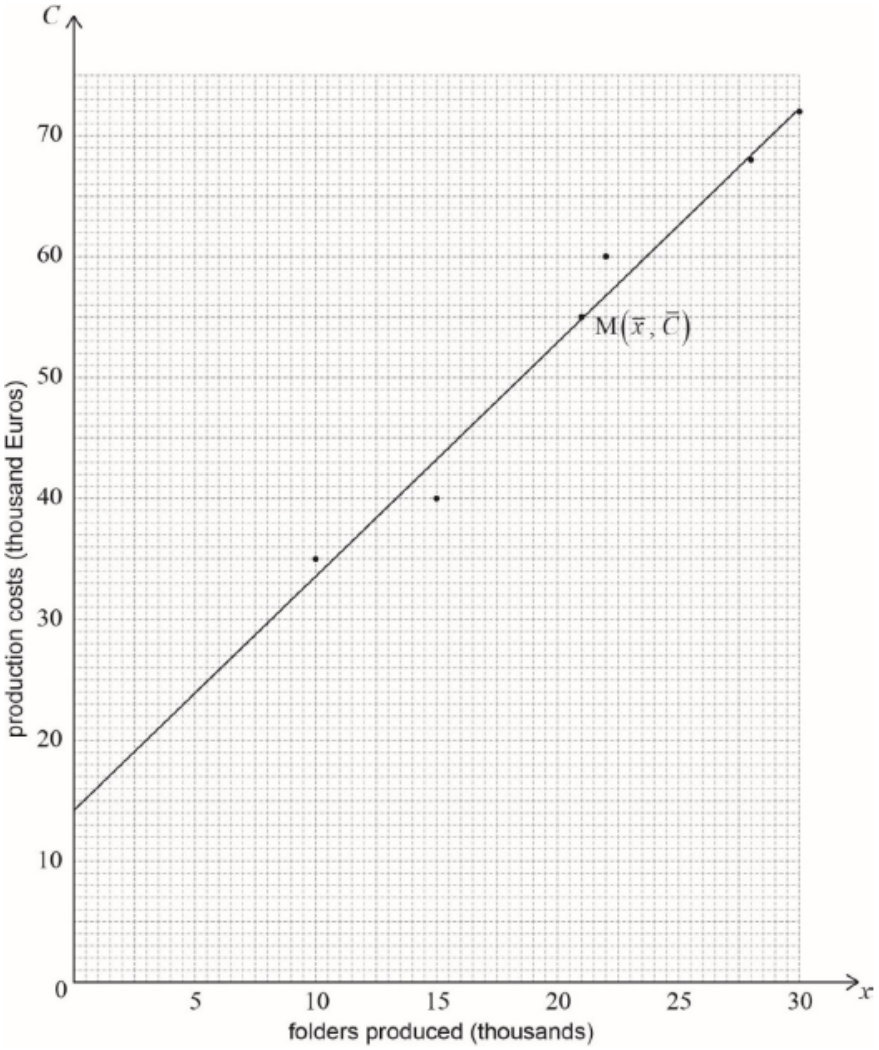
f. Use your graphic display calculator to find the equation of the regression line C on x. [2]

g. Draw the regression line C on x on the scatter diagram. [2]

h. Use the equation of the regression line to estimate the least number of folders that the factory needs to sell in a month to exceed its production cost for that month. [4]

Markscheme

a.



(A4)

Notes: Award ***(A1)*** for correct scales and labels. Award ***(A0)*** if axes are reversed and follow through for their points.

Award ***(A3)*** for all six points correctly plotted, ***(A2)*** for four or five points correctly plotted, ***(A1)*** for two or three points correctly plotted.

If graph paper has not been used, award at most ***(A1)(A0)(A0)(A0)***. If accuracy cannot be determined award ***(A0)(A0)(A0)(A0)***.

[4 marks]

b.i. $(\bar{x} =)\ 21$ ***(A1)(G1)***

[1 mark]

b.ii. $(\bar{C} =)\ 55$ ***(A1)(G1)***

Note: Accept (i) 21000 and (ii) 55000 seen.

[1 mark]

c. their mean point M labelled on diagram ***(A1)*****(ft)*****(G1)***

Note: Follow through from part (b).

Award ***(A1)*****(ft)** if their part (b) is correct and their attempt at plotting $(21,\ 55)$ in part (a) is labelled M.

If graph paper not used, award ***(A1)*** if $(21,\ 55)$ is labelled. If their answer from part (b) is incorrect and accuracy cannot be determined, award ***(A0)***.

[1 mark]

d. $(r =)\ 0.990\ (0.989568\ldots)$ ***(G2)***

Note: Award ***(G2)*** for 0.99 seen. Award ***(G1)*** for 0.98 or 0.989. Do not accept 1.00.

[2 marks]

e. the correlation coefficient/*r* is (very) close to 1 ***(R1)*****(ft)**

OR

the correlation is (very) strong ***(R1)*****(ft)**

Note: Follow through from their answer to part (d).

OR

the position of the data points on the scatter graphs suggests that the tendency is linear ***(R1)*****(ft)**

Note: Follow through from their scatter graph in part (a).

[1 mark]

f. $C = 1.94x + 14.2\ (C = 1.94097\ldots x + 14.2395\ldots)$ ***(G2)***

Notes: Award ***(G1)*** for $1.94x$, ***(G1)*** for 14.2.

Award a maximum of ***(G0)(G1)*** if the answer is not an equation.

Award ***(G0)(G1)*****(ft)** if gradient and C-intercept are swapped in the equation.

[2 marks]

g. straight line through their $\mathrm{M}(21,\ 55)$ ***(A1)*****(ft)**

C-intercept of the line (or extension of line) passing through $14.2\ (\pm 1)$ ***(A1)*****(ft)**

Notes: Follow through from part (f). In the event that the regression line is not straight (ruler not used), award ***(A0)(A1)*****(ft)** if line passes through both their $(21,\ 55)$ and $(0,\ 14.2)$, otherwise award ***(A0)(A0)***. The line must pass *through* the midpoint, not *near* this point. If it is not clear award ***(A0)***.

If graph paper is not used, award at most (A1)(ft)(A0).

[2 marks]

h. $2.99x = 1.94097\ldots x + 14.2395\ldots$ ***(M1)(M1)***

Note: Award ***(M1)*** for $2.99x$ seen and ***(M1)*** for equating to their equation of the regression line. Accept an inequality sign.

Accept a correct graphical method involving their part (f) and $2.99x$.

Accept $C = 2.99x$ drawn on their scatter graph.

$x = 13.5739\ldots$ (this step may be implied by their final answer) ***(A1)*(ft)(G2)**

$13\,600\ (13\,574)$ ***(A1)*(ft)*(G3)***

Note: Follow through from their answer to (f). Use of 3 sf gives an answer of $13\,524$.

Award ***(G2)*** for $13.5739\ldots$ or 13.524 or a value which rounds to 13500 seen without workings.

Award the last ***(A1)*(ft)** for correct multiplication by 1000 **and** an answer satisfying revenue > **their** production cost.

Accept 13.6 thousand (folders).

[4 marks]

Examiners report

a. [N/A]
b.i. [N/A]
b.ii. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]
h. [N/A]

A manufacturer produces 1500 boxes of breakfast cereal every day.

The weights of these boxes are normally distributed with a mean of 502 grams and a standard deviation of 2 grams.

All boxes of cereal with a weight between 497.5 grams and 505 grams are sold. The manufacturer's income from the sale of each box of cereal is $2.00.

The manufacturer recycles any box of cereal with a weight **not** between 497.5 grams and 505 grams. The manufacturer's recycling cost is $0.16 per box.

A **different** manufacturer produces boxes of cereal with weights that are normally distributed with a mean of 350 grams and a standard deviation of 1.8 grams.

This manufacturer sells all boxes of cereal that are above a minimum weight, w.

They sell 97% of the cereal boxes produced.

a. Draw a diagram that shows this information. [2]

b. (i) Find the probability that a box of cereal, chosen at random, is sold. [4]

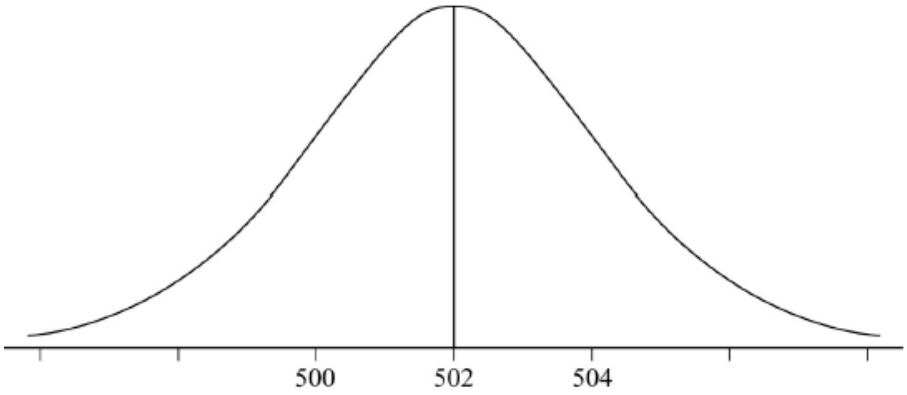
(ii) Calculate the manufacturer's expected daily income from these sales.

c. Calculate the manufacturer's expected daily recycling cost. [2]

d. Calculate the value of w. [3]

Markscheme

a.



(A1)(A1)

Notes: Award ***(A1)*** for bell shape with mean of 502.

Award ***(A1)*** for an indication of standard deviation *eg* 500 and 504.

[2 marks]

b. (i) $0.921\ (0.920968\ldots,\ 92.0968\ldots\%)$ ***(G2)***

Note: Award ***(M1)*** for a diagram showing the correct shaded region.

(ii) $1500 \times 2 \times 0.920968\ldots$ ***(M1)***

$= (\$)\ 2760\ (2762.90\ldots)$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from their answer to part (b)(i).

[4 marks]

c. $1500 \times 0.16 \times 0.079031\ldots$ ***(M1)***

Notes: Award ***(A1)*** for $1500 \times 0.16 \times$ their $(1 - 0.920968\ldots)$.

OR

$(1500 - 1381.45) \times 0.16$ ***(M1)***

Notes: Award ***(M1)*** for $(1500 -$ their $1381.45) \times 0.16$.

$= (\$)19.0\ (18.9676\ldots)$ ***(A1)*****(ft)*****(G2)***

[2 marks]

d. $347\ (\text{grams})\ (346.614\ldots)$ ***(G3)***

Notes: Award ***(G2)*** for an answer that rounds to 346.

Award ***(G1)*** for $353.385\ldots$ seen without working (for finding the top 3%).

[3 marks]

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]

As part of his IB Biology field work, Barry was asked to measure the circumference of trees, in centimetres, that were growing at different distances, in metres, from a river bank. His results are summarized in the following table.

Distance, x (metres)	5	12	17	21	24	30	34	44	47
Circumference, y (centimetres)	82	76	70	68	67	60	62	50	50

a. State whether *distance from the river bank* is a continuous **or** discrete variable. [1]

b. **On graph paper,** draw a scatter diagram to show Barry's results. Use a scale of 1 cm to represent 5 m on the *x*-axis and 1 cm to represent 10 cm on the *y*-axis. [4]

c. Write down [2]

(i) the mean distance, $\bar{x}$, of the trees from the river bank;

(ii) the mean circumference, $\bar{y}$, of the trees.

d. Plot and label the point $\mathrm{M}(\bar{x},\ \bar{y})$ on your graph. [2]

e. Write down [4]

(i) the Pearson's product–moment correlation coefficient, r, for Barry's results;

(ii) the equation of the regression line y on x, for Barry's results.

f. Draw the regression line y on x on your graph. [2]

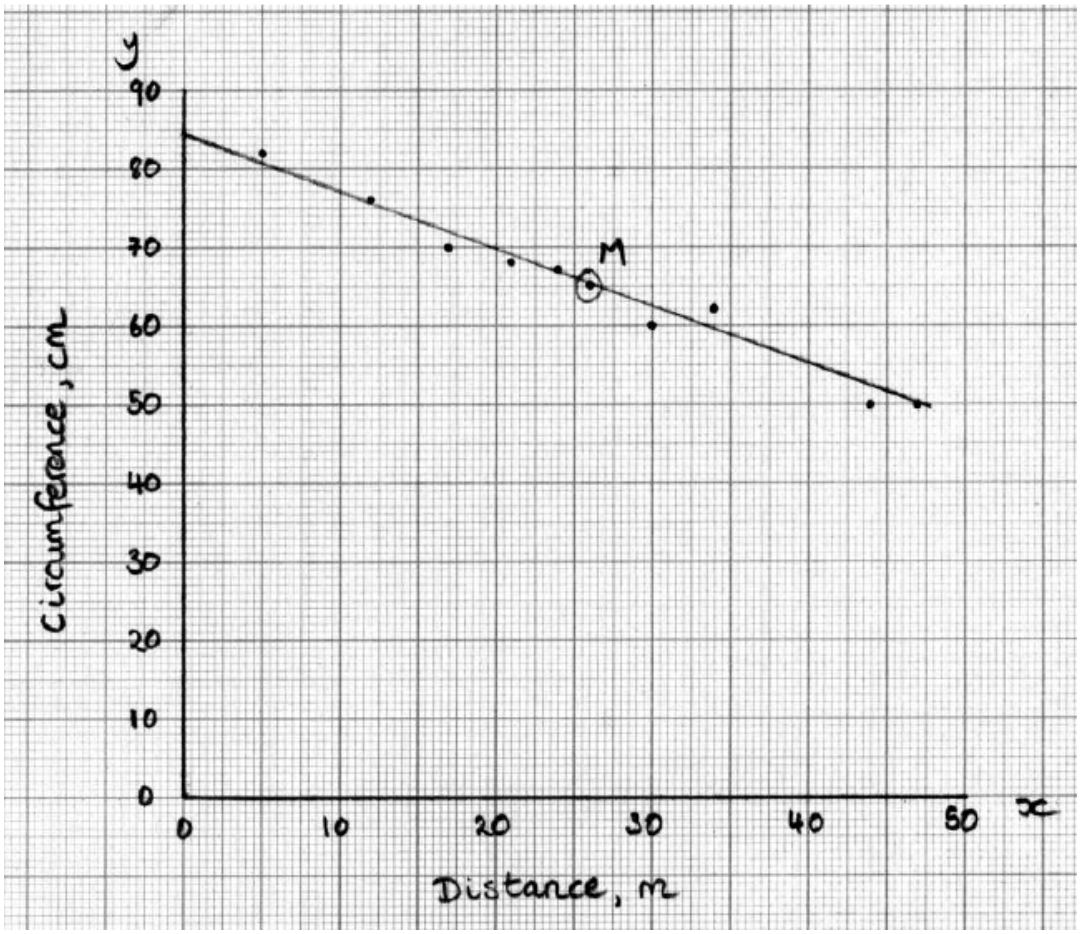
g. **Use the equation of the regression line** y on x to estimate the circumference of a tree that is 40 m from the river bank. [2]

Markscheme

a. continuous ***(A1)***

[1 mark]

b.



(A1)(A1)(A1)(A1)

Notes: Award ***(A1)*** for labelled axes and correct scales; if axes are reversed award ***(A0)*** and follow through for their points. Award ***(A1)*** for at least 3 correct points, ***(A2)*** for at least 6 correct points, ***(A3)*** for all 9 correct points. If scales are too small or graph paper has not been used, accuracy cannot be determined; award ***(A0)***. Do not penalize if extra points are seen.

[4 marks]

c. (i) 26 (m) ***(A1)***

(ii) 65 (cm) ***(A1)***

[2 marks]

d. point M labelled, in correct position ***(A1)(A1)*****(ft)**

Notes: Award ***(A1)*****(ft)** for point plotted in correct position, ***(A1)*** for point labelled M or $(\bar{x},\ \bar{y})$. Follow through from their answers to part (c).

[2 marks]

e. (i) $-0.988\ \ (-0.988432\ldots)$ ***(G2)***

Note: Award ***(G2)*** for -0.99. Award ***(G1)*** for -0.990.

Award ***(A1)(A0)*** if minus sign is omitted.

(ii) $y = -0.756x + 84.7 \quad (y = -0.756281\ldots x + 84.6633\ldots)$ ***(G2)***

Notes: Award ***(A1)*** for $-0.756x$, ***(A1)*** for 84.7. If the answer is not given as an equation, award a maximum of ***(A1)(A0)***.

[4 marks]

f. regression **line** through their M ***(A1)*****(ft)**

regression **line** through their $(0, 85)$ (accept 85 ± 1) ***(A1)*****(ft)**

Notes: Follow through from part (d). Award a maximum of ***(A1)(A0)*** if the line is not straight. Do not penalize if either the line does not meet the *y*-axis or extends into quadrants other than the first.

If M is not plotted or labelled, then follow through from part (c).

Follow through from their *y*-intercept in part (e)(ii).

[2 marks]

g. $-0.756281(40) + 84.6633$ ***(M1)***

$= 54.4\ (\mathrm{cm})\ (54.4120\ldots)$ ***(A1)*****(ft)*****(G2)***

Notes: Accept 54.5 (54.46) for use of 3 sf. Accept 54.3 from use of -0.76 and 84.7.

Follow through from their equation in part (e)(ii) **irrespective of working shown**; the final answer seen must be consistent with that equation for the final ***(A1)*** to be awarded.

Do not accept answers taken from the graph.

[2 marks]

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]

The following table shows the number of bicycles, x, produced daily by a factory and their total production cost, y, in US dollars (USD). The table shows data recorded over seven days.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Number of bicycles, x	12	15	14	17	20	18	21
Production cost, y	3900	4600	4100	5300	6000	5400	6000

a. (i) Write down the Pearson's product–moment correlation coefficient, r, for these data. [4]

(ii) Hence comment on the result.

b. Write down the equation of the regression line y on x for these data, in the form $y = ax + b$. [2]

c. Estimate the total cost, **to the nearest USD**, of producing 13 bicycles on a particular day. [3]

d. All the bicycles that are produced are sold. The bicycles are sold for 304 USD **each**. [2]

Explain why the factory does **not** make a profit when producing 13 bicycles on a particular day.

e. All the bicycles that are produced are sold. The bicycles are sold for 304 USD **each**. [5]

(i) Write down an expression for the total selling price of x bicycles.

(ii) Write down an expression for the **profit** the factory makes when producing x bicycles on a particular day.

(iii) Find the least number of bicycles that the factory should produce, on a particular day, in order to make a profit.

Markscheme

a. (i) $r = 0.985 \quad (0.984905\ldots)$ ***(G2)***

Notes: If unrounded answer is not seen, award ***(G1)(G0)*** for 0.99 or 0.984. Award ***(G2)*** for 0.98.

(ii) strong, positive ***(A1)(A1)***

b. $y = 259.909\ldots x + 698.648\ldots \quad (y = 260x + 699)$ ***(G1)(G1)***

Notes: Award ***(G1)*** for $260x$ and ***(G1)*** for 699. If the answer is not an equation award a maximum of ***(G1)(G0)***.

c. $y = 259.909\ldots \times 13 + 698.648\ldots$ ***(M1)***

Note: Award ***(M1)*** for substitution of 13 into their regression line equation from part (b).

$y = 4077.47\ldots$ ***(A1)*****(ft)*****(G2)***

$y = 4077\ (\text{USD})$ ***(A1)*****(ft)**

Notes: Follow through from their answer to part (b). If rounded values from part (b) used, answer is 4079. Award the final ***(A1)*****(ft)** for a correct rounding to the nearest USD of their answer. The unrounded answer may not be seen.

If answer is 4077 and no working is seen, award ***(G2)***.

d. $13 \times 304 - (4077.47) = -125.477\ldots \quad (-125)$ **OR**

$4077.47 - (13 \times 304) = 125.477\ldots \quad (125)$ ***(M1)***

Notes: Award ***(M1)*** for calculating the difference between 13×304 and their answer to part (c).

If rounded values are used in equation, answer is -127.

profit is negative **OR** $\text{cost} > \text{sales}$ ***(A1)***

OR

$13 \times 304 = 3952$ ***(M1)***

Note: Award ***(M1)*** for calculating the price of 13 bikes.

$3952 < 4077.47$ ***(A1)*****(ft)**

Note: Award ***(A1)*** for showing 3952 is less than their part (c). This may be communicated in words. Follow through from part (c), but only if value is greater than 3952.

OR

$\frac{4077}{13} = 313.62$ ***(M1)***

Note: Award ***(M1)*** for calculating the cost of 1 bicycle.

$313.62 > 304$ ***(A1)*****(ft)**

Note: Award ***(A1)*** for showing 313.62 is greater than 304. This may be communicated in words. Follow through from part (c), but only if value is greater than 304.

OR

$\frac{4077}{304} = 13.41$ ***(M1)***

Note: Award ***(M1)*** for calculating the number of bicycles that should have been be sold to cover total cost.

$13.41 > 13$ ***(A1)*****(ft)**

Note: Award ***(A1)*** for showing 13.41 is greater than 13. This may be communicated in words. Follow through from part (c), but only if value is greater than 13.

e. (i) $304x$ ***(A1)***

(ii) $304x - (259.909\ldots x + 698.648\ldots)$ ***(A1)*****(ft)*****(A1)*****(ft)**

Note: Award ***(A1)*****(ft)** for difference between their answers to parts (b) and (e)(i), ***(A1)*****(ft)** for correct expression.

(iii) $304x - (259.909\ldots x + 698.648\ldots) > 0$ ***(M1)***

Notes: Award ***(M1)*** for comparing their expression in part (e)(ii) to 0. Accept an equation. Accept $3040x - y > 0$ or equivalent.

$x = 16$ bicycles ***(A1)*****(ft)*****(G2)***

Notes: Follow through from their answer to part (b). Answer must be a positive integer greater than 13 for the ***(A1)*****(ft)** to be awarded.

Award ***(G1)*** for an answer of 15.84.

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]

d. [N/A]
e. [N/A]

The following table shows the average body weight, x, and the average weight of the brain, y, of seven species of mammal. Both measured in kilograms (kg).

Species	Average body weight, x (kg)	Average weight of the brain, y (kg)
Cat	3	0.026
Cow	465	0.423
Donkey	187	0.419
Giraffe	529	0.680
Goat	28	0.115
Jaguar	100	0.157
Sheep	56	0.175

The average body weight of grey wolves is 36 kg.

In fact, the average weight of the brain of grey wolves is 0.120 kg.

The average body weight of mice is 0.023 kg.

a. Find the range of the average body weights for these seven species of mammal. [2]

b.i. For the data from these seven species calculate r, the Pearson's product–moment correlation coefficient; [2]

b.ii. For the data from these seven species describe the correlation between the average body weight and the average weight of the brain. [2]

c. Write down the equation of the regression line y on x, in the form $y = mx + c$. [2]

d. Use your regression line to estimate the average weight of the brain of grey wolves. [2]

e. Find the percentage error in your estimate in part (d). [2]

f. State whether it is valid to use the regression line to estimate the average weight of the brain of mice. Give a reason for your answer. [2]

Markscheme

a. $529 - 3$ ***(M1)***

$= 526\ (\text{kg})$ ***(A1)(G2)***

[2 marks]

b.i. $0.922\ (0.921857\ldots)$ ***(G2)***

[2 marks]

b.ii (very) strong, positive ***(A1)*****(ft)*****(A1)*****(ft)**

Note: Follow through from part (b)(i).

[2 marks]

c. $y = 0.000986x + 0.0923$ $(y = 0.000985837\ldots x + 0.0923391\ldots)$ ***(A1)(A1)***

Note: Award ***(A1)*** for $0.000986x$, ***(A1)*** for 0.0923.

Award a maximum of ***(A1)(A0)*** if the answer is not an equation in the form $y = mx + c$.

[2 marks]

d. $0.000985837\ldots(36) + 0.0923391\ldots$ ***(M1)***

Note: Award ***(M1)*** for substituting 36 into their equation.

$0.128\ (\text{kg})\ \ (0.127829\ldots\ (\text{kg}))$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (c). The final ***(A1)*** is awarded only if their answer is positive.

[2 marks]

e. $\left|\frac{0.127829\ldots - 0.120}{0.120}\right| \times 100$ ***(M1)***

Note: Award ***(M1)*** for their correct substitution into percentage error formula.

$6.52\ (\%)\ \ (6.52442...\ (\%))$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (d). Do not accept a negative answer.

[2 marks]

f. Not valid ***(A1)***

the mouse is smaller/lighter/weighs less than the cat (lightest mammal) ***(R1)***

OR

as it would mean the mouse's brain is heavier than the whole mouse ***(R1)***

OR

0.023 kg is outside the given data range. ***(R1)***

OR

Extrapolation ***(R1)***

Note: Do not award ***(A1)(R0)***. Do not accept percentage error as a reason for validity.

[2 marks]

Examiners report

a. [N/A]
b.i. [N/A]
b.ii. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]

A group of candidates sat a Chemistry examination and a Physics examination. The candidates' marks in the Chemistry examination are normally distributed with a mean of 60 and a standard deviation of 12.

a. Draw a diagram that shows this information. [2]

b. Write down the probability that a randomly chosen candidate who sat the Chemistry examination scored at most 60 marks. [1]

c. Hee Jin scored 80 marks in the Chemistry examination. [2]

Find the probability that a randomly chosen candidate who sat the Chemistry examination scored **more** than Hee Jin.

d. The candidates' marks in the Physics examination are normally distributed with a mean of 63 and a standard deviation of 10. Hee Jin also scored 80 marks in the Physics examination. [2]

Find the probability that a randomly chosen candidate who sat the Physics examination scored **less** than Hee Jin.

e. The candidates' marks in the Physics examination are normally distributed with a mean of 63 and a standard deviation of 10. Hee Jin also scored 80 marks in the Physics examination. [2]

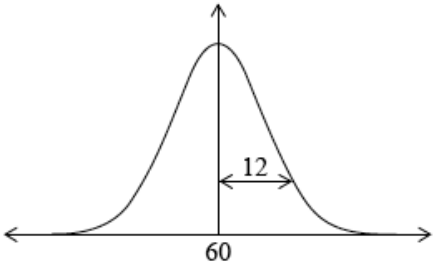
Determine whether Hee Jin's Physics mark, **compared to the other candidates**, is better than her mark in Chemistry. Give a reason for your answer.

f. To obtain a "grade A" a candidate must be in the top 10% of the candidates who sat the Physics examination. [3]

Find the minimum possible mark to obtain a "grade A". Give your answer correct to the nearest integer.

Markscheme

a.



(A1)(A1)

Notes: Award ***(A1)*** for rough sketch of normal curve centred at 60, ***(A1)*** for some indication of 12 as the standard deviation *eg*, as diagram, or with 72 and 48 shown on the horizontal axis in appropriate places, or for 96 and 24 shown on the horizontal axis in appropriate places.

[2 marks]

b. $0.5 \left(\frac{1}{2},\ 50\%\right)$ ***(A1)***

Note: Accept only the exact answer.

[1 mark]

c. $0.0478\ (0.0477903...)$ ***(G2)***

Note: Award ***(G1)*** for $0.952209\ldots$, award ***(M1)(G0)*** for diagram with correct area shown but incorrect answer.

[2 marks]

d. $0.955\ (0.955434...)$ ***(G2)***

Note: Award ***(G1)*** for $0.044565\ldots$, award ***(M1)(G0)*** for diagram with correct area shown but incorrect answer.

[2 marks]

e. $0.0446 < 0.0478$ ***(R1)***

Notes: Award ***(R1)*** for correct comparison seen. Accept alternative methods, for example, $1-$ (their answer to part (c)) used in comparison or a comparison based on z scores.

the Physics result is better ***(A1)*****(ft)**

Notes: Do not award ***(R0)(A1)***. Follow through from their answers to part (c) and part (d).

[2 marks]

f. 76 ***(G3)***

Notes: Award ***(G1)*** for $75.8155\ldots$, award ***(G2)*** for 75.

Award ***(M1)(G0)*** for diagram with correct area shown but incorrect answer.

[3 marks]

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]

The table below shows the distribution of test grades for 50 IB students at Greendale School.

Test grade	1	2	3	4	5	6	7
Frequency	1	3	7	13	11	10	5

A student is chosen at random from these 50 students.

A second student is chosen at random from these 50 students.

The number of minutes that the 50 students spent preparing for the test was normally distributed with a mean of 105 minutes and a standard deviation of 20 minutes.

a.i. Calculate the mean test grade of the students; [2]

a.ii. Calculate the standard deviation. [1]

b. Find the median test grade of the students. [1]

c. Find the interquartile range. [2]

d. Find the probability that this student scored a grade 5 or higher. [2]

e. Given that the first student chosen at random scored a grade 5 or higher, find the probability that both students scored a grade 6. [3]

f.i. Calculate the probability that a student chosen at random spent at least 90 minutes preparing for the test. [2]

f.ii. Calculate the expected number of students that spent at least 90 minutes preparing for the test. [2]

Markscheme

a.i. $\frac{1(1)+3(2)+7(3)+13(4)+11(5)+10(6)+5(7)}{50} = \frac{230}{50}$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into mean formula.

$= 4.6$ ***(A1)*** ***(G2)***

[2 marks]

a.ii. $1.46\ (1.45602\ldots)$ ***(G1)***

[1 mark]

b. 5 ***(A1)***

[1 mark]

c. $6 - 4$ ***(M1)***

Note: Award ***(M1)*** for 6 and 4 seen.

$= 2$ ***(A1)*** ***(G2)***

[2 marks]

d. $\frac{11+10+5}{50}$ ***(M1)***

Note: Award ***(M1)*** for $11 + 10 + 5$ seen.

$= \frac{26}{50}\ \left(\frac{13}{25},\ 0.52,\ 52\%\right)$ ***(A1)*** ***(G2)***

[2 marks]

e. $\frac{10}{\text{their } 26} \times \frac{9}{49}$ ***(M1)(M1)***

Note: Award ***(M1)*** for $\frac{10}{\text{their } 26}$ seen, ***(M1)*** for multiplying their first probability by $\frac{9}{49}$.

OR

$\frac{\frac{10}{50} \times \frac{9}{49}}{\frac{26}{50}}$

Note: Award ***(M1)*** for $\frac{10}{50} \times \frac{9}{49}$ seen, ***(M1)*** for dividing their first probability by $\frac{\text{their } 26}{50}$.

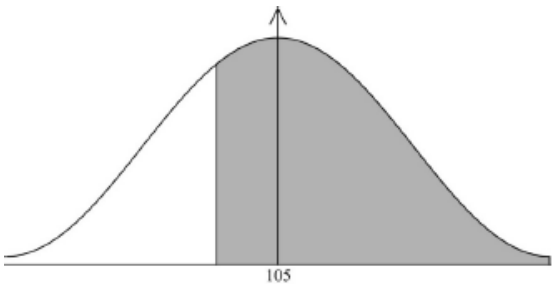
$= \frac{45}{637}\ (0.0706,\ 0.0706436\ldots,\ 7.06436\ldots\%)$ ***(A1)*(ft)** ***(G3)***

Note: Follow through from part (d).

[3 marks]

f.i. $\mathrm{P}(X \geqslant 90)$ ***(M1)***

OR



(M1)

Note: Award ***(M1)*** for a diagram showing the correct shaded region (> 0.5).

$0.773\ (0.773372\ldots)$ $0.773\ (0.773372\ldots,\ 77.3372\ldots\%)$ ***(A1)*** ***(G2)***

[2 marks]

f.ii. $0.773372\ldots \times 50$ ***(M1)***

$= 38.7\ (38.6686\ldots)$ ***(A1)*****(ft)** ***(G2)***

Note: Follow through from part (f)(i).

[2 marks]

Examiners report

a.i. [N/A]
a.ii. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f.i. [N/A]
f.ii. [N/A]

In a school, all Mathematical Studies SL students were given a test. The test contained four questions, each one on a different topic from the syllabus. The quality of each response was classified as satisfactory or not satisfactory. Each student answered only three of the four questions, each on a separate answer sheet.

The table below shows the number of satisfactory and not satisfactory responses for each question.

		Topic from the syllabus				
		Calculus	**Probability**	**Geometry**	**Logic**	**Total**
Quality of response	**Satisfactory**	10	16	20	14	60
	Not satisfactory	8	6	10	6	30
	Total	18	22	30	20	90

A χ^2 test is carried out at the 5% significance level for the data in the table.

The critical value for this test is 7.815.

a.i. If the teacher chooses a response at random, find the probability that it is a response to the Calculus question; [2]

a.ii. If the teacher chooses a response at random, find the probability that it is a satisfactory response to the Calculus question; [2]

a.iiiIf the teacher chooses a response at random, find the probability that it is a satisfactory response, given that it is a response to the Calculus question. [2]

b. The teacher groups the responses by topic, and chooses two responses to the Logic question. Find the probability that both are not satisfactory. [3]

c. State the null hypothesis for this test. [1]

d. Show that the expected frequency of satisfactory Calculus responses is 12. [1]

e. Write down the number of degrees of freedom for this test. [1]

f. Use your graphic display calculator to find the χ^2 statistic for this data. [2]

g. State the conclusion of this χ^2 test. Give a reason for your answer. [2]

Markscheme

a.i. $\frac{1}{5}$ $\left(\frac{18}{90};\ 0.2;\ 20\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[2 marks]

a.ii. $\frac{1}{9}$ $\left(\frac{10}{90};\ 0.\overline{1};\ 0.111111\ldots;\ 11.1\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[2 marks]

a.iii $\frac{5}{9}$ $\left(\frac{10}{18};\ 0.\overline{5};\ 0.555556\ldots;\ 55.6\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[2 marks]

b. $\frac{6}{20} \times \frac{5}{19}$ ***(A1)(M1)***

Note: Award ***(A1)*** for two correct fractions seen, ***(M1)*** for multiplying their two fractions.

$\frac{3}{38}$ $\left(\frac{30}{380};\ 0.0789473\ldots;\ 7.89\%\right)$ ***(A1)(G2)***

[3 marks]

c. H_0: quality (of response) and topic (from the syllabus) are independent ***(A1)***

Note: Accept there is no association between quality (of response) and topic (from the syllabus). Do not accept "not related" or "not correlated" or "influenced".

[1 mark]

d. $\frac{18}{90} \times \frac{60}{90} \times 90$ **OR** $\frac{18 \times 60}{90}$ ***(M1)***

Note: Award ***(M1)*** for correct substitution in expected value formula.

$(=) 12$ ***(AG)***

Note: The conclusion, $(=) 12$, must be seen for the ***(A1)*** to be awarded.

[1 mark]

e. 3 ***(A1)***

[1 mark]

f. $(\chi^2_{calc} =) 1.46\ (1.46\overline{36};\ 1.46363\ldots)$ ***(G2)***

[2 marks]

g. $1.46 < 7.815$ **OR** $0.690688\ldots > 0.05$ ***(R1)***

the null hypothesis is not rejected ***(A1)*****(ft)**

OR

the quality of the response and the topic are independent ***(A1)*****(ft)**

Note: Award ***(R1)*** for a correct comparison of either their χ^2 statistic to the χ^2 critical value or the correct p-value 0.690688… to the test level, award ***(A1)*****(ft)** for the correct result from that comparison. Accept "$\chi^2_{\text{calc}} < \chi^2_{\text{crit}}$" for the comparison, but only if their χ^2_{calc} value is explicitly seen in part (f). Follow through from their answers to part (f) and part (c). Do not award ***(R0)(A1)***.

[2 marks]

Examiners report

a.i. [N/A]
a.ii. [N/A]
a.iii. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]

In the month before their IB Diploma examinations, eight male students recorded the number of hours they spent on social media.

For each student, the number of hours spent on social media (x) and the number of IB Diploma points obtained (y) are shown in the following table.

Hours on social media (x)	6	15	26	12	13	40	33	23
IB Diploma points (y)	43	33	27	36	39	17	20	33

Use your graphic display calculator to find

Ten female students also recorded the number of hours they spent on social media in the month before their IB Diploma examinations. Each of these female students spent between 3 and 30 hours on social media.

The equation of the regression line *y* on *x* for these ten female students is

$$y = -\frac{2}{3}x + \frac{125}{3}.$$

An eleventh girl spent 34 hours on social media in the month before her IB Diploma examinations.

a. On graph paper, draw a scatter diagram for these data. Use a scale of 2 cm to represent 5 hours on the x-axis and 2 cm to represent 10 points on the y-axis. [4]

b. (i) $\bar{x}$, the mean number of hours spent on social media; [2]

(ii) $\bar{y}$, the mean number of IB Diploma points.

c. Plot the point $(\bar{x},\ \bar{y})$ on your scatter diagram and label this point M. [2]

d. Write down the value of r, the Pearson's product–moment correlation coefficient, for these data. [2]

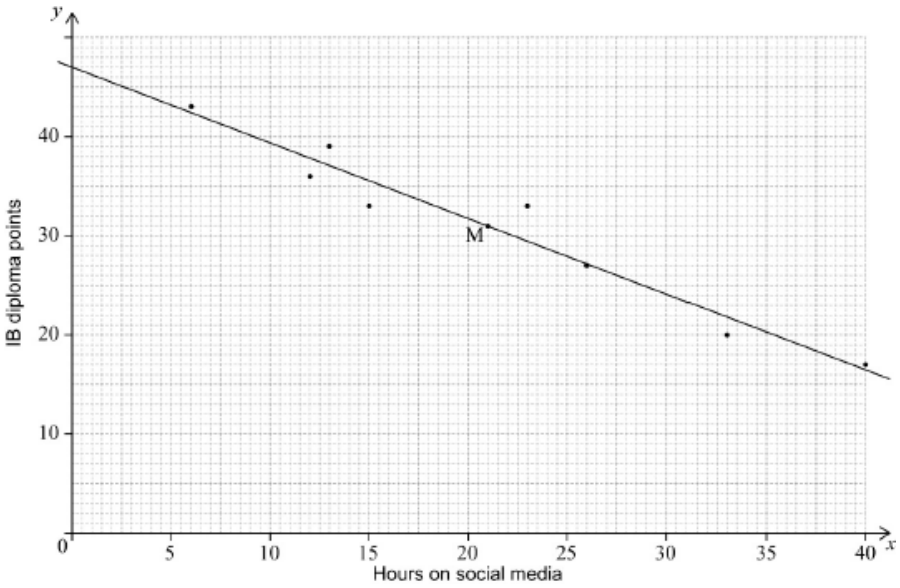
e. Write down the equation of the regression line y on x for these eight male students. [2]

f. Draw the regression line, from part (e), on your scatter diagram. [2]

g. Use the given equation of the regression line to estimate the number of IB Diploma points that this girl obtained. [2]

h. Write down a reason why this estimate is not reliable. [1]

Markscheme

a. 

(A4)

Notes: Award ***(A1)*** for correct scale and labelled axes.

Award ***(A3)*** for 7 or 8 points correctly plotted,

(A2) for 5 or 6 points correctly plotted,

(A1) for 3 or 4 points correctly plotted.

Award at most ***(A0)(A3)*** if axes reversed.

Accept x and y sufficient for labelling.

If graph paper is not used, award ***(A0)***.

If an inconsistent scale is used, award ***(A0)***. Candidates' points should be read from this scale **where possible** and awarded accordingly.

A scale which is too small to be meaningful (ie mm instead of cm) earns ***(A0)*** for plotted points.

[4 marks]

b. (i) $\bar{x} = 21$ ***(A1)***

(ii) $\bar{y} = 31$ ***(A1)***

[2 marks]

c. $(\bar{x},\ \bar{y})$ correctly plotted on graph ***(A1)*****(ft)**

this point labelled M ***(A1)***

Note: Follow through from parts (b)(i) and (b)(ii).

Only accept M for labelling.

[2 marks]

d. $-0.973\ (-0.973388\ldots)$ ***(G2)***

Note: Award ***(G1)*** for 0.973, without minus sign.

[2 marks]

e. $y = -0.761x + 47.0$ $(y = -0.760638\ldots x + 46.9734\ldots)$ ***(A1)(A1)(G2)***

Notes: Award ***(A1)*** for $-0.761x$ and ***(A1)*** $+47.0$. Award a maximum of ***(A1)(A0)*** if answer is not an equation.

[2 marks]

f. line on graph ***(A1)*****(ft)*****(A1)*****(ft)**

Notes: Award ***(A1)*****(ft)** for **straight line** that passes through their M, ***(A1)*****(ft)** for line (extrapolated if necessary) that passes through $(0,\ 47.0)$.

If M is not plotted or labelled, follow through from part (e).

[2 marks]

g. $y = -\frac{2}{3}(34) + \frac{125}{3}$ ***(M1)***

Note: Award ***(M1)*** for correct substitution.

19 (points) ***(A1)(G2)***

[2 marks]

h. extrapolation ***(R1)***

OR

34 hours is outside the given range of data ***(R1)***

Note: Do not accept 'outlier'.

[1 mark]

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]
h. [N/A]

The table below shows the scores for 12 golfers for their first two rounds in a local golf tournament.

Round 1 (x)	71	79	66	73	69	76	68	75	82	67	69	74
Round 2 (y)	73	81	68	75	70	79	69	77	83	68	72	76

a. (i) Write down the mean score in Round 1. [5]

(ii) Write down the standard deviation in Round 1.

(iii) Find the number of these golfers that had a score of more than one standard deviation above the mean in Round 1.

b. Write down the correlation coefficient, *r*. [2]

c. Write down the equation of the regression line of *y* on *x*. [2]

d. Another golfer scored 70 in Round 1. [2]

Calculate an estimate of his score in Round 2.

e. Another golfer scored 89 in Round 1. [2]

Determine whether you can use the equation of the regression line to estimate his score in Round 2. Give a reason for your answer.

Markscheme

a. (i) $\frac{71+79+...}{12}$ ***(M1)***

$72.4\left(72.4166..., \frac{869}{12}\right)$ ***(A1)(G2)***

Note: Award ***(M1)*** for correct substitution into the mean formula.

(ii) 4.77 (4.76896…) ***(G1)***

(iii) 72.4 + 4.77 = 77.17 ***(M1)***

Note: Award ***(M1)*** for adding their mean to their standard deviation.

Two golfers ***(A1)*****(ft)*****(G2)***

Note: Follow through from their answers to parts (i) and (ii).

[5 marks]

b. 0.990 (0.99014…) ***(G2)***

[2 marks]

c. $y = 1.01x + 0.816$ ($y = 1.01404...x + 0.81618...$) ***(G1)(G1)***

Notes: Award ***(G1)*** for $1.01x$ and ***(G1)*** for 0.816. If the answer is not an equation award a maximum of ***(G1)(G0)***.

OR

$y - 74.25 = 1.01(x - 72.4)$ ($y - 74.25 = 1.01404...(x - 72.4166...)$) ***(A1)(A1)***

Notes: Award ***(A1)*** for 1.01 correctly substituted in the equation, and ***(A1)*****(ft)** for correct substitution of (72.4, 74.25) in the equation. Follow through from their part (a)(i). If the final answer is not an equation award a maximum of ***(A1)(A0)***.

[2 marks]

d. $y = 1.01404... \times 70 + 0.81618...$ ***(M1)***

Note: Award ***(M1)*** for substitution of 70 into their regression line equation from part (c).

y = 72 (71.7989...) ***(A1)*****(ft)*****(G2)***

Note: Follow through from their part (c).

[2 marks]

e. No, equation cannot be (reliably) used as 89 is outside the data range. ***(A1)(R1)***

OR

Yes, but the result is not valid/not reliable as 89 is outside the data range/as we extrapolate ***(A1)(R1)***

Note: Do not award ***(A1)(R0)***.

[2 marks]

Examiners report

a. The question was for the most part approached by almost all candidates and answered relatively well. The question in part (e) related to the use of the equation of the regression line for predicting, although regularly asked on exams, was still found to be a difficult one by some candidates. Some answers still suggested mathematical thinking and language unaccustomed to drawing conclusions and providing justifications.

b. The question was for the most part approached by almost all candidates and answered relatively well. The question in part (e) related to the use of the equation of the regression line for predicting, although regularly asked on exams, was still found to be a difficult one by some candidates. Some answers still suggested mathematical thinking and language unaccustomed to drawing conclusions and providing justifications.

c. The question was for the most part approached by almost all candidates and answered relatively well. The question in part (e) related to the use of the equation of the regression line for predicting, although regularly asked on exams, was still found to be a difficult one by some candidates. Some answers still suggested mathematical thinking and language unaccustomed to drawing conclusions and providing justifications.

d. The question was for the most part approached by almost all candidates and answered relatively well. The question in part (e) related to the use of the equation of the regression line for predicting, although regularly asked on exams, was still found to be a difficult one by some candidates. Some answers still suggested mathematical thinking and language unaccustomed to drawing conclusions and providing justifications.

e. The question was for the most part approached by almost all candidates and answered relatively well. The question in part (e) related to the use of the equation of the regression line for predicting, although regularly asked on exams, was still found to be a difficult one by some candidates. Some answers still suggested mathematical thinking and language unaccustomed to drawing conclusions and providing justifications.

In a mountain region there appears to be a relationship between the number of trees growing in the region and the depth of snow in winter. A set of 10 areas was chosen, and in each area the number of trees was counted and the depth of snow measured. The results are given in the table below.

Number of trees (x)	Depth of snow in cm (y)
45	30
75	50
66	40
27	25
44	30
28	5
60	35
35	20
73	45
47	25

In a study on 100 students there seemed to be a difference between males and females in their choice of favourite car colour. The results are given in the table below. A χ^2 test was conducted.

	Blue	Red	Green
Males	14	6	8
Females	31	24	17

A, a(i) Use your graphic display calculator to find the mean number of trees. [1]

A, a(ii) Use your graphic display calculator to find the mean depth of snow. [1]

A, a(iii) Use your graphic display calculator to find the standard deviation of the depth of snow. [1]

A, b The covariance, $S_{xy} = 188.5$. [2]

Write down the product-moment correlation coefficient, r.

A, c Write down the equation of the regression line of y on x. [2]

A, d(i) If the number of trees in an area is 55, estimate the depth of snow. [2]

A, e(i) Use the equation of the regression line to estimate the depth of snow in an area with 100 trees. [1]

A, e(ii) Decide whether the answer in (e)(i) is a valid estimate of the depth of snow in the area. Give a reason for your answer. [2]

B, a Write down the total number of male students. [1]

B, b Show that the expected frequency for males, whose favourite car colour is blue, is 12.6. [2]

B, c The calculated value of χ^2 is 1.367 and the critical value of χ^2 is 5.99 at the 5% significance level. [1]

Write down the null hypothesis for this test.

B, c The calculated value of χ^2 is 1.367 and the critical value of χ^2 is 5.99 at the 5% significance level. [1]

Write down the number of degrees of freedom.

B, c, iv. The calculated value of χ^2 is 1.367 and the critical value of χ^2 is 5.99 at the 5% significance level. [2]

Determine whether the null hypothesis should be accepted at the 5% significance level. Give a reason for your answer.

Markscheme

A, a, i. 50. ***(G1)***

[1 mark]

A, a, ii. 30 ***(G1)***

[1 mark]

A, a, iii. 12.3 ***(G1)***

Note: Award ***(A1)*****(ft)** for 13.0 in (iv) but only if 17.7 seen in (a)(ii).

[1 mark]

A, b. $r = \frac{188.5}{(16.79 \times 12.33)}$ ***(M1)***

Note: Award ***(M1)*** for using their values in the correct formula.

= 0.911 (accept 0.912, 0.910) ***(A1)*****(ft)*****(G2)***

[2 marks]

A, c. $y = 0.669x - 2.95$ ***(G1)(G1)***

Note: Award ***(G1)*** for $0.669x$, ***(G1)*** for -2.95. If the answer is not in the form of an equation, award at most ***(G1)(G0)***.

[2 marks]

A, d. Depth = 0.669 × 55 − 2.95 ***(M1)***

= 33.8 ***(A1)*****(ft)*****(G2)*****(ft)**

Note: Follow through from their (c) even if no working seen.

[2 marks]

A, e. 64.0 (accept 63.95, 63.9) ***(A1)*****(ft)*****(G1)*****(ft)**

Note: Follow through from their (c) even if no working seen.

[1 mark]

A, e, i. It is not valid. It lies too far outside the values that are given. *Or equivalent.* ***(A1)(R1)***

Note: Do not award ***(A1)(R0)***.

[2 marks]

B, a. 28 ***(A1)***

[1 mark]

B, b. $\frac{28\times45}{100}\left(\frac{28}{100}\times\frac{45}{100}\times100\right)$ ***(M1)(A1)(ft)***

Note: Award ***(M1)*** for correct formula, ***(A1)*** for correct substitution.

= 12.6 ***(AG)***

Note: Do not award ***(A1)*** unless 12.6 seen.

[2 marks]

B, c. the favourite car colour is **independent** of gender. ***(A1)***

Note: Accept there is no association between gender and favourite car colour.

Do not accept 'not related' or 'not correlated'.

[1 mark]

B, d. ii. ***(A1)***

[1 marks]

B, e. Accept the null hypothesis since $1.367 < 5.991$ ***(A1)*****(ft)*****(R1)***

Note: Allow "Do not reject". Follow through from their null hypothesis and their critical value.

Full credit for use of p-values from GDC [$p = 0.505$].

Do not award ***(A1)(R0)***. Award ***(R1)*** for valid comparison.

[2 marks]

Examiners report

A, a. A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

The use of the inappropriate standard deviation was seen, but infrequently.

A, a(i) A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

The use of the inappropriate standard deviation was seen, but infrequently.

A, a(iv) A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

The use of the inappropriate standard deviation was seen, but infrequently.

A, b A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

It is expected that the GDC is used to calculate the correlation coefficient; the covariance was given to aid those candidates for whom the reset process removes this function from the display. It is anticipated that this hint will not be given in future papers.

A, c A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

A, d A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

A, e(i) A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

A, e(ii) A straightforward question that saw many fine attempts. Given its nature – where much of the work was done on the GDC – it must be emphasised to candidates that incorrect entry of data into the calculator will result in considerable penalties; they must check their data entry most carefully.

The dangers of extrapolation should be clearly explained to students.

B, a Once again, a straightforward question on chi-squared testing that was either highly successful (for the majority) or showed a lack of syllabus coverage.

B, b Once again, a straightforward question on chi-squared testing that was either highly successful (for the majority) or showed a lack of syllabus coverage. A surprising number of candidates lacked knowledge of the theory underlying the test and were thus unable to attempt (b).

B, Once again, a straightforward question on chi-squared testing that was either highly successful (for the majority) or showed a lack of syllabus coverage. In (c)(i) it is worth stressing that the test is for the mathematical **independence** of two characteristics and this determines the null hypothesis.

B, Once again, a straightforward question on chi-squared testing that was either highly successful (for the majority) or showed a lack of syllabus coverage.

B, Once again, a straightforward question on chi-squared testing that was either highly successful (for the majority) or showed a lack of syllabus coverage. A number of candidates confuse the critical value and p-value approach to the test and thus lost marks in (c)(iv).

A group of 800 students answered 40 questions on a category of their choice out of History, Science and Literature.

For each student the category and the number of correct answers, N, was recorded. The results obtained are represented in the following table.

		Number of correct answers, N				
		$1 \leq N \leq 10$	$11 \leq N \leq 20$	$21 \leq N \leq 30$	$31 \leq N \leq 40$	**Total number of students**
Category	History	46	80	68	39	**233**
	Science	37	82	85	56	**260**
	Literature	31	110	104	62	**307**
	Total number of students	**114**	**272**	**257**	**157**	**800**

A χ^2 test at the 5% significance level is carried out on the results. The critical value for this test is 12.592.

a. State whether N is a discrete or a continuous variable. [1]

b.i. Write down, for N, the modal class; [1]

b.ii. Write down, for N, the mid-interval value of the modal class. [1]

c.i. Use your graphic display calculator to estimate the mean of N; [2]

c.ii. Use your graphic display calculator to estimate the standard deviation of N. [1]

d. Find the expected frequency of students choosing the Science category and obtaining 31 to 40 correct answers. [2]

e.i. Write down the null hypothesis for this test; [1]

e.ii. Write down the number of degrees of freedom. [1]

f.i. Write down the p-value for the test; [1]

f.ii. Write down the χ^2 statistic. [2]

g. State the result of the test. Give a reason for your answer. [2]

Markscheme

a. discrete ***(A1)***

[1 mark]

b.i. $11 \leqslant N \leqslant 20$ ***(A1)***

[1 mark]

b.ii. 15.5 ***(A1)*****(ft)**

Note: Follow through from part (b)(i).

[1 mark]

c.i. $21.2\ (21.2125)$ ***(G2)***

[2 marks]

c.ii. $9.60\ (9.60428\ldots)$ ***(G1)***

[1 marks]

d. $\frac{260}{800} \times \frac{157}{800} \times 800$ **OR** $\frac{260 \times 157}{800}$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into expected frequency formula.

$= 51.0\ (51.025)$ ***(A1)(G2)***

[2 marks]

e.i. choice of category and number of correct answers are independent ***(A1)***

Notes: Accept “no association” between (choice of) category and number of correct answers. Do not accept “not related” or “not correlated” or “influenced”.

[1 mark]

e.ii. 6 ***(A1)***

[1 mark]

f.i. $0.0644\ (0.0644123\ldots)$ ***(G1)***

[1 mark]

f.ii. $11.9\ (11.8924\ldots)$ ***(G2)***

[2 marks]

g. the null hypothesis is not rejected (the null hypothesis is accepted) ***(A1)*****(ft)**

OR

(choice of) category and number of correct answers are independent ***(A1)*****(ft)**

as $11.9 < 12.592$ **OR** $0.0644 > 0.05$ ***(R1)***

Notes: Award ***(R1)*** for a correct comparison of either their χ^2 statistic to the χ^2 critical value or their p-value to the significance level. Award ***(A1)*****(ft)** from that comparison.

Follow through from part (f). Do not award ***(A1)*****(ft)*****(R0)***.

[2 marks]

Examiners report

a. [N/A]
b.i. [N/A]
b.ii. [N/A]
c.i. [N/A]
c.ii. [N/A]
d. [N/A]
e.i. [N/A]
e.ii. [N/A]
f.i. [N/A]
f.ii. [N/A]
g. [N/A]

Part A

100 students are asked what they had for breakfast on a particular morning. There were three choices: cereal (X) , bread (Y) and fruit (Z). It is found that

10 students had all three

17 students had bread and fruit only

15 students had cereal and fruit only

12 students had cereal and bread only

13 students had only bread

8 students had only cereal

9 students had only fruit

Part B

The same 100 students are also asked how many meals on average they have per day. The data collected is organized in the following table.

	3 or fewer meals per day	4 or 5 meals per day	More than 5 meals per day	Total
Male	15	25	15	55
Female	12	20	13	45
Total	27	45	28	100

A χ^2 test is carried out at the 5 % level of significance.

A.a. Represent this information on a Venn diagram. [4]

A.b. Find the number of students who had none of the three choices for breakfast. [2]

A.c. Write down the percentage of students who had fruit for breakfast. [2]

A.d. Describe in words what the students in the set $X \cap Y'$ had for breakfast. [2]

A.e. Find the probability that a student had **at least** two of the three choices for breakfast. [2]

A.f. Two students are chosen at random. Find the probability that both students had all three choices for breakfast. [3]

B.a. Write down the null hypothesis, H_0, for this test. [1]

B.b. Write down the number of degrees of freedom for this test. [1]

B.c. Write down the critical value for this test. [1]

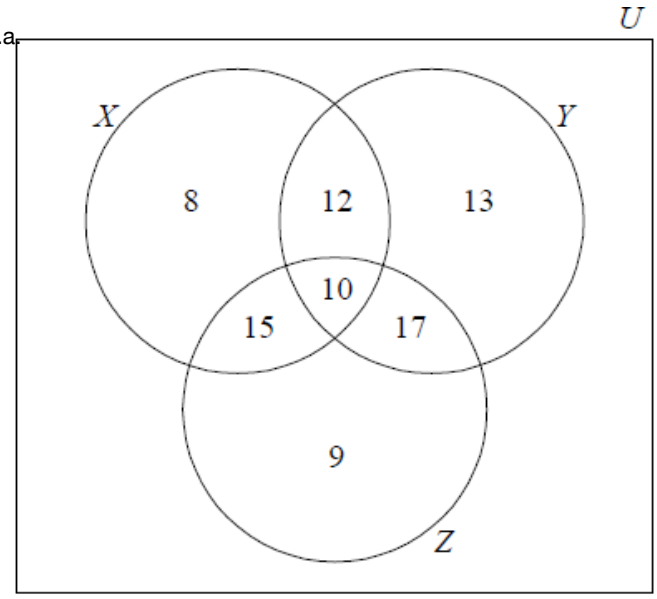
B.d. Show that the expected number of females that have more than 5 meals per day is 13, correct to the nearest integer. [2]

B.e. Use your graphic display calculator to find the χ^2_{calc} for this data. [2]

B.f. Decide whether H_0 must be accepted. Justify your answer. [2]

Markscheme

A.a.



(A1) for rectangle and three intersecting circles

(A1) for 10, ***(A1)*** for 8, 13 and 9, ***(A1)*** for 12, 15 and 17 ***(A4)***

[4 marks]

A.b. 100 – (9 +12 +13 +15 +10 +17 + 8) =16 ***(M1)(A1)*****(ft)*****(G2)***

Note: Follow through from their diagram.

[2 marks]

A.c. $\frac{51}{100}(0.51)$ ***(A1)*****(ft)**

= 51% ***(A1)*****(ft)*****(G2)***

Note: Follow through from their diagram.

[2 marks]

A.d. **Note:** The following statements are correct. Please note that the connectives are important. It is not the same (had cereal) and (not bread) and (had cereal) or (not bread). The parentheses are not needed but are there to facilitate the understanding of the propositions.

(had cereal) and (did not have bread)

(had cereal only) or (had cereal and fruit only)

(had either cereal or (fruit and cereal)) and (did not have bread) ***(A1)(A1)***

Notes: If the statements are correct but the connectives are wrong then award at most ***(A1)(A0)***. For the statement (had only cereal) and (cereal and fruit) award ***(A1)(A0)***. For the statement had cereal and fruit award ***(A0)(A0)***.

[2 marks]

A.e. $\frac{54}{100}(0.54,\ 54\ \%)$ ***(A1)*****(ft)*****(A1)*****(ft)*****(G2)***

Note: Award ***(A1)*****(ft)** for numerator, follow through from their diagram, ***(A1)*****(ft)** for denominator. Follow through from total or denominator used in part (c).

[2 marks]

A.f. $\frac{10}{100} \times \frac{9}{99} = \frac{1}{110}(0.00909,\ 0.909\ \%)$ ***(A1)*****(ft)*****(M1)*****(A1)*****(ft)*****(G2)***

Notes: Award ***(A1)*****(ft)** for their correct fractions, ***(M1)*** for multiplying two fractions, ***(A1)*****(ft)** for their correct answer. Answer 0.009 with no working receives no marks. Follow through from denominator in parts (c) and (e) and from their diagram.

[3 marks]

B.a. H_0 : The (average) number of meals per day a student has and gender are independent ***(A1)***

Note: For “independent” accept “not associated” but do not accept “not related” or “not correlated”.

[1 mark]

B.b. 2 ***(A1)***

[1 mark]

B.c. 5.99 (accept 5.991) ***(A1)*****(ft)**

Note: Follow through from their part (b).

[1 mark]

B.d. $\frac{28\times45}{100} = 12.6 = 13$ or $\frac{28}{100} \times \frac{25}{100} \times 100 = 12.6 = 13$ ***(M1)(A1)(AG)***

Notes: Award ***(M1)*** for correct formula and ***(A1)*** for correct substitution. Unrounded answer must be seen for the ***(A1)*** to be awarded.

[2 marks]

B.e 0.0321 ***(G2)***

Note: For 0.032 award ***(G1)(G1)(AP)***. For 0.03 with no working award ***(G0)***.

[2 marks]

B.f 0.0321 < 5.99 or 0.984 > 0.05 ***(R1)***

accept H_0 ***(A1)*****(ft)**

Note: If reason is incorrect both marks are lost, do not award ***(R0)(A1)***.

[2 marks]

Examiners report

A.a This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates. Some misread what Y was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

A.b This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates. Some misread what Y was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

A.c This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates. Some misread what Y was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

A.d This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates.

Some misread what *Y* was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

A.e. This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates. Some misread what *Y* was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

A.f. This question was in general well done. Candidates began the paper well by drawing the Venn diagram correctly. Some students omitted the rectangle (universal set) around the three circles. There were quite a few errors in (c) as some students forgot to convert their answers to percentages. Also describing in words what the students in $X \cap Y'$ had for breakfast seemed to be difficult for the majority of the candidates. Some misread what *Y* was and even more missed the complement sign. However, the main problem in answering this question seemed to be the lack of knowledge in the relationship between set theory and logic (use of "and" and "or"). Combining probabilities caused problems to many. Common wrong answers were $\frac{10}{100}$, $\frac{10}{100} \times \frac{10}{100}$ or $\frac{10}{100} + \frac{9}{99}$.

B.a. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

B.b. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

B.c. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

B.d. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

B.e. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

B.f. In general this part question was well answered. The major concerns of the examining team were the following:

- In (f) many students wrote down the expected values table (from the GDC) and highlighted the correct expected value, 12.6. As this is a "show that" question the use of the GDC is not expected and therefore no marks are awarded for this working. Instead it is expected the use of the formula for the expected value with the correct substitutions.
- In (e) surprisingly many candidates found the χ^2_{calc} through the use of the formula. Unfortunately this led to some incorrect answers and also to a bad use of time. The question clearly says "use your graphic display calculator" and it is worth 2 marks therefore a student should not spend more than 2 minutes to answer this part question. Time management is essential in this type of examinations and the IB rule is one minute – one mark.

The table shows the distance, in km, of eight regional railway stations from a city centre terminus and the price, in \$, of a return ticket from each regional station to the terminus.

Distance in km (x)	3	15	23	42	56	62	74	93
Price in \$ (y)	5	24	43	56	68	74	86	100

a. Draw a scatter diagram for the above data. Use a scale of 1 cm to represent 10 km on the x-axis and 1 cm to represent $\$10$ on the y-axis. [4]

b. Use your graphic display calculator to find [2]

(i) $\bar{x}$, the mean of the distances;

(ii) $\bar{y}$, the mean of the prices.

c. Plot and label the point $\mathrm{M}\ (\bar{x},\ \bar{y})$ on your scatter diagram. [1]

d. Use your graphic display calculator to find [3]

(i) the product–moment correlation coefficient, r ;

(ii) the equation of the regression line y on x.

e. Draw the regression line y on x on your scatter diagram. [2]

f. A ninth regional station is 76 km from the city centre terminus. [3]

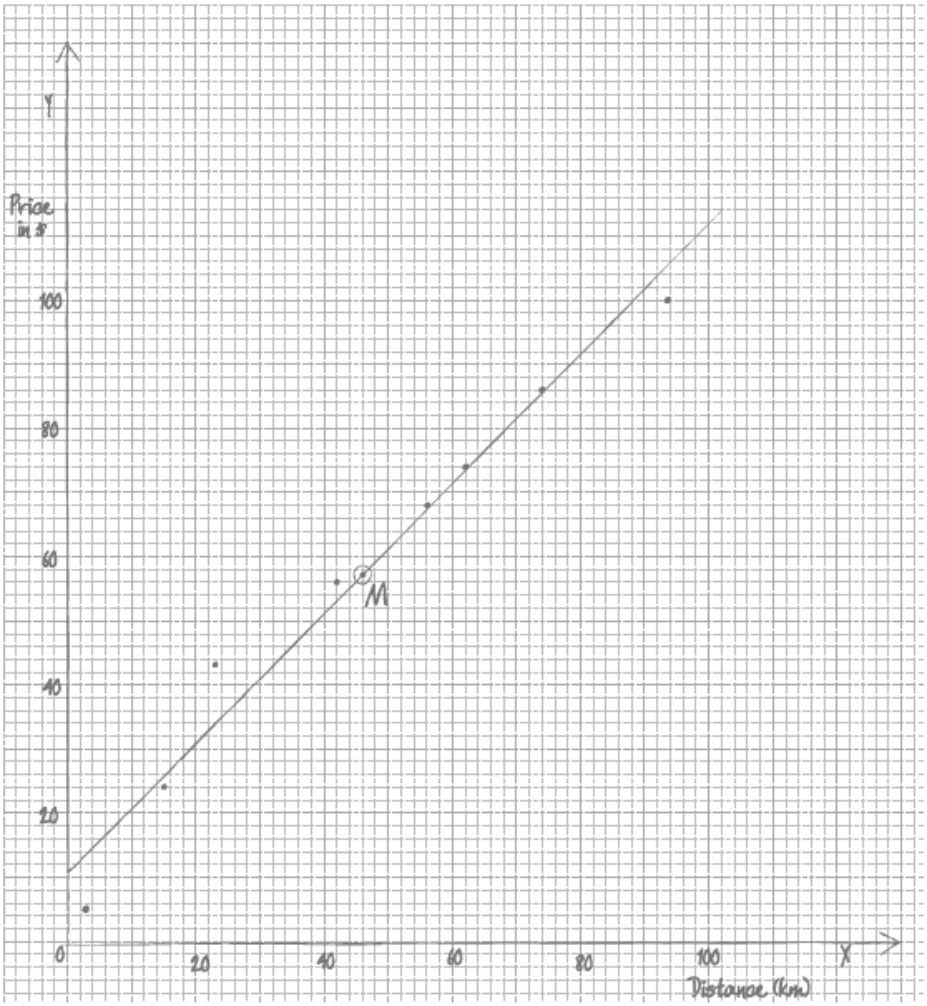
Use the equation of the regression line to estimate the price of a return ticket to the city centre terminus from this regional station. **Give your answer correct to the nearest $\$$.**

g. Give a reason why it is valid to use your regression line to estimate the price of this return ticket. [1]

h. The actual price of the return ticket is $\$80$. [2]

Using your answer to part (f), calculate the percentage error in the estimated price of the ticket.

Markscheme

a. 

(A4)

Notes: Award ***(A1)*** for correct scale and labels (accept x and y).

Award ***(A3)*** for 7 or 8 points plotted correctly.

Award ***(A2)*** for 5 or 6 points plotted correctly.

Award ***(A1)*** for 3 or 4 points plotted correctly.

Award at most ***(A1)(A2)*** if points are joined up.

If axes are reversed, award at most ***(A0)(A3)***.

If graph paper is not used, award at most ***(A1)(A0)***.

[4 marks]

b. (i) $(\bar{x} =) 46$ ***(G1)***

(ii) $(\bar{y} =) 57$ ***(G1)***

[2 marks]

c. $\mathrm{M}(46, 57)$ plotted and labelled on the scatter diagram ***(A1)*****(ft)**

Notes: Follow through from their part (b).

Accept $(\bar{x},\ \bar{y})$ as the label.

[1 mark]

d. (i) $0.986 \quad (0.986322...)$ ***(G1)***

(ii) $y = 1.01x + 10.3 \quad (y = 1.01431\ldots x + 10.3412\ldots)$ ***(G1)(G1)***

Notes: Award ***(G1)*** for $1.01x$, ***(G1)*** for 10.3.

Award ***(G1)(G0)*** if not written in the form of an equation.

OR

$(y - 57) = 1.01(x - 46) \quad (y - 57 = 1.01431...(x - 46))$ ***(G1)(G1)*****(ft)**

Note: Award ***(G1)*** for 1.01, ***(G1)*** for their 57 and 46.

[3 marks]

e. straight line drawn on the scatter diagram ***(A1)*****(ft)*****(A1)*****(ft)**

Notes: The line must be straight for either of the two marks to be awarded.

Award ***(A1)*****(ft)** passing through their M plotted in (c).

Award ***(A1)*****(ft)** for correct y-intercept (between 9 and 12).

Follow through from their y-intercept found in part (d).

If part (d) is used, award ***(A1)*****(ft)** for their intercept (± 1).

[2 marks]

f. $y = 1.01431... \times 76 + 10.3412\ldots$ ***(M1)***

Note: Award ***(M1)*** for substitution of 76 into their regression line.

$= 87.4295\ldots$ ***(A1)*****(ft)**

Note: Follow through from part (d). If 3 sf values are used the value is 87.06.

$\$87$ ***(A1)*****(ft)*****(G2)***

Notes: The final ***(A1)*** is awarded for their answer given correct to the nearest dollar.

Method, followed by the answer of 87 earns ***(M1)(G2).*** It is not necessary to see the interim step.

Where the candidate uses their graph instead of the equation, and arrives at an answer other than 87, award, at most, ***(G1)*(ft)**.

If the candidate uses their graph and arrives at the required answer of 87, award ***(G2)*(ft)**.

[3 marks]

g. 76 is within the range of distances given in the data **OR** the correlation coefficient is close to 1. ***(R1)***

Notes: Award ***(R1)*** if **either** condition is given.

Sufficient to indicate that 76 is 'within the data range' and the correlation is 'strong'.

Allow r^2 close to 1.

Do **not** accept "within the range of prices".

[1 mark]

h. $\text{Percentage error} = \frac{87-80}{80} \times 100$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into formula.

8.75% ***(A1)*(ft)*(G2)***

Notes: Follow through from their answer to part (f).

Accept either the rounded or unrounded answer to part (f).

If no integer value seen in part (f), follow through from their unrounded answer to part (f).

Answer must be positive.

[2 marks]

Examiners report

a. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point M, a significant number of candidates seemed to draw their line 'by eye' rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and 'within the data range' or a 'strong correlation' were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

b. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point *M*, a significant number of candidates seemed to draw their line 'by eye' rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and 'within the data range' or a 'strong correlation' were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

c. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point *M*, a significant number of candidates seemed to draw their line 'by eye' rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and 'within the data range' or a 'strong correlation' were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

d. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point *M*, a significant number of candidates seemed to draw their line 'by eye' rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and 'within the data range' or a 'strong correlation' were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

e. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point *M*, a significant number of candidates seemed to draw their line 'by eye' rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and

full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and ‘within the data range’ or a ‘strong correlation’ were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

f. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point M, a significant number of candidates seemed to draw their line ‘by eye’ rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and ‘within the data range’ or a ‘strong correlation’ were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

g. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point M, a significant number of candidates seemed to draw their line ‘by eye’ rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and ‘within the data range’ or a ‘strong correlation’ were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

h. This question was very well attempted by a significant majority of candidates. Many good and accurate attempts at plotting a scatter diagram were seen in part (a). However, a minority of candidates chose not to use graph paper but instead used their answer book. These candidates achieved, at most, one mark for that part question. Many correct answers were seen in parts (b) and (d) reflecting good use of the graphic display calculator. Whilst many candidates realized that the line of regression passes through the point M, a significant number of candidates seemed to draw their line ‘by eye’ rather than using the equation found in part (d) and, as a consequence for many, their straight line (or projected line) did not fall within the required tolerances for the second mark. Many candidates understood the requirements for part (f) and full marks were seen on a majority of scripts. Those candidates, however, who used their graph instead scored, at most, two marks here. Many candidates seemed to be well-drilled in giving a suitable reason in part (f) and ‘within the data range’ or a ‘strong correlation’ were frequently seen. Percentage error caused very few problems for candidates and many correct answers were seen in part (h).

In a debate on voting, a survey was conducted. The survey asked people's opinion on whether or not the minimum voting age should be reduced to 16 years of age. The results are shown as follows.

	Age 18–25	Age 26–40	Age 41+	Total
Oppose the reduction	12	20	48	80
Favour the reduction	18	15	17	50
Total	30	35	65	130

A χ^2 test at the 1% significance level was conducted. The χ^2 critical value of the test is 9.21.

a. State [2]

(i) H_0, the null hypothesis for the test;

(ii) H_1, the alternative hypothesis for the test.

b. Write down the number of degrees of freedom. [1]

c. Show that the expected frequency of those between the ages of 26 and 40 who oppose the reduction in the voting age is 21.5, correct to three significant figures. [2]

d. Find [3]

(i) the χ^2 statistic;

(ii) the associated p-value for the test.

e. Determine, giving a reason, whether H_0 should be accepted. [2]

Markscheme

a. (i) H_0 age and opinion (about the reduction) are independent. ***(A1)***

Notes: Accept "not associated" instead of independent.

(ii) H_1 age and opinion are not independent. ***(A1)*****(ft)**

Notes: Follow through from part (a)(i). Accept "associated" or "dependent".

Award ***(A1)*****(ft)** for their correct H_1 worded consistently with their part (a)(i).

b. 2 ***(A1)***

c. $\frac{80}{130} \times \frac{35}{130} \times 130$ **OR** $\frac{80 \times 35}{130}$ ***(M1)***

Note: Award ***(M1)*** for $\frac{80}{130} \times \frac{35}{130} \times 130$ **OR** $\frac{80 \times 35}{130}$ seen. The following ***(A1)*** cannot be awarded without this statement.

$= 21.5384\ldots$ ***(A1)***

$= 21.5$ ***(AG)***

Note: Both an unrounded answer that rounds to the given answer and rounded must be seen for the ***(A1)*** to be awarded. Accept 21.54 or 21.53 as an unrounded answer.

d. (i) χ^2 statistic $= 10.3$ $(10.3257\ldots)$ ***(G2)***

Note: Accept 10 as a correct 2 significant figure answer.

(ii) p-value $= 0.00573 \quad (0.00572531\ldots)$ ***(G1)***

e. since p-value $< 0.01,\ \mathrm{H}_0$ should not be accepted ***(R1)(A1)*****(ft)**

OR

since χ^2 statistic $> \chi^2$ critical value, H_0 should not be accepted ***(R1)(A1)*****(ft)**

Note: Do not award ***(R0)(A1)***. Follow through from their answer to part (d). Award ***(R0)(A0)*** if part (d) is unanswered.

Award ***(R1)*** for a correct comparison of either their p-value to the test level or their χ^2 statistic to the χ^2 critical value, award ***(A1)*** for the correct result from that comparison.

Examiners report

a. The great majority of candidates found this question to be a good start to the paper, with many perfect scores accruing. A common problem was the inability to form consistent null and alternative hypotheses. Also, calculating the expected value “by hand” as part of a “show that” question was left blank by a number of candidates; to reiterate again – to attain full marks, both the unrounded and the consistent and correctly rounded answer must be stated.

And, lastly, incorrect comparison of statistics when forming a conclusion was a common fault.

b. The great majority of candidates found this question to be a good start to the paper, with many perfect scores accruing. A common problem was the inability to form consistent null and alternative hypotheses. Also, calculating the expected value “by hand” as part of a “show that” question was left blank by a number of candidates; to reiterate again – to attain full marks, both the unrounded and the consistent and correctly rounded answer must be stated.

And, lastly, incorrect comparison of statistics when forming a conclusion was a common fault.

c. The great majority of candidates found this question to be a good start to the paper, with many perfect scores accruing. A common problem was the inability to form consistent null and alternative hypotheses. Also, calculating the expected value “by hand” as part of a “show that” question was left blank by a number of candidates; to reiterate again – to attain full marks, both the unrounded and the consistent and correctly rounded answer must be stated.

And, lastly, incorrect comparison of statistics when forming a conclusion was a common fault.

d. The great majority of candidates found this question to be a good start to the paper, with many perfect scores accruing. A common problem was the inability to form consistent null and alternative hypotheses. Also, calculating the expected value “by hand” as part of a “show that” question was left blank by a number of candidates; to reiterate again – to attain full marks, both the unrounded and the consistent and correctly rounded answer must be stated.

And, lastly, incorrect comparison of statistics when forming a conclusion was a common fault.

e. The great majority of candidates found this question to be a good start to the paper, with many perfect scores accruing. A common problem was the inability to form consistent null and alternative hypotheses. Also, calculating the expected value “by hand” as part of a “show that” question was left blank by a number of candidates; to reiterate again – to attain full marks, both the unrounded and the consistent and correctly rounded

answer must be stated.

And, lastly, incorrect comparison of statistics when forming a conclusion was a common fault.

The seniors from Gulf High School are required to participate in exactly one after-school sport. Data were gathered from a sample of 120 students regarding their choice of sport. The following data were recorded.

	Sport			
Gender	**Football**	**Tennis**	**Basketball**	**Total**
Male	17	8	10	35
Female	31	17	37	85
Total	48	25	47	120

A χ^2 test was carried out at the 5 % significance level to analyse the relationship between gender and choice of after-school sport.

a. Write down the null hypothesis, H_0, for this test. [1]

b. Find the expected value of female footballers. [2]

c. Write down the number of degrees of freedom. [1]

d. Write down the critical value of χ^2, at the 5 % level of significance. [1]

e. Use your graphic display calculator to determine the χ^2_{calc} value. [2]

f. Determine whether H_0 should be accepted. Justify your answer. [2]

g. One student is chosen at random from the 120 students. [2]

Find the probability that this student

(i) is male;

(ii) plays tennis.

h. Two students are chosen at random from the 120 students. [5]

Find the probability that

(i) both play football;

(ii) neither play basketball.

Markscheme

a. H_0 : Gender and choice of afterschool sport are independent. ***(A1)***

Note: Accept “not associated”, do not accept “not related”, “not correlated”, or “not linked”. Accept “the relation between gender and sport is independent”.

[1 mark]

b. $\frac{85}{120} \times \frac{48}{120} \times 120 \left(\frac{85 \times 48}{120}\right)$ ***(M1)***

Note: Award ***(M1)*** for correct expression.

= 34 ***(A1)(G2)***

[2 marks]

c. 2 ***(A1)***

[1 mark]

d. 5.99 (5.991) ***(A1)*****(ft)**

Note: Follow through from part (c).

[1 mark]

e. 2.42 (2.42094…) ***(G2)***

[2 marks]

f. Since 2.42 < 5.99 therefore accept (do not reject) H_0 ***(R1)(A1)*****(ft)**

Note: The numerical values need not be seen, but must be consistent with their parts (d) and (e).

OR

p-value 0.298 > 0.05 therefore accept (do not reject) H_0 ***(R1)(A1)***

Note: *p-value* comparison may **not** be used as part of a follow through solution. Do not award ***(A1)(R0)***. Follow through from parts (c), (d) and (e).

[2 marks]

g. (i) $\frac{35}{120}\left(\frac{7}{24},\ 0.292,\ 29.2\%\right)$ (0.291666...) ***(A1)***

(ii) $\frac{25}{120}\left(\frac{5}{24},\ 0.208,\ 20.8\%\right)$ (0.208333...) ***(A1)***

[2 marks]

h. (i) $\frac{48}{120} \times \frac{47}{119}$ ***(A1)(M1)***

Note: Award ***(A1)*** for two correct fractions, ***(M1)*** for multiplying their two fractions.

$= \frac{94}{595}(0.158,\ 15.8\%)$ (0.157983...) ***(A1)(G2)***

(ii) $\frac{73}{120} \times \frac{72}{119}$ ***(M1)***

Note: Award ***(M1)*** for multiplying correct fractions. If sampling with replacement has been used in both parts (h)(i) and (h)(ii) do not penalise in part (h)(ii). Award a maximum of ***(M1)(A1)*****(ft)**.

$= \frac{219}{595}(0.368,\ 36.8\%)$ (0.368067...) ***(A1)(G2)***

[5 marks]

Examiners report

a. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

b. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

c. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

d. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

e. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

f. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

g. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

h. This question was successfully attempted by the great majority. However, the test is for the mathematical independence of the two variables; it does not address "correlation" or whether there is "no relation" between them. Further, the result of the test should be determined by the comparison of the **numerical values** of either the chi-squared calculated and critical values or the associated *p*-value and the significance level of the test. The creeping use of *k* as the critical value is the notation used in one text book; it is **not** standard notation and its use is not accepted. Comments were made on the G2 forms as to whether the the null hypothesis should be "accepted" or not rejected; both forms are acceptable.

In the compound probability questions, the lack of an explicit tree diagram determined that many candidates were not able to proceed. Determining an appropriate technique is a skill that should be taught.

Daniel grows apples and chooses at random a sample of 100 apples from his harvest.

He measures the diameters of the apples to the nearest cm. The following table shows the distribution of the diameters.

Diameter (to the nearest cm)	5	6	7	8	9
Frequency	15	27	33	17	8

a. Using your graphic display calculator, write down the value of [3]

(i) the mean of the diameters in this sample;

(ii) the standard deviation of the diameters in this sample.

b. Daniel assumes that the diameters of all of the apples from his harvest are normally distributed with a mean of 7 cm and a standard deviation of [3]

1.2 cm. He classifies the apples according to their diameters as shown in the following table.

Classification	Diameter (cm)
Small	Diameter < 6.5
Medium	6.5 ≤ Diameter < a
Large	Diameter ≥ a

Calculate the percentage of **small** apples in Daniel's harvest.

c. Daniel assumes that the diameters of all of the apples from his harvest are normally distributed with a mean of 7 cm and a standard deviation of 1.2 cm. He classifies the apples according to their diameters as shown in the following table. [2]

Classification	Diameter (cm)
Small	Diameter < 6.5
Medium	6.5 ≤ Diameter < a
Large	Diameter ≥ a

Of the apples harvested, 5% are **large** apples.

Find the value of a.

d. Daniel assumes that the diameters of all of the apples from his harvest are normally distributed with a mean of 7 cm and a standard deviation of 1.2 cm. He classifies the apples according to their diameters as shown in the following table. [2]

Classification	Diameter (cm)
Small	Diameter < 6.5
Medium	6.5 ≤ Diameter < a
Large	Diameter ≥ a

Find the percentage of **medium** apples.

e. Daniel assumes that the diameters of all of the apples from his harvest are normally distributed with a mean of 7 cm and a standard deviation of 1.2 cm. He classifies the apples according to their diameters as shown in the following table. [2]

Classification	Diameter (cm)
Small	Diameter < 6.5
Medium	6.5 ≤ Diameter < a
Large	Diameter ≥ a

This year, Daniel estimates that he will grow $100\,000$ apples.

Estimate the number of **large** apples that Daniel will grow this year.

Markscheme

a. (i) $6.76\ (\text{cm})$ ***(G2)***

Notes: Award ***(M1)*** for an attempt to use the formula for the mean with a least two rows from the table.

(ii) $1.14\ (\text{cm})\quad (1.14122\ldots\ (\text{cm}))$ ***(G1)***

b. $\text{P}(\text{diameter} < 6.5) = 0.338\quad (0.338461)$ ***(M1)(A1)***

Notes: Award ***(M1)*** for attempting to use the normal distribution to find the probability **or** for correct region indicated on labelled diagram. Award ***(A1)*** for correct probability.

$33.8(\%)$ ***(A1)*(ft)*(G3)***

Notes: Award ***(A1)*(ft)** for converting their probability into a percentage.

c. $\text{P}(\text{diameter} \geqslant a) = 0.05$ ***(M1)***

Note: Award ***(M1)*** for attempting to use the normal distribution to find the probability **or** for correct region indicated on labelled diagram.

$a = 8.97\ (\text{cm})\quad (8.97382\ldots)$ ***(A1)(G2)***

d. $100 - (5 + 33.8461\ldots)$ ***(M1)***

Note: Award ***(M1)*** for subtracting "$5+$ their part (b)" from 100 **or** ***(M1)*** for attempting to use the normal distribution to find the probability $\text{P}\left(6.5 \leqslant \text{diameter} < \text{their part (c)}\right)$ **or** for correct region indicated on labelled diagram.

$= 61.2(\%)\quad (61.1538\ldots(\%))$ ***(A1)*(ft)*(G2)***

Notes: Follow through from their answer to part (b). Percentage symbol is not required. Accept $61.1(\%)$ $(61.1209\ldots(\%))$ if 8.97 used.

e. $100\,000 \times 0.05$ ***(M1)***

Note: Award ***(M1)*** for multiplying by 0.05 (or 5%).

$= 5000$ ***(A1)(G2)***

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]

An agricultural cooperative uses three brands of fertilizer, A, B and C, on 120 different crops. The crop yields are classified as High, Medium or Low.

The data collected are organized in the table below.

	Fertilizer			Total
	A	B	C	
High Yield	10	8	12	**30**
Medium Yield	24	14	12	**50**
Low Yield	16	8	16	**40**
Total	**50**	**30**	**40**	**120**

The agricultural cooperative decides to conduct a chi-squared test at the 1 % significance level using the data.

a. State the null hypothesis, H_0, for the test. [2]

b. Write down the number of degrees of freedom. [1]

c. Write down the critical value for the test. [1]

d. Show that the expected number of Medium Yield crops using Fertilizer C is 17, correct to the nearest integer. [2]

e. Use your graphic display calculator to find for the data [3]

(i) the χ^2 calculated value, χ^2_{calc};

(ii) the *p*-value.

f. State the conclusion of the test. Give a reason for your decision. [2]

Markscheme

a. The (crop) yield is independent of the (type of) fertilizer used. ***(A1)(A1)***

Note: Award ***(A1)*** for (crop) yield and (type of) fertilizer, ***(A1)*** for "independent" or "not dependent" or "not associated".

Do not accept "not correlated" or "not related" or "not connected" or "does not depend on".

b. 4 ***(A1)***

c. 13.277 ***(A1)*(ft)**

Note: Accept 13.3. Follow through from part (b).

d. $\frac{50}{120} \times \frac{40}{120} \times 120$ **or** $\frac{50 \times 40}{120}$ ***(M1)***

Note: Award ***(M1)*** for correct substitution in the expected value formula.

= 16.6666... ***(A1)***
= 17 ***(AG)***

Note: Both unrounded and rounded answers must be seen to award ***(A1)***.

e. (i) $\chi^2_{calc} = 3.86 (3.86133...)$ ***(G2)***

(ii) *p*-value $= 0.425\ (0.425097...)$ ***(G1)***

f. Since χ^2_{calc} < Critical Value ***(R1)***

Accept (do not reject) the Null Hypothesis. ***(A1)*****(ft)**

Note: Accept decision based on *p*-value with comparison to 1 % (0.425097... > 0.01) . Do not award ***(R0)(A1)***. Follow through from parts (c) and (e). Numerical answers must be present in the question for a valid comparison to be made.

Examiners report

a. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

b. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

c. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

d. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

e. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

f. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect terminology in the null hypothesis, (2) use of the 5% level, (3) an inability to find the expected value by hand, (4) comparison of incorrect values. Note, candidates will never be asked to calculate the chi-squared statistic other than from the GDC.

The number of bottles of water sold at a railway station on each day is given in the following table.

Day	0	1	2	3	4	5	6	7	8	9	10	11	12
Temperature ($T°$)	21	20.7	20	19	18	17.3	17	17.3	18	19	20	20.7	21
Number of bottles sold (n)	150	141	126	125	98	101	93	99	116	121	119	134	141

a. Write down [2]

(i) the mean temperature;

(ii) the standard deviation of the temperatures.

b. Write down the correlation coefficient, r, for the variables n and T. [1]

c. Comment on your value for r. [2]

d. The equation of the line of regression for n on T is $n = dT - 100$. [2]

(i) Write down the value of d.

(ii) Estimate how many bottles of water will be sold when the temperature is 19.6°.

e. On a day when the temperature was 36° Peter calculates that 314 bottles would be sold. Give one reason why his answer might be unreliable. [1]

Markscheme

a. (i) 19.2 ***(G1)***

(ii) 1.45 ***(G1)***

[2 marks]

b. $r = 0.942$ ***(G1)***

[1 mark]

c. Strong, positive correlation. ***(A1)*****(ft)*****(A1)*****(ft)**

[2 marks]

d. (i) $d = 11.5$ ***(G1)***

(ii) $n = 11.5 \times 19.6 - 100$

$= 125$ (*accept* 126) ***(A1)*****(ft)**

Note: Answer must be a whole number.

[2 marks]

e. It is unreliable to extrapolate outside the values given (outlier). ***(R1)***

[1 mark]

Examiners report

a. (i) Generally well done but many lost an AP here

(ii) Only correct if the candidate knew how to use their GDC and even then several gave the wrong standard deviation.

b. Again, only correct if the candidate could use their GDC. Many answers given were greater than 1 and the candidates did not see anything wrong with this.

c. Many received a ft mark for this part. The word “positive” was often omitted.

d. (i) Most candidates substituted the first set of points into the equation instead of finding the regression line on their GDC.

(ii) Most managed to score a ft point here. But some did not give their answer as a whole number.

e. Not many candidates mentioned the idea of an outlier. Most came up with some creative reason, albeit wrong, as to why the answer might be unreliable. Some of them made interesting reading.

---

a. A speed camera on Peterson Road records the speed of each passing vehicle. The speeds are found to be normally distributed with a mean of $67\,\text{km h}^{-1}$ and a standard deviation of $3.4\,\text{km h}^{-1}$. [2]

Sketch a diagram of this normal distribution and shade the region representing the probability that the speed of a vehicle is between 60 and $70\,\text{km h}^{-1}$.

b. A vehicle on Peterson Road is chosen at random. [3]

Find the probability that the speed of this vehicle is

(i) more than $60\,\text{km h}^{-1}$;

(ii) less than $70\,\text{km h}^{-1}$;

(iii) between 60 and $70\,\text{km h}^{-1}$.

c. It is found that $19\,\%$ of the vehicles are exceeding the speed limit of $s\,\text{km h}^{-1}$. [2]

Find the value of s , correct to the nearest integer.

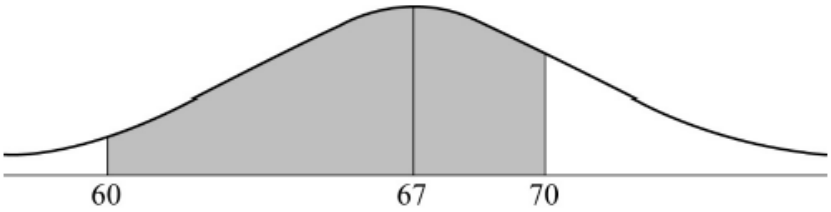
d. There is a fine of $\text{US\$65}$ for exceeding the speed limit on Peterson Road. On a particular day the total value of fines issued was $\text{US\$14 820}$. [4]

(i) Calculate the number of fines that were issued on this day.

(ii) Estimate the total number of vehicles that passed the speed camera on Peterson Road on this day.

Markscheme

a.



(A1)(A1)

Note: Award ***(A1)*** for normal curve with mean of 67 indicated or two vertical lines drawn approximately in correct place. Award ***(A1)*** for correct shaded region (between the vertical lines.).

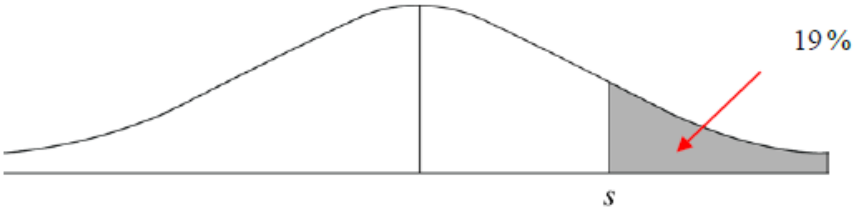
b. (i) $0.980\ \ (0.980244...,\ 98.0\,\%)$ ***(G1)***

(ii) $0.811\ \ (0.811207...,\ 81.1\,\%)$ ***(G1)***

(iii) $0.791\ \ (0.791451...,\ 79.1\,\%)$ ***(G1)***

c. $\text{P}\ (S > s) = 19\,\%\ (0.19)$ **OR** $\text{P}\ (S > s) = 81\,\%\ (0.81)$ **(M1)**

OR



(M1)

Note: Award ***(M1)*** for the correct probability equation **OR** for a correct region indicated on labelled diagram.

$(s =)\ 70.0\ \ (69.9848...)$ ***(A1)(G2)***

Note: Award ***(M1)*** for any correct method.

d. (i) $\frac{14\,820}{65}$ ***(M1)***

Note: Award ***(M1)*** for dividing $14\,820$ by 65.

$= 228$ ***(A1)(G2)***

(ii) $\frac{\text{their } 228}{0.19}$ (or equivalent) ***(M1)***

$= 1200$ (vehicles) ***(A1)*****(ft)*****(G2)***

Note: Award ***(M1)*** for correct method. Follow through from their part (d)(i).

Examiners report

a. Question 3: The normal distribution

Candidates showed comprehensive understanding of the normal distribution. The graphic display calculator was used efficiently by most of the candidates. There was much variability in the ability to sketch the curve in part (a). Instead of drawing the straight-forward sketch with the mean line and two vertical lines as required at 60 and 70, many linked it to standard deviations. It was very rare to see any method in part (c). Most candidates managed part (d)(i) but few went on to complete part (d)(ii).

b. Question 3: The normal distribution

Candidates showed comprehensive understanding of the normal distribution. The graphic display calculator was used efficiently by most of the candidates. There was much variability in the ability to sketch the curve in part (a). Instead of drawing the straight-forward sketch with the mean line and two vertical lines as required at 60 and 70, many linked it to standard deviations. It was very rare to see any method in part (c). Most candidates managed part (d)(i) but few went on to complete part (d)(ii).

c. Question 3: The normal distribution

Candidates showed comprehensive understanding of the normal distribution. The graphic display calculator was used efficiently by most of the candidates. There was much variability in the ability to sketch the curve in part (a). Instead of drawing the straight-forward sketch with the mean line and two vertical lines as required at 60 and 70, many linked it to standard deviations. It was very rare to see any method in part (c). Most candidates managed part (d)(i) but few went on to complete part (d)(ii).

d. Question 3: The normal distribution

Candidates showed comprehensive understanding of the normal distribution. The graphic display calculator was used efficiently by most of the candidates. There was much variability in the ability to sketch the curve in part (a). Instead of drawing the straight-forward sketch with the mean

line and two vertical lines as required at 60 and 70, many linked it to standard deviations. It was very rare to see any method in part (c). Most candidates managed part (d)(i) but few went on to complete part (d)(ii).

One day the numbers of customers at three cafés, "Alan's Diner" (A), "Sarah's Snackbar" (S) and "Pete's Eats" (P), were recorded and are given below.

17 were customers of Pete's Eats only
27 were customers of Sarah's Snackbar only
15 were customers of Alan's Diner only
10 were customers of Pete's Eats **and** Sarah's Snackbar **but not** Alan's Diner
8 were customers of Pete's Eats **and** Alan's Diner **but not** Sarah's Snackbar

Some of the customers in each café were given survey forms to complete to find out if they were satisfied with the standard of service they received.

	Pete's Eats	Alan's Diner	Sarah's Snackbar	Total
Dissatisfied	16	8	16	40
Satisfied	26	20	34	80
Total	42	28	50	120

A.a Draw a Venn Diagram, using sets labelled A , S and P , that shows this information. [3]

A.b There were 48 customers of Pete's Eats that day. Calculate the number of people who were customers of all three cafés. [2]

A.c There were 50 customers of Sarah's Snackbar that day. Calculate the total number of people who were customers of Alan's Diner. [3]

A.d Write down the number of customers of Alan's Diner that were also customers of Pete's Eats. [1]

A.e Find $n[(S \cup P) \cap A']$. [2]

B.a One of the survey forms was chosen at random, find the probability that the form showed "Dissatisfied"; [2]

B.b One of the survey forms was chosen at random, find the probability that the form showed "Satisfied" and was completed at Sarah's Snackbar; [2]

B.c One of the survey forms was chosen at random, find the probability that the form showed "Dissatisfied", given that it was completed at Alan's Diner. [2]

B.d A χ^2 test at the 5% significance level was carried out to determine whether there was any difference in the level of customer satisfaction in each of the cafés. [1]

Write down the null hypothesis, H_0 , for the χ^2 test.

B.e A χ^2 test at the 5% significance level was carried out to determine whether there was any difference in the level of customer satisfaction in each of the cafés. [1]

Write down the number of degrees of freedom for the test.

B.f A χ^2 test at the 5% significance level was carried out to determine whether there was any difference in the level of customer satisfaction in each of the cafés. [2]

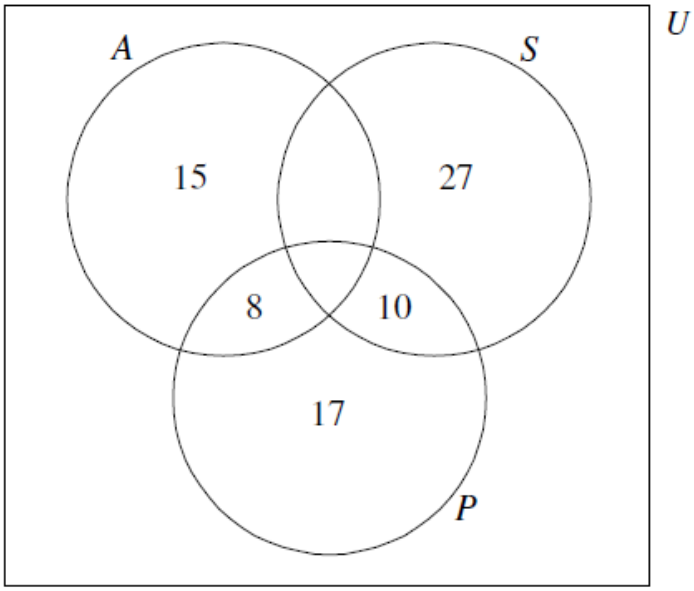
Using your graphic display calculator, find ${\chi^2}_{calc}$.

B.g A χ^2 test at the 5% significance level was carried out to determine whether there was any difference in the level of customer satisfaction in each of the cafés. [2]

State, giving a reason, the conclusion to the test.

Markscheme

A.a.



(A1) for rectangle and three labelled intersecting circles
(A1) for 15, 27 and 17
(A1) for 10 and 8 ***(A3)***

[3 marks]

A.b $48 - (8 + 10 + 17)$ *or equivalent* ***(M1)***

$= 13$ ***(A1)*****(ft)*****(G2)***

[2 marks]

A.c $50 - (27 + 10 + 13)$ ***(M1)***

Note: Award ***(M1)*** for working seen.

$= 0$ ***(A1)***
number of elements in A $= 36$ ***(A1)*****(ft)*****(G3)***

Note: Follow through from (b).

[3 marks]

A.d 21 ***(A1)*****(ft)**

Note: Follow through from (b) even if no working seen.

[1 mark]

A.e 54 ***(M1)(A1)*****(ft)*****(G2)***

Note: Award ***(M1)*** for $17, 10, 27$ seen. Follow through from (a).

[2 marks]

B.a $\frac{40}{120}$ $\left(\frac{1}{3}, 0.333, 33.3\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator.

[2 marks]

B.b $\frac{34}{120}$ $\left(\frac{17}{60}, 0.283, 28.3\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator.

[2 marks]

B.c $\frac{8}{28}$ $\left(\frac{2}{7}, 0.286, 28.6\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator.

[2 marks]

B.d customer satisfaction is **independent** of café ***(A1)***

Note: Accept “customer satisfaction is **not associated with** the café”.

[1 mark]

B.e 2 ***(A1)***

[1 mark]

B.f 0.754 ***(G2)***

Note: Award ***(G1)(G1)(AP)*** for 0.75 or for correct answer incorrectly rounded to 3 s.f. or more, ***(G0)*** for 0.7.

[2 marks]

B.g since \({\chi ^2}_{calc} < {\chi ^2}_{crit}5.991 accept (or Do not reject) H_0 ***(R1)(A1)*****(ft)**

Note: Follow through from their value in (e).

OR

Accept (or Do not reject) H_0 as p-value $(0.686) > 0.05$ ***(R1)(A1)*****(ft)**

Notes: Do not award ***(A1)(R0)***. Award the ***(R1)*** for comparison of appropriate values.

[2 marks]

Examiners report

A.a **Part A**

This part was successfully attempted by the great majority. A common mistake was the failure to intersect all three sets.

A.b **Part A**

This part was successfully attempted by the great majority. A common mistake was the failure to intersect all three sets.

A.c **Part A**

This part was successfully attempted by the great majority. A common mistake was the failure to intersect all three sets.

A.d **Part A**

This part was successfully attempted by the great majority. A common mistake was the failure to intersect all three sets.

A.e **Part A**

This part was successfully attempted by the great majority. A common mistake was the failure to intersect all three sets.

A surprising number seemed unfamiliar with set notation in (e) and thus did not attempt this part.

B.a **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

B.b **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

B.c **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

B.d **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

The chi-squared test was well done by the great majority, however, it was clear that a number of centres do not teach this subject, since there were a number of scripts which either were left blank or showed no understanding in the responses seen.

B.e **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

The chi-squared test was well done by the great majority, however, it was clear that a number of centres do not teach this subject, since there were a number of scripts which either were left blank or showed no understanding in the responses seen.

B.f. **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

The chi-squared test was well done by the great majority, however, it was clear that a number of centres do not teach this subject, since there were a number of scripts which either were left blank or showed no understanding in the responses seen.

B.g **Part B**

The work on probability also proved accessible to the great majority with a large number of candidates attaining full marks. Most errors occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table.

The chi-squared test was well done by the great majority, however, it was clear that a number of centres do not teach this subject, since there were a number of scripts which either were left blank or showed no understanding in the responses seen.

The Brahma chicken produces eggs with weights in grams that are normally distributed about a mean of 55 g with a standard deviation of 7 g. The eggs are classified as small, medium, large or extra large according to their weight, as shown in the table below.

Size	Weight (g)
Small	Weight < 53
Medium	$53 \leq$ Weight < 63
Large	$63 \leq$ Weight < 73
Extra Large	Weight ≥ 73

a. Sketch a diagram of the distribution of the weight of Brahma chicken eggs. On your diagram, show clearly the boundaries for the classification of the eggs. [3]

b. An egg is chosen at random. Find the probability that the egg is [4]

(i) medium;

(ii) extra large.

c. There is a probability of 0.3 that a randomly chosen egg weighs more than w grams. [2]

Find w .

d. The probability that a Brahma chicken produces a large size egg is 0.121. Frank's Brahma chickens produce 2000 eggs each month. [2]

Calculate an estimate of the number of large size eggs produced by Frank's chickens each month.

e. The selling price, in US dollars (USD), of each size is shown in the table below. [3]

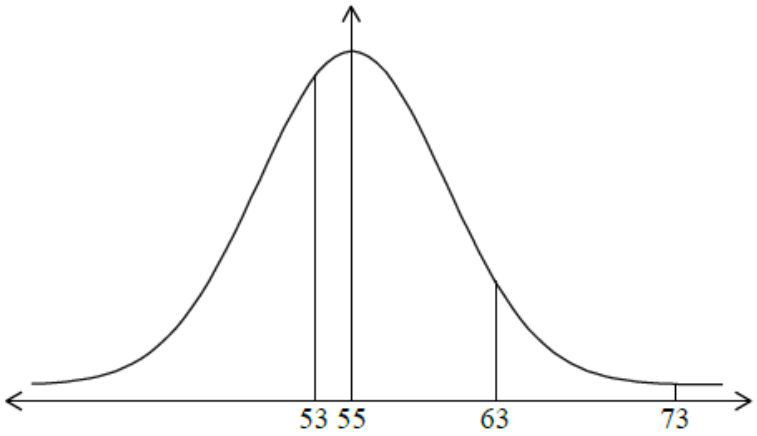
Size	Selling price (USD)
Small	0.30
Medium	0.50
Large	0.65
Extra Large	0.80

The probability that a Brahma chicken produces a small size egg is 0.388.

Estimate the monthly income, in USD, earned by selling the 2000 eggs. Give your answer correct to two decimal places.

Markscheme

a.



(A1) for normal curve with mean of 55 indicated
(A1) for three lines in approximately the correct position
(A1) for labels on the three lines ***(A1)(A1)(A1)***

b. (i) $\mathrm{P}(53 \leqslant \text{Weight} < 63) = 0.486\ (0.485902\ldots)$ ***(M1)(A1)(G2)***

Note: Award ***(M1)*** for correct region indicated on labelled diagram.

(ii) $\mathrm{P}(\text{Weight} > 73) = 0.00506\ (0.00506402)$ ***(M1)(A1)(G2)***

Note: Award ***(M1)*** for correct region indicated on labelled diagram.

c. $\mathrm{P}(\text{Weight} > w) = 0.3$ ***(M1)***

$w = 58.7\ (58.6708\ldots)$ ***(A1)(G2)***

Note: Award ***(M1)*** for correct region indicated on labelled diagram.

d. Expected number of large size eggs

$= 2000(0.121)$ ***(M1)***
$= 242$ ***(A1)(G2)***

e. Expected income

$= 2000 \times 0.30 \times 0.388 + 2000 \times 0.50 \times 0.486 + 2000 \times 0.65 \times 0.121 + 2000 \times 0.80 \times 0.00506$ ***(M1)(M1)***

Note: Award ***(M1)*** for their correct products, ***(M1)*** for addition of 4 terms.

$= 884.20\ \text{USD}$ ***(A1)*****(ft)*****(G3)***

Note: Follow through from part (b).

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]

Alex and Kris are riding their bicycles together along a bicycle trail and note the following distance markers at the given times.

Time (t hours)	1	2	3	4	5	6	7
Distance (d km)	57	65	72	81	89	97	107

a. Draw a scatter diagram of the data. Use 1 cm to represent 1 hour and 1 cm to represent 10 km. [3]

b.i. Write down for this set of data the mean time, $\bar{t}$. [1]

b.ii. Write down for this set of data the mean distance, $\bar{d}$. [1]

c. Mark and label the point $M(\bar{t},\ \bar{d})$ on your scatter diagram. [2]

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

e. **Using your graph**, estimate the time when Alex and Kris pass the 85 km distance marker. Give your answer correct to **one decimal place**. [2]

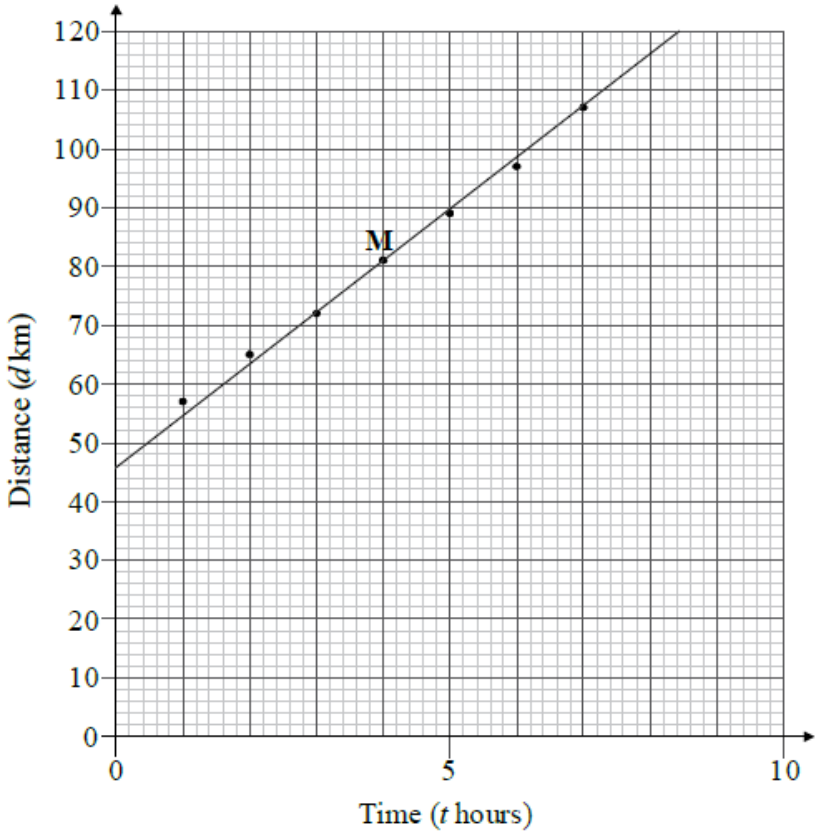
f. Write down the equation of the regression line for the data given. [2]

g.i. **Using your equation** calculate the distance marker passed by the cyclists at 10.3 hours. [2]

g.ii. Is this estimate of the distance reliable? Give a reason for your answer. [2]

Markscheme

a.



(A1)(A2)

Notes: Award ***(A1)*** for axes labelled with d and t and correct scale, ***(A2)*** for 6 or 7 points correctly plotted, ***(A1)*** for 4 or 5 points, ***(A0)*** for 3 or less points correctly plotted. Award at most ***(A1)(A1)*** if points are joined up. If axes are reversed award at most ***(A0)(A2)***

[3 marks]

b.i. $\bar{t} = 4$ ***(G1)***

[1 mark]

b.ii.

$\bar{d} = 81.1\left(\frac{568}{7}\right)$ ***(G1)***

Note: If answers are the wrong way around award in (i) ***(G0)*** and in (ii) ***(G1)*****(ft)**.

[1 mark]

c. Point marked and labelled with M or $\bar{t}$, $\bar{d}$ on their graph ***(A1)*****(ft)*****(A1)*****(ft)**

[2 marks]

d. Line of best fit drawn that passes through their M and (0, 48) ***(A1)*****(ft)*****(A1)*****(ft)**

Notes: Award ***(A1)*****(ft)** for straight line that passes through their M, ***(A1)*** for line (extrapolated if necessary) that passes through (0, 48).

Accept error of ±3. If ruler not used award a maximum of ***(A1)*****(ft)*****(A0)***.

[2 marks]

e. 4.5h (their answer ±0.2) ***(M1)(A1)*****(ft)*****(G2)***

Note: Follow through from their graph. If method shown by some indication on graph of point but answer is incorrect, award ***(M1)(A0)***.

[2 marks]

f. $d = 8.25t + 48.1$ ***(G1)(G1)***

Notes: Award ***(G1)*** for 8.25, ***(G1)*** for 48.1.

Award at most ***(G1)(G0)*** if $d =$ (*or* $y =$) is not seen.

Accept $d - 81.1 = 8.25(t - 4)$ or equivalent.

[2 marks]

g.i. $d = 8.25 \times 10.3 + 48.1$ ***(M1)***

$d = 133$ km ***(A1)*****(ft)*****(G2)***

[2 marks]

g.ii. No ***(A1)***

Outside the set of values of t or equivalent. ***(R1)***

Note: Do not award ***(A1)(R0)***.

[2 marks]

Examiners report

a. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

b.i. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

b.ii. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

c. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

d. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

e. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

f. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

g.i. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

g.ii. This question was well answered by most of the candidates. Diagrams were in general well drawn except for some students that reversed the axes or did not use the stated scales. They were able to use the GDC to find the means and the equation of the regression line. Very few students could take the correct decision in (g) (ii) by stating that the value was outside the range of the data set. The majority inclined their answers towards the context of the question and forgot what they had been taught about how wrong extrapolation can be.

A store recorded their sales of televisions during the 2010 football World Cup. They looked at the numbers of televisions bought by gender and the size of the television screens.

This information is shown in the table below; S represents the size of the television screen in inches.

	$S \leq 22$	$22 < S \leq 32$	$32 < S \leq 46$	$S > 46$	Total
Female	65	100	40	15	220
Male	20	65	140	55	280
Total	85	165	180	70	500

The store wants to use this information to predict the probability of selling these sizes of televisions for the 2014 football World Cup.

a. Use the table to find the probability that [6]

(i) a television will be bought by a female;

(ii) a television with a screen size of $32 < S \leq 46$ will be bought;

(iii) a television with a screen size of $32 < S \leq 46$ will be bought by a female;

(iv) a television with a screen size greater than 46 inches will be bought, given that it is bought by a male.

b. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % [1]
significance level.

Write down the null hypothesis.

c. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % [2]
significance level.

Show that the expected frequency for females who bought a screen size of $32 < S \leq 46$, is 79, correct to the nearest integer.

d. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % [1]
significance level.

Write down the number of degrees of freedom.

e. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % significance level. [2]

Write down the χ^2 calculated value.

f. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % significance level. [1]

Write down the critical value for this test.

g. The manager of the store wants to determine whether the screen size is independent of gender. A Chi-squared test is performed at the 1 % significance level. [2]

Determine if the null hypothesis should be accepted. Give a reason for your answer.

Markscheme

a. (i) $\frac{220}{500}\left(\frac{11}{25},\ 0.44, 44\%\right)$ ***(A1)(G1)***

(ii) $\frac{180}{500}\left(\frac{9}{25},\ 0.36, 36\%\right)$ ***(A1)(G1)***

(iii) $\frac{40}{500}\left(\frac{2}{25},\ 0.08, 8\%\right)$ ***(A1)(A1)(G2)***

(iv) $\frac{55}{500}\left(\frac{11}{56},\ 0.196, 19.6\%\right)$ ***(A1)(A1)(G2)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator. Award ***(A0)(A0)*** if answers are given as incorrect reduced fractions without working.

[6 marks]

b. "The size of the television screen is independent of gender." ***(A1)***

Note: Accept "not associated", do not accept "not correlated".

[1 mark]

c. $\frac{180}{500} \times \frac{220}{500} \times 500$ **OR** $\frac{180 \times 220}{500}$ ***(M1)***

= 79.2 ***(A1)***

= 79 ***(AG)***

Note: Both the unrounded and the given answer must be seen for the final ***(A1)*** to be awarded.

[2 marks]

d. 3 ***(A1)***

[1 mark]

e. χ^2_{calc} = 104(103.957...) ***(G2)***

Note: Award ***(M1)*** if an attempt at using the formula is seen but incorrect answer obtained.

[2 marks]

f. 11.345 ***(A1)*****(ft)**

Notes: Follow through from their degrees of freedom.

[1 mark]

g. $\chi^2_{calc} > \chi^2_{crit}$ **OR** $p < 0.01$ ***(R1)***

Do not accept H_0. ***(A1)*****(ft)**

Note: Do not award ***(R0)(A1)*****(ft)**. Follow through from their parts (d), (e) and (f).

[2 marks]

Examiners report

a. Part (a) was generally well answered by most of the students, except for part (a)(iv) which called for conditional probability.

b. Most students correctly stated the null hypothesis in part (b), and answered parts (d), (e), (f) and (g).

c. In some responses to part (c) it seemed that the difference between calculation of the expected value and showing that the value is 79 was not clear to the candidates. It is important that teachers explain to their students that in a ***"show that"*** *question* they are expected to demonstrate the mathematical reasoning through which the given answer is obtained.

d. Most students correctly stated the null hypothesis in part (b), and answered parts (d), (e), (f) and (g).

e. Most students correctly stated the null hypothesis in part (b), and answered parts (d), (e), (f) and (g).

f. Most students correctly stated the null hypothesis in part (b), and answered parts (d), (e), (f) and (g).

g. Most students correctly stated the null hypothesis in part (b), and answered parts (d), (e), (f) and (g).

Pam has collected data from a group of 400 IB Diploma students about the Mathematics course they studied and the language in which they were examined (English, Spanish or French). The summary of her data is given below.

	Mathematics HL	Mathematics SL	Mathematical Studies SL	**Total**
English	50	70	80	200
Spanish	30	50	30	110
French	20	30	40	90
Total	100	150	150	400

a. A student is chosen at random from the group. Find the probability that the student [8]

(i) studied Mathematics HL;

(ii) was examined in French;

(iii) studied Mathematics HL and was examined in French;

(iv) did not study Mathematics SL and was not examined in English;

(v) studied Mathematical Studies SL given that the student was examined in Spanish.

b. Pam believes that the Mathematics course a student chooses is independent of the language in which the student is examined. [2]

Using your answers to parts (a) (i), (ii) and (iii) above, state whether there is any evidence for Pam's belief. Give a reason for your answer.

c. Pam decides to test her belief using a Chi-squared test at the 5% level of significance. [3]

(i) State the null hypothesis for this test.

(ii) Show that the expected number of Mathematical Studies SL students who took the examination in Spanish is 41.3, correct to 3 significant figures.

d. Write down [4]

(i) the Chi-squared calculated value;

(ii) the number of degrees of freedom;

(iii) the Chi-squared critical value.

e. State, giving a reason, whether there is sufficient evidence at the 5% level of significance that Pam's belief is correct. [2]

Markscheme

a. (i) $\frac{100}{400}$ $\left(\frac{1}{4}, 0.25, 25\%\right)$ ***(A1)***

(ii) $\frac{90}{400}$ $\left(\frac{9}{40}, 0.225, 22.5\%\right)$ ***(A1)***

(iii) $\frac{20}{400}$ $\left(\frac{1}{20}, 0.05, 5\%\right)$ ***(A1)(A1)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator.

(iv) $\frac{120}{400}$ $\left(\frac{3}{10}, 0.3, 30\%\right)$ ***(A1)(A1)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator.

(v) $\frac{30}{110}$ $\left(\frac{3}{11}, 0.273, 27.3\%\right)$ $(0.272727\ldots)$ ***(A1)(A1)***

Note: Award ***(A1)*** for numerator, ***(A1)*** for denominator. Accept 0.27, do not accept 0.272, do not accept 0.3.

[8 marks]

b. $\frac{1}{20} \neq \frac{1}{4} \times \frac{9}{40}$ ***(R1)*****(ft)**

Note: The fractions must be used as part of the reason. Follow through from (a)(i), (a)(ii) and (a)(iii).

Pam is not correct. ***(A1)*****(ft)**

Notes: Do not award ***(R0)(A1)***. Accept the events are not independent (dependent).

[2 marks]

c. (i) The mathematics course and language of examination are independent. ***(A1)***

Notes: Accept “There is no association between Mathematics course and language”. Do not accept “not related”, “not correlated”, “not influenced”.

(ii) $\frac{110}{400} \times \frac{150}{400} \times 400 \left(= \frac{110 \times 150}{400}\right)$ ***(M1)***

$= 41.25$ ***(A1)***

$= 41.3$ ***(AG)***

Note: 41.25 **and** 41.3 must be seen to award final ***(A1)***.

[3 marks]

d. (i) 7.67 $(7.67003\ldots)$ ***(G2)***

Note: Accept 7.7, do not accept 8 or 7.6. Award ***(G1)*** if formula with all nine terms seen but their answer is not one of those above.

(ii) 4 ***(G1)***

(iii) 9.488 ***(A1)*****(ft)**

Notes: Accept 9.49 or 9.5, do not accept 9.4 or 9. Follow through from their degrees of freedom.

[4 marks]

e. $7.67 < 9.488$ ***(R1)***

OR

$p = 0.104\ldots, p > 0.05$ ***(R1)***

Accept (Do not reject) H_0 (Pam’s belief is correct) ***(A1)*****(ft)**

Notes: Follow through from part (d). Do not award ***(R0)(A1)***.

[2 marks]

Examiners report

a. The simple probabilities beginning this question were successfully attempted by the great majority. Most errors in the latter parts occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table. Probability questions in this course are, in the main, contextual and the reliance of formulas is not always beneficial to the candidates. Only the best candidates realized the significance of part (b) as a link to the chi-squared test.

This was well attempted by the majority, the weakness being the sole reliance of the calculator to calculate expected value. However, there still remains confusion between critical and *p*-values as the basis for accepting the null hypothesis.

b. The simple probabilities beginning this question were successfully attempted by the great majority. Most errors in the latter parts occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table. Probability questions in this course are, in the main, contextual and the reliance of formulas is not always beneficial to the candidates. Only the best candidates realized the significance of part (b) as a link to the chi-squared test.

   This was well attempted by the majority, the weakness being the sole reliance of the calculator to calculate expected value. However, there still remains confusion between critical and p-values as the basis for accepting the null hypothesis.

c. The simple probabilities beginning this question were successfully attempted by the great majority. Most errors in the latter parts occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table. Probability questions in this course are, in the main, contextual and the reliance of formulas is not always beneficial to the candidates. Only the best candidates realized the significance of part (b) as a link to the chi-squared test.

   This was well attempted by the majority, the weakness being the sole reliance of the calculator to calculate expected value. However, there still remains confusion between critical and p-values as the basis for accepting the null hypothesis.

d. The simple probabilities beginning this question were successfully attempted by the great majority. Most errors in the latter parts occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table. Probability questions in this course are, in the main, contextual and the reliance of formulas is not always beneficial to the candidates. Only the best candidates realized the significance of part (b) as a link to the chi-squared test.

   This was well attempted by the majority, the weakness being the sole reliance of the calculator to calculate expected value. However, there still remains confusion between critical and p-values as the basis for accepting the null hypothesis.

e. The simple probabilities beginning this question were successfully attempted by the great majority. Most errors in the latter parts occurred due to candidates trying to use the algebraic form of laws of probability, rather than by interpreting the contingency table. Probability questions in this course are, in the main, contextual and the reliance of formulas is not always beneficial to the candidates. Only the best candidates realized the significance of part (b) as a link to the chi-squared test.

   This was well attempted by the majority, the weakness being the sole reliance of the calculator to calculate expected value. However, there still remains confusion between critical and p-values as the basis for accepting the null hypothesis.

Jorge conducted a survey of 200 drivers. He asked two questions:

> How long have you had your driving licence?
> Do you wear a seat belt when driving?

The replies are summarized in the table below.

	Wear a seat belt	Do not wear a seat belt
Licence less than 2 years	38	42
Licence between 2 and 15 years	30	45
Licence more than 15 years	30	15

a. Jorge applies a χ^2 test at the 5% level to investigate whether wearing a seat belt is associated with the time a driver has had their licence. [8]

(i) Write down the null hypothesis, H_0.

(ii) Write down the number of degrees of freedom.

(iii) Show that the expected number of drivers that wear a seat belt and have had their driving licence for more than 15 years is 22, correct to the nearest whole number.

(iv) Write down the χ^2 test statistic for this data.

(v) Does Jorge accept H_0 ? Give a reason for your answer.

b. Consider the 200 drivers surveyed. One driver is chosen at random. Calculate the probability that [4]

(i) this driver wears a seat belt;

(ii) the driver does not wear a seat belt, **given that** the driver has held a licence for more than 15 years.

c. Two drivers are chosen at random. Calculate the probability that [6]

(i) both wear a seat belt.

(ii) at least one wears a seat belt.

Markscheme

a. (i) $\mathrm{H}_0 =$ wearing of a seat belt and the time a driver has held a licence are independent. ***(A1)***

Note: For independent accept 'not associated' but do not accept 'not related' or 'not correlated'

(ii) 2 ***(A1)***

(iii) $\frac{98\times45}{200} = 22.05 = 22$ (*correct to the nearest whole number*) ***(M1)(A1)(AG)***

Note: ***(M1)*** for correct formula and ***(A1)*** for correct substitution. Unrounded answer must be seen for the ***(A1)*** to be awarded.

(iv) $\chi^2 = 8.12$ ***(G2)***

Note: For unrounded answer award ***(G1)(G0)(AP)***. If formula used award ***(M1)*** for correct substituted formula with correct substitution (6 terms) ***(A1)*** for correct answer.

(v) "Does not accept H_0" ***(A1)*****(ft)**

$p\text{-}value < 0.05$ ***(R1)*****(ft)**

Note: Allow "Reject H_0" or equivalent. Follow through from their χ^2 statistic. Award ***(R1)*****(ft)** for comparing the appropriate values. The ***(A1)*****(ft)** can be awarded only if the conclusion is valid according to the comparison given. If no reason given or if reason is wrong the two marks are lost.

[8 marks]

b. (i) $\frac{98}{200} (= 0.49, 49\%)$ ***(A1)(A1)(G2)***

Note: ***(A1)*** for numerator, ***(A1)*** for denominator.

(ii) $\frac{15}{45} (= 0.333, 33.3\%)$ ***(A1)(A1)(G2)***

Note: ***(A1)*** for numerator, ***(A1)*** for denominator.

[4 marks]

c. (i) $\frac{98}{200} \times \frac{97}{199} = 0.239\ (23.9\%)$ ***(A1)(M1)(A1)(G3)***

Note: ***(A1)*** for correct probabilities seen, ***(M1)*** for multiplying two probabilities, ***(A1)*** for correct answer.

(ii) $1 - \frac{102}{200} \times \frac{101}{199} = 0.741\ (74.1\%)$ ***(M1)(M1)(A1)*****(ft)*****(G2)***

Note: ***(M1)*** for showing the product, ***(M1)*** for using the probability of the complement, ***(A1)*** for correct answer. Follow through for consistent use of with replacement.

OR

$\frac{98}{200} \times \frac{97}{199} + \frac{98}{200} \times \frac{102}{199} + \frac{102}{200} \times \frac{98}{199} = 0.741\ (74.1\%)$ ***(M1)(M1)(A1)*****(ft)*****(G2)***

Note: ***(M1)*** for adding three products of fractions (or equivalent), ***(M1)*** for using the correct fractions, ***(A1)*** for correct answer. Follow through for consistent use of with replacement.

[6 marks]

Examiners report

a. The first part of the question was relatively well done. The null hypothesis and the degrees of freedom were well answered by the majority of the students. In the show that question some students used the GDC to find the expected values table and highlighted the correct value 22.05. This procedure gained no mark; the expected value formula was expected to be used here. Also those who did use the formula were expected to show the unrounded value 22.05 to gain full marks in this part question. Many lost the answer mark for not doing so. GDC was used by most of the students to find the chi-squared test though some students attempted to find this value by hand which made them waste time. Correct values were compared when deciding whether to accept or not the null hypothesis and follow through marks were awarded from their degrees of freedom and chi-squared test when incorrect.

The second part was not as successful as the first one. Simple probability was well answered. Not all the students changed the denominator to 45 for the second probability showing their weaknesses in conditional probability. It would have been useful for the students to use a tree diagram to help them solve the last part of this question but very few did so. Some of those students that reached the last part of the question forgot to add one of the three terms. Very few used the probability of the complement.

b. The first part of the question was relatively well done. The null hypothesis and the degrees of freedom were well answered by the majority of the students. In the show that question some students used the GDC to find the expected values table and highlighted the correct value 22.05. This procedure gained no mark; the expected value formula was expected to be used here. Also those who did use the formula were expected

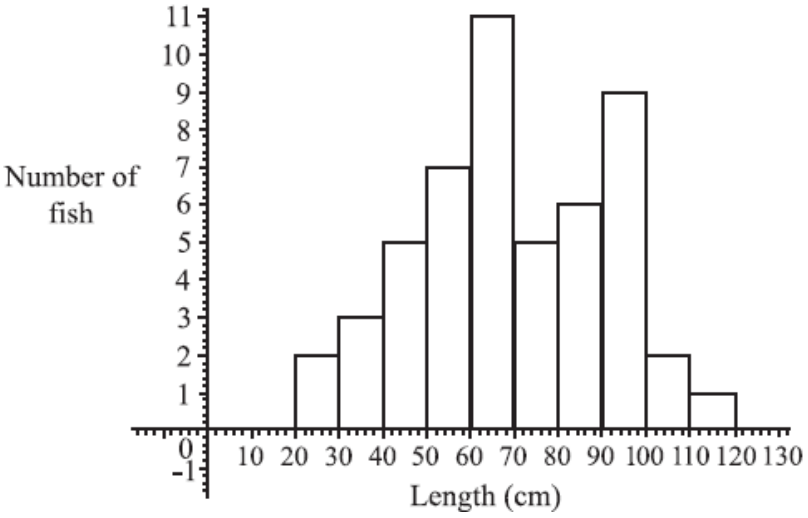
to show the unrounded value 22.05 to gain full marks in this part question. Many lost the answer mark for not doing so. GDC was used by most of the students to find the chi-squared test though some students attempted to find this value by hand which made them waste time. Correct values were compared when deciding whether to accept or not the null hypothesis and follow through marks were awarded from their degrees of freedom and chi-squared test when incorrect.

The second part was not as successful as the first one. Simple probability was well answered. Not all the students changed the denominator to 45 for the second probability showing their weaknesses in conditional probability. It would have been useful for the students to use a tree diagram to help them solve the last part of this question but very few did so. Some of those students that reached the last part of the question forgot to add one of the three terms. Very few used the probability of the complement.

c. The first part of the question was relatively well done. The null hypothesis and the degrees of freedom were well answered by the majority of the students. In the show that question some students used the GDC to find the expected values table and highlighted the correct value 22.05. This procedure gained no mark; the expected value formula was expected to be used here. Also those who did use the formula were expected to show the unrounded value 22.05 to gain full marks in this part question. Many lost the answer mark for not doing so. GDC was used by most of the students to find the chi-squared test though some students attempted to find this value by hand which made them waste time. Correct values were compared when deciding whether to accept or not the null hypothesis and follow through marks were awarded from their degrees of freedom and chi-squared test when incorrect.

The second part was not as successful as the first one. Simple probability was well answered. Not all the students changed the denominator to 45 for the second probability showing their weaknesses in conditional probability. It would have been useful for the students to use a tree diagram to help them solve the last part of this question but very few did so. Some of those students that reached the last part of the question forgot to add one of the three terms. Very few used the probability of the complement.

The figure below shows the lengths in centimetres of fish found in the net of a small trawler.



a. Find the total number of fish in the net. [2]

b. Find (i) the modal length interval, [5]

(ii) the interval containing the median length,

(iii) an estimate of the mean length.

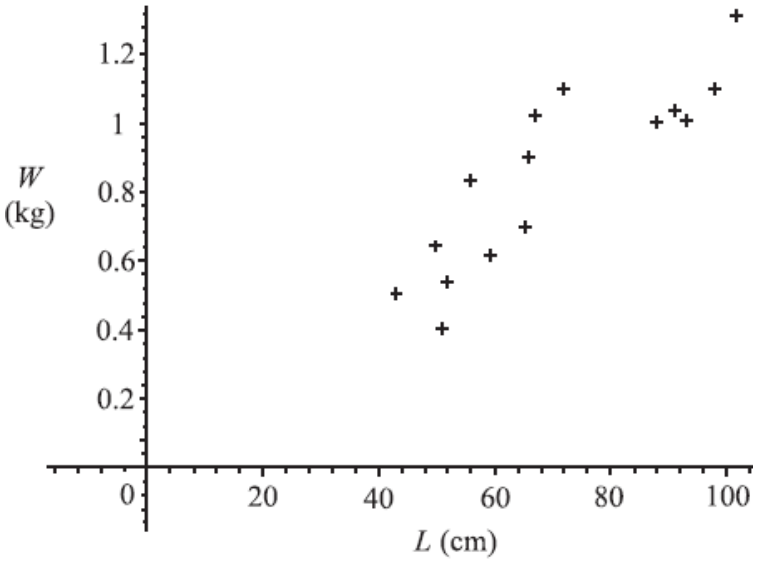
c. (i) Write down an estimate for the standard deviation of the lengths. [3]

(ii) How many fish (if any) have length **greater than** three standard deviations **above** the mean?

d. The fishing company must pay a fine if more than 10% of the catch have lengths less than 40cm. [2]

Do a calculation to decide whether the company is fined.

e. A sample of 15 of the fish was weighed. The weight, W was plotted against length, L as shown below. [2]



Exactly **two** of the following statements about the plot could be correct. Identify the two correct statements.

Note: You do **not** need to enter data in a GDC **or** to calculate r exactly.

(i) The value of r, the correlation coefficient, is approximately 0.871.

(ii) There is an exact linear relation between W and L.

(iii) The line of regression of W on L has equation $W = 0.012L + 0.008$.

(iv) There is negative correlation between the length and weight.

(v) The value of r, the correlation coefficient, is approximately 0.998.

(vi) The line of regression of W on L has equation $W = 63.5L + 16.5$.

Markscheme

a. Total = 2 + 3 + 5 + 7 + 11 + 5 + 6 + 9 + 2 + 1 ***(M1)***

(M1) *is for a sum of frequencies.*

= 51 ***(A1)(G2)***

[2 marks]

b. *Unit penalty* ***(UP)*** *is applicable where indicated in the left hand column.*

(i) modal interval is 60 – 70

Award ***(A0)*** *for 65* ***(A1)***

(ii) median is length of fish no. 26, ***(M1)(A1)***

also 60 – 70 ***(G2)***

Can award ***(A1)*****(ft)** *or* ***(G2)*****(ft)** *for 65 if* ***(A0)*** *was awarded for 65 in part (i).*

(iii) mean is $\frac{2\times25+3\times35+5\times45+7\times55+...}{51}$ ***(M1)***

(UP) = 69.5 cm (3sf) ***(A1)*(ft)*(G1)***

*Note: **(M1)** is for a sum of (frequencies multiplied by midpoint values) divided by candidate's answer from part (a). Accept mid-points 25.5, 35.5 etc or 24.5, 34.5 etc, leading to answers 70.0 or 69.0 (3sf) respectively. Answers of 69.0, 69.5 or 70.0 (3sf) with no working can be awarded **(G1).***

[5 marks]

c. *Unit penalty **(UP)** is applicable where indicated in the left hand column.*

(UP) (i) standard deviation is 21.8 cm ***(G1)***

*For any other answer without working, award **(G0).** If working is present then **(G0)(AP)** is possible.*

(ii) $69.5 + 3 \times 21.8 = 134.9 > 120$ ***(M1)***

no fish ***(A1)*(ft)*(G1)***

*For 'no fish' without working, award **(G1)** regardless of answer to (c)(i). Follow through from (c)(i) only if method is shown.*

[3 marks]

d. 5 fish are less than 40 cm in length, ***(M1)***

*Award **(M1)** for any of $\frac{5}{51}$, $\frac{46}{51}$, 0.098 or 9.8%, 0.902, 90.2% or 5.1 seen.*

hence no fine. ***(A1)*(ft)***

*Note: There is no G mark here and **(M0)(A1)** is never allowed. The follow-through is from answer in part (a).*

[2 marks]

e. (i) and (iii) are correct. ***(A1)(A1)***

[2 marks]

Examiners report

a. a) b), c) There was much confusion about how to present the intervals. Often the mid-point only was seen. (eg. 65 instead of 60-70). Understanding of mode, median and mean was usually good but too many candidates wasted time calculating standard deviation by hand and got it wrong. In c(ii) 'greater than three' caused no problems but 'above the mean' was often ignored.

b. a) b), c) There was much confusion about how to present the intervals. Often the mid-point only was seen. (eg. 65 instead of 60-70). Understanding of mode, median and mean was usually good but too many candidates wasted time calculating standard deviation by hand and got it wrong. In c(ii) 'greater than three' caused no problems but 'above the mean' was often ignored.

c. a) b), c) There was much confusion about how to present the intervals. Often the mid-point only was seen. (eg. 65 instead of 60-70). Understanding of mode, median and mean was usually good but too many candidates wasted time calculating standard deviation by hand and got it wrong. In c(ii) 'greater than three' caused no problems but 'above the mean' was often ignored.

d. d) This was often well done, even if earlier parts were poorly done.

e. e) Rather mixed performance here. It was hard to identify any consistency in the errors made.

Too much time was spent on this question. It was only worth two marks and candidates ought to have realised that it relied on a general pictorial understanding of the concepts, possibly supplemented by a little elementary arithmetic only, to compare (iii) and (vi). With good understanding, many of the options could be ruled out in a few seconds.

On one day 180 flights arrived at a particular airport. The distance travelled and the arrival status for each incoming flight was recorded. The flight was then classified as on time, slightly delayed, or heavily delayed.

The results are shown in the following table.

		Distance travelled			
		At most 500 km	Between 500 km and 5000 km	At least 5000 km	TOTAL
Arrival Status	On time	19	17	16	52
	Slightly delayed	13	18	14	45
	Heavily delayed	28	15	40	83
	TOTAL	60	50	70	180

A χ^2 test is carried out at the 10 % significance level to determine whether the arrival status of incoming flights is independent of the distance travelled.

The critical value for this test is 7.779.

A flight is chosen at random from the 180 recorded flights.

a. State the alternative hypothesis. [1]

b. Calculate the expected frequency of flights travelling at most 500 km and arriving slightly delayed. [2]

c. Write down the number of degrees of freedom. [1]

d.i. Write down the χ^2 statistic. [2]

d.ii. Write down the associated p-value. [1]

e. State, with a reason, whether you would reject the null hypothesis. [2]

f. Write down the probability that this flight arrived on time. [2]

g. Given that this flight was not heavily delayed, find the probability that it travelled between 500 km and 5000 km. [2]

h. Two flights are chosen at random from those which were slightly delayed. [3]

Find the probability that each of these flights travelled at least 5000 km.

Markscheme

a. The arrival status is dependent on the distance travelled by the incoming flight ***(A1)***

Note: Accept “associated” or “not independent”.

[1 mark]

b. $\frac{60\times45}{180}$ **OR** $\frac{60}{180}\times\frac{45}{180}\times180$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into expected value formula.

= 15 ***(A1) (G2)***

[2 marks]

c. 4 ***(A1)***

Note: Award ***(A0)*** if “2 + 2 = 4” is seen.

[1 mark]

d.i. 9.55 (9.54671…) ***(G2)***

Note: Award ***(G1)*** for an answer of 9.54.

[2 marks]

d.ii. 0.0488 (0.0487961…) ***(G1)***

[1 mark]

e. Reject the Null Hypothesis ***(A1)*****(ft)**

Note: Follow through from their hypothesis in part (a).

9.55 (9.54671…) > 7.779 ***(R1)*****(ft)**

OR

0.0488 (0.0487961…) < 0.1 ***(R1)*****(ft)**

Note: Do not award ***(A1)*****(ft)*****(R0)*****(ft)**. Follow through from part (d). Award ***(R1)*****(ft)** for a correct comparison, ***(A1)*****(ft)** for a consistent conclusion with the answers to parts (a) and (d). Award ***(R1)*****(ft)** for $\chi^2_{calc} > \chi^2_{crit}$, provided the calculated value is explicitly seen in part (d)(i).

[2 marks]

f. $\frac{52}{180}\ \left(0.289,\ \frac{13}{45},\ 28.9\,\%\right)$ ***(A1)(A1) (G2)***

Note: Award ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[2 marks]

g. $\frac{35}{97}\ (0.361,\ 36.1\,\%)$ ***(A1)(A1) (G2)***

Note: Award ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[2 marks]

h. $\frac{14}{45}\times\frac{13}{44}$ ***(A1)(M1)***

Note: Award ***(A1)*** for two correct fractions and ***(M1)*** for multiplying their two fractions.

$= \frac{182}{1980}\left(0.0919,\ \frac{91}{990},\ 0.091919\ldots,\ 9.19\,\%\right)$ ***(A1) (G2)***

[3 marks]

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d.i. [N/A]
d.ii. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]
h. [N/A]

A random sample of 167 people who own mobile phones was used to collect data on the amount of time they spent per day using their phones.

The results are displayed in the table below.

Time spent per day (t minutes)	$0 \le t < 15$	$15 \le t < 30$	$30 \le t < 45$	$45 \le t < 60$	$60 \le t < 75$	$75 \le t < 90$
Number of people	21	32	35	41	27	11

Manuel conducts a survey on a random sample of 751 people to see which television programme type they watch most from the following: Drama, Comedy, Film, News. The results are as follows.

	Drama	Comedy	Film	News
Males under 25	22	65	90	35
Males 25 and over	36	54	67	17
Females under 25	22	59	82	15
Females 25 and over	64	39	38	46

Manuel decides to ignore the ages and to test at the 5 % level of significance whether the most watched programme type is independent of **gender.**

i.a. State the modal group. [1]

i.b. Use your graphic display calculator to calculate approximate values of the mean and standard deviation of the time spent per day on these mobile phones. [3]

i.c. On graph paper, draw a fully labelled histogram to represent the data. [4]

ii.a. Draw a table with 2 rows and 4 columns of data so that Manuel can perform a chi-squared test. [3]

ii.b. State Manuel's null hypothesis and alternative hypothesis. [1]

ii.c. Find the expected frequency for the number of females who had 'Comedy' as their most-watched programme type. Give your answer to the nearest whole number. [2]

ii.d. Using your graphic display calculator, or otherwise, find the chi-squared statistic for Manuel's data. [3]

ii.e. (i) State the number of degrees of freedom available for this calculation. [3]

(ii) State his conclusion.

Markscheme

i.a. $45 \leqslant t < 60$ ***(A1)***

[1 mark]

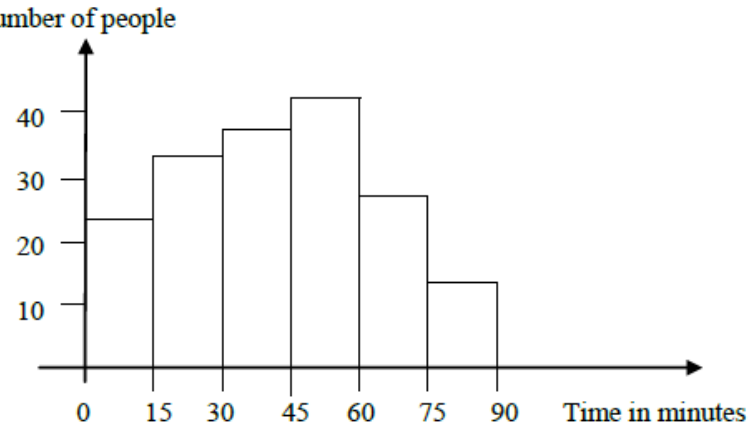
i.b. *Unit penalty **(UP)** is applicable in question part (i)(b) **only**.*

(UP) 42.4 minutes ***(G2)***

21.6 minutes ***(G1)***

[3 marks]

i.c.



(A4)

[4 marks]

ii.a.

	Drama	Comedy	Film	News
Males	58	119	157	52
Females	86	98	120	61

(M1)(M1)(A1)

[3 marks]

ii.b. H_0: favourite TV programme is independent of gender or no association between favourite TV programme and gender

H_1: favourite TV programme is dependent on gender *(must have both)* ***(A1)***

[1 mark]

ii.c. $\frac{365 \times 217}{751}$ ***(M1)***

$= 105$ ***(A1)*****(ft)*****(G2)***

[2 marks]

ii.d. 12.6 *(accept* 12.558*)* ***(G3)***

[3 marks]

ii.e. (i) 3 ***(A1)***

(ii) reject H_0 *or equivalent statement (e.g. accept H_1)* ***(A1)*****(ft)**

[3 marks]

Examiners report

i.a. Many candidates who had survived the previous two unit penalties, fell here with omission of units for the mean and standard deviation. The modal group was answered well. Part (b), finding the mean and standard deviation by GDC, was answered very poorly. Most did put the midpoints in one list and the frequencies in a second list but then either used the 2-Var stats button or 1-var stats button but only named L1 instead of L1, L2. Candidates who showed midpoints in their working did at least score a method mark.

i.b. Many candidates who had survived the previous two unit penalties, fell here with omission of units for the mean and standard deviation. The modal group was answered well. Part (b), finding the mean and standard deviation by GDC, was answered very poorly. Most did put the midpoints in one list and the frequencies in a second list but then either used the 2-Var stats button or 1-var stats button but only named L1 instead of L1, L2. Candidates who showed midpoints in their working did at least score a method mark.

i.c. Many candidates who had survived the previous two unit penalties, fell here with omission of units for the mean and standard deviation. The modal group was answered well. Part (b), finding the mean and standard deviation by GDC, was answered very poorly. Most did put the midpoints in one list and the frequencies in a second list but then either used the 2-Var stats button or 1-var stats button but only named L1 instead of L1, L2. Candidates who showed midpoints in their working did at least score a method mark.

ii.a. The chi-squared question was answered well by the majority of candidates and almost all found the chi-squared statistic correctly by GDC, though many could not look up the correct critical value.

ii.b. The chi-squared question was answered well by the majority of candidates and almost all found the chi-squared statistic correctly by GDC, though many could not look up the correct critical value.

ii.c. The chi-squared question was answered well by the majority of candidates and almost all found the chi-squared statistic correctly by GDC, though many could not look up the correct critical value.

ii.d. The chi-squared question was answered well by the majority of candidates and almost all found the chi-squared statistic correctly by GDC, though many could not look up the correct critical value.

ii.e. The chi-squared question was answered well by the majority of candidates and almost all found the chi-squared statistic correctly by GDC, though many could not look up the correct critical value.

Francesca is a chef in a restaurant. She cooks eight chickens and records their masses and cooking times. The mass m of each chicken, in kg, and its cooking time t, in minutes, are shown in the following table.

Mass m (kg)	Cooking time t (minutes)
1.5	62
1.6	75
1.8	82
1.9	83
2.0	86
2.1	87
2.1	91
2.3	98

a. Draw a scatter diagram to show the relationship between the mass of a chicken and its cooking time. Use 2 cm to represent 0.5 kg on the horizontal axis and 1 cm to represent 10 minutes on the vertical axis. [4]

b. Write down for this set of data [2]

(i) the mean mass, $\bar{m}$;

(ii) the mean cooking time, $\bar{t}$.

c. Label the point $\mathrm{M}(\bar{m}, \bar{t})$ on the scatter diagram. [1]

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

e. Using your line of best fit, estimate the cooking time, in minutes, for a 1.7 kg chicken. [2]

f. Write down the Pearson's product–moment correlation coefficient, r . [2]

g. Using your value for r , comment on the correlation. [2]

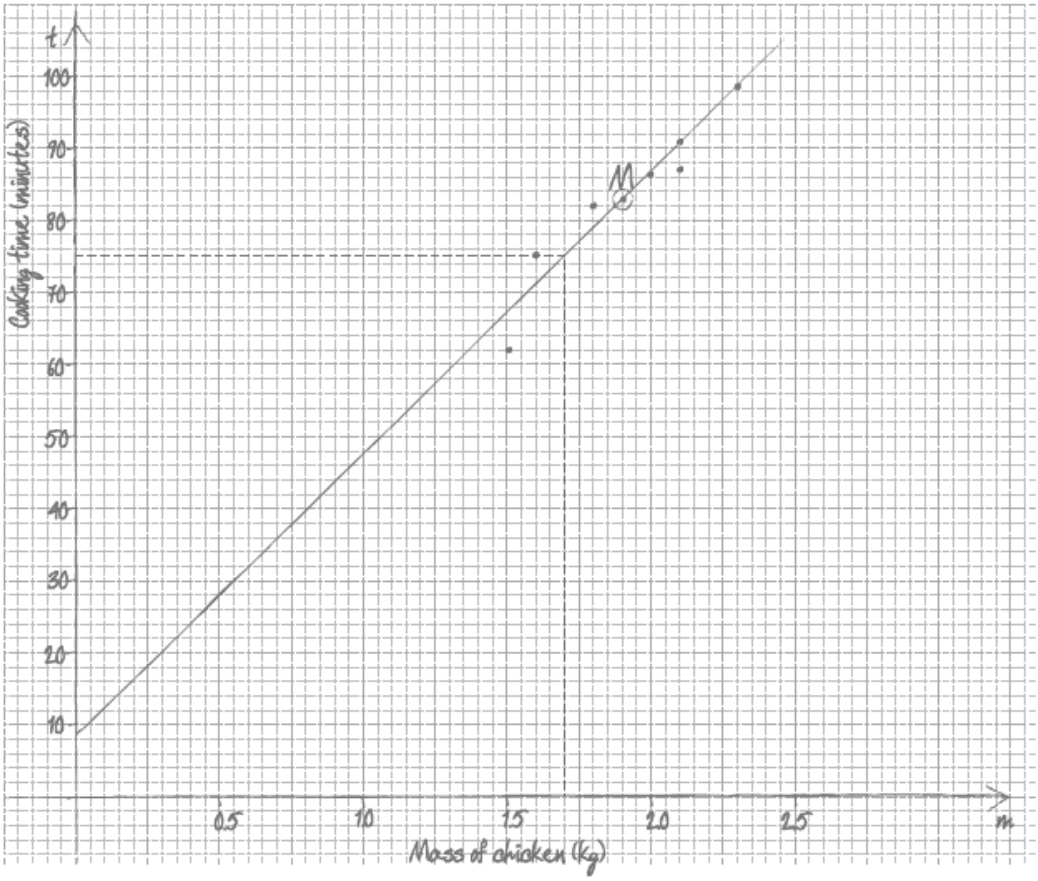
h. The cooking time of an additional 2.0 kg chicken is recorded. If the mass and cooking time of this chicken is included in the data, the correlation is weak. [2]

(i) Explain how the cooking time of this additional chicken might differ from that of the other eight chickens.

(ii) Explain how a new line of best fit might differ from that drawn in part (d).

Markscheme

a.



(A1) *for correct scales and labels (mass or m on the horizontals axis, time or t on the vertical axis)*

(A3) *for 7 or 8 correctly placed data points*

(A2) *for 5 or 6 correctly placed data points*

(A1) *for 3 or 4 correctly placed data points,* ***(A0)*** *otherwise.* ***(A4)***

Note: If axes reversed award at most ***(A0)(A3)*****(ft)**. If graph paper not used, award at most ***(A1)(A0)***.

b. (i) 1.91 (kg) (1.9125 kg) ***(G1)***

(ii) 83 (minutes) ***(G1)***

c. Their mean point labelled. ***(A1)*****(ft)**

Note: Follow through from part (b). Accept any clear indication of the mean point. For example: circle around point, (*m*, *t*), M , etc.

d. Line of best fit drawn on scatter diagram. ***(A1)*****(ft)*****(A1)*****(ft)**

Notes:Award ***(A1)*****(ft)** for straight line through their mean point, ***(A1)*****(ft)** for line of best fit with intercept 9(±2) . The second ***(A1)*****(ft)** can be awarded even if the line does not reach the *t*-axis but, if extended, the *t*-intercept is correct.

e. 75 ***(M1)(A1)*****(ft)*****(G2)***

Notes: Accept 74.77 from the regression line equation. Award ***(M1)*** for indication of the use of their graph to get an estimate **OR** for correct substitution of 1.7 in the correct regression line equation $t = 38.5m + 9.32$.

f. 0.960 (0.959614...) ***(G2)***

Note: Award ***(G0)(G1)*****(ft)** for 0.95, 0.959

g. Strong and positive ***(A1)*(ft)*(A1)*(ft)**

Note: Follow through from their correlation coefficient in part (f).

h. (i) Cooking time is much larger (or smaller) than the other eight ***(A1)***

(ii) The gradient of the new line of best fit will be larger (or smaller) ***(A1)***

Note: Some acceptable explanations may include but are not limited to:

The line of best fit may be further away from the plotted points
It may be steeper than the previous line (as the mean would change)
The t-intercept of the new line is smaller (larger)

Do not accept vague explanations, like:

The new line would vary
It would not go through all points
It would not fit the patterns
The line may be slightly tilted

Examiners report

a. [N/A]
b. [N/A]
c. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]
h. [N/A]

A survey of 400 people is carried out by a market research organization in two different cities, Buenos Aires and Montevideo. The people are asked which brand of cereal they prefer out of Chocos, Zucos or Fruti. The table below summarizes their responses.

	Chocos	Zucos	Fruti	Total
Buenos Aires	43	85	62	190
Montevideo	57	35	118	210
Total	100	120	180	400

The following table shows the cost in AUD of seven paperback books chosen at random, together with the number of pages in each book.

Book	1	2	3	4	5	6	7
Number of pages (x)	50	120	200	330	400	450	630
Cost (y AUD)	6.00	5.40	7.20	4.60	7.60	5.80	5.20

i.a. One person is chosen at random from those surveyed. Find the probability that this person [4]

(i) does not prefer Zucos;

(ii) prefers Chocos, given that they live in Montevideo.

i.b. Two people are chosen at random from those surveyed. Find the probability that they both prefer Fruti. [3]

i.c. The market research organization tests the survey data to determine whether the brand of cereal preferred is associated with a city. A chi-squared test at the 5% level of significance is performed. [1]

State the null hypothesis.

i.d. The market research organization tests the survey data to determine whether the brand of cereal preferred is associated with a city. A chi-squared test at the 5% level of significance is performed. [1]

State the number of degrees of freedom.

i.e. The market research organization tests the survey data to determine whether the brand of cereal preferred is associated with a city. A chi-squared test at the 5% level of significance is performed. [2]

Show that the expected frequency for the number of people who live in Montevideo and prefer Zucos is 63.

i.f. The market research organization tests the survey data to determine whether the brand of cereal preferred is associated with a city. A chi-squared test at the 5% level of significance is performed. [2]

Write down the chi-squared statistic for this data.

i.g. The market research organization tests the survey data to determine whether the brand of cereal preferred is associated with a city. A chi-squared test at the 5% level of significance is performed. [2]

State whether the market research organization would accept the null hypothesis. Clearly justify your answer.

ii.a. Plot these pairs of values on a scatter diagram. Use a scale of $1\ \text{cm}$ to represent 50 pages on the horizontal axis and $1\ \text{cm}$ to represent $1\ \text{AUD}$ on the vertical axis. [3]

ii.b. Write down the linear correlation coefficient, r, for the data. [2]

ii.c. Stephen wishes to buy a paperback book which has 350 pages in it. He plans to draw a line of best fit to determine the price. State whether or not this is an appropriate method in this case and justify your answer. [2]

Markscheme

i.a. (i) $\frac{280}{400}$ $(0.7,\ 70\%\ \text{or equivalent})$ ***(A1)(A1)(G2)***

Note: ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

(ii) $\frac{57}{210}$ $\left(\frac{19}{70},\ 0.271,\ 27.1\%\right)$ ***(A1)(A1)(G2)***

Note: ***(A1)*** for correct numerator, ***(A1)*** for correct denominator.

[4 marks]

i.b. $\frac{180}{400} \times \frac{179}{399}$ ***(A1)(M1)***

Note: ***(A1)*** for correct values seen, ***(M1)*** for multiplying their two values, ***(A1)*** for correct answer.

$= \frac{537}{2660}$ $(= 0.202)$ ***(A1)(G3)***

[3 marks]

i.c. H_0 : 'the preference of brand of cereal is independent of the city'. ***(A1)***

OR

H_0 : 'there is no association between the brand of cereal and city'.

[1 mark]

i.d. $df = 2$ ***(A1)***

[1 mark]

i.e. $\frac{210 \times 120}{400}$ ***(M1)(A1)***

Note: ***(M1)*** for substituting in correct formula, ***(A1)*** for correct values.

$= 63$ ***(AG)***

Note: Final line must be seen or previous ***(A1)*** mark is lost.

[2 marks]

i.f. 39.3 ***(G2)***

Note: Award ***(G1)(A0)(AP)*** if answers not to 3 significant figures.

[2 marks]

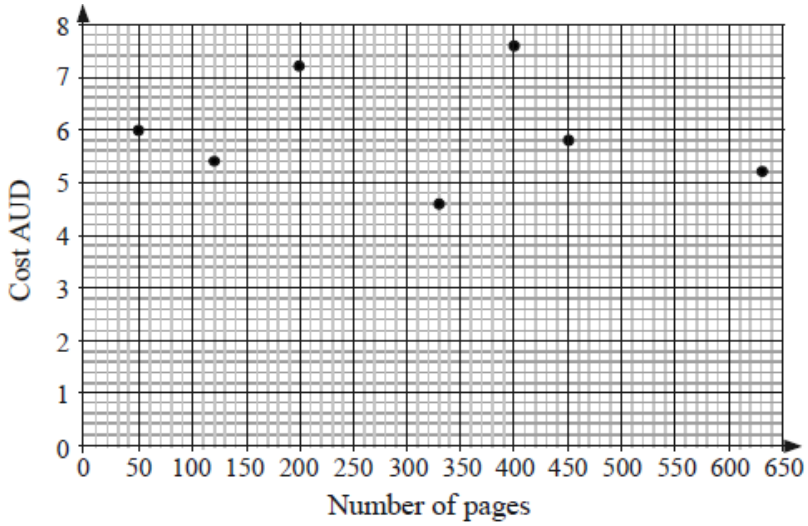
i.g. $p - \text{value} < 0.05$ ***(R1)*****(ft)**

Do not accept H_0 . ***(A1)*****(ft)**

Notes: Allow 'Reject H_0 or equivalent'. **(ft)** from their χ^2 statistic.
Award ***(R1)*****(ft)** for comparing the appropriate values. ***(A1)*****(ft)** can be awarded only if the conclusion is valid according to the comparison given. If no reason given or if reason is wrong both marks are lost. Note that ***(R1)(A0)*****(ft)** can be awarded but ***(R0)(A1)*****(ft)** cannot.

[2 marks]

ii.a.



(A1)(A1)(A1)

Notes: ***(A1)*** for label and scales, ***(A2)*** for all points correct, ***(A1)*** for 5 or 6 correct. Award a maximum of ***(A2)*** if points are joined.

[3 marks]

ii.b $r = -0.141$ ***(G2)***

Note: If negative sign is missing award ***(G1)(G0)***.

[2 marks]

ii.c. 'The coefficient of correlation is too low, (very) weak (linear) relationship'. ***(R1)***

Not a sensible thing to do, *accept 'no'.* ***(A1)***

Note: Do not award ***(R0)(A1)***. The correlation coefficient has to be mentioned in their reasoning.

[2 marks]

Examiners report

i.a. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.b. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.c. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.d. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.e. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.f. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected value formula should have been used instead.

i.g. Candidates answered part (a) correctly. Some lost one out of the 4 marks for making an error in the denominator of the conditional probability. In (b) many students failed to see that (b) was 'without replacement'. Parts (c), (d) and (e) seemed to be very well done by some centres and uniformly badly by others. In (e) many gave the table from the GDC and highlighted the value 63 for which no mark was gained. Expected

value formula should have been used instead.

ii.a. The graph was well done with almost all candidates labelling and scaling the axes correctly. A minority of students joined the points or drew the graph on lined paper which prevented them from gaining full marks in this part of the question.

In (b) some candidates were not able to calculate the linear correlation coefficient. A few G2 comments pointed out that the command term used may have been ambiguous to some candidates and they did not think that they could use their GDC to find *r*. Some attempted to use the formula even though the value of S_{xy} was not given. The guide says that 'A GDC can be used to calculate *r* when raw data is given'. This potential unfairness was taken into consideration during the setting of boundaries so that no candidate was disadvantaged by the possible ambiguous wording of the question. In future the command term 'Using your GDC' or 'Write down' will be used in similar questions.

Some students who did use the GDC gave r^2 instead of r. This really caught the attention of many examiners as r^2 is not in the syllabus.

ii.b. The graph was well done with almost all candidates labelling and scaling the axes correctly. A minority of students joined the points or drew the graph on lined paper which prevented them from gaining full marks in this part of the question.

In (b) some candidates were not able to calculate the linear correlation coefficient. A few G2 comments pointed out that the command term used may have been ambiguous to some candidates and they did not think that they could use their GDC to find *r*. Some attempted to use the formula even though the value of S_{xy} was not given. The guide says that 'A GDC can be used to calculate *r* when raw data is given'. This potential unfairness was taken into consideration during the setting of boundaries so that no candidate was disadvantaged by the possible ambiguous wording of the question. In future the command term 'Using your GDC' or 'Write down' will be used in similar questions.

Some students who did use the GDC gave r^2 instead of r. This really caught the attention of many examiners as r^2 is not in the syllabus.

ii.c. The graph was well done with almost all candidates labelling and scaling the axes correctly. A minority of students joined the points or drew the graph on lined paper which prevented them from gaining full marks in this part of the question.

In (b) some candidates were not able to calculate the linear correlation coefficient. A few G2 comments pointed out that the command term used may have been ambiguous to some candidates and they did not think that they could use their GDC to find *r*. Some attempted to use the formula even though the value of S_{xy} was not given. The guide says that 'A GDC can be used to calculate *r* when raw data is given'. This potential unfairness was taken into consideration during the setting of boundaries so that no candidate was disadvantaged by the possible ambiguous wording of the question. In future the command term 'Using your GDC' or 'Write down' will be used in similar questions.

Some students who did use the GDC gave r^2 instead of r. This really caught the attention of many examiners as r^2 is not in the syllabus.

Part A

A university required all Science students to study one language for one year. A survey was carried out at the university amongst the 150 Science students. These students all studied one of either French, Spanish or Russian. The results of the survey are shown below.

	French	Spanish	Russian
Female	9	29	12
Male	31	40	29

Ludmila decides to use the χ^2 test at the 5% level of significance to determine whether the choice of language is independent of gender.

At the end of the year, only seven of the female Science students sat examinations in Science and French. The marks for these seven students are shown in the following table.

Science (S)	23	51	56	62	12	73	72
French (F)	65	45	45	40	70	36	30

A.a State Ludmila's null hypothesis. [1]

A.b Write down the number of degrees of freedom. [1]

A.c Find the expected frequency for the females studying Spanish. [2]

A.d Use your graphic display calculator to find the χ^2 test statistic for this data. [2]

A.e State whether Ludmila accepts the null hypothesis. Give a reason for your answer. [2]

B.a Draw a labelled scatter diagram for this data. Use a scale of $2\ \text{cm}$ to represent $10\ \text{marks}$ on the x-axis (S) and $10\ \text{marks}$ on the y-axis (F). [4]

B.b Use your graphic calculator to find [2]

(i) $\bar{S}$, the mean of S ;

(ii) $\bar{F}$, the mean of F .

B.c Plot the point $\text{M}(\bar{S}, \bar{F})$ on your scatter diagram. [1]

B.d Use your graphic display calculator to find the equation of the regression line of F on S . [2]

B.e Draw the regression line on your scatter diagram. [2]

B.f. Carletta's mark on the Science examination was 44. She did not sit the French examination. [2]

Estimate Carletta's mark for the French examination.

B.g Monique's mark on the Science examination was 85. She did not sit the French examination. Her French teacher wants to use the regression line to estimate Monique's mark. [2]

State whether the mark obtained from the regression line for Monique's French examination is reliable. Justify your answer.

Markscheme

A.a.

H_0 : Choice of language is independent of gender. ***(A1)***

Notes: Do not accept "not related" or "not correlated".

[1 mark]

A.b.

2 ***(A1)***

[1 mark]

A.c.

$\frac{50\times69}{150} = 23$ ***(M1)(A1)(G2)***

Notes: Award ***(M1)*** for correct substituted formula, ***(A1)*** for 23.

[2 marks]

A.d.

$\chi^2 = 4.77$ ***(G2)***

Notes: If answer is incorrect, award ***(M1)*** for correct substitution in the correct formula (all terms).

[2 marks]

A.e.

Accept H_0 since

$\chi^2_{calc} < \chi^2_{crit}(5.99)$ **or** p-value $(0.0923) > 0.05$ ***(R1)(A1)*****(ft)**

Notes: Do not award ***(R0)(A1)***. Follow through from their (d) and (b).

B.a.

Award ***(A1)*** for correct scale and labels.

Award ***(A3)*** for all seven points plotted correctly, ***(A2)*** for 5 or 6 points plotted correctly, ***(A1)*** for 3 or 4 points plotted correctly.

(A4)

[4 marks]

B.b.(i) $\bar{S} = 49.9,$ ***(G1)***

(ii) $\bar{F} = 47.3$ ***(G1)***

[2 marks]

B.c. $\mathrm{M}(49.9, 47.3)$ plotted on scatter diagram ***(A1)*****(ft)**

Notes: Follow through from (a) and (b).

[1 mark]

B.d.

$F = -0.619S + 78.2$ ***(G1)(G1)***

Notes: Award ***(G1)*** for $-0.619S$, ***(G1)*** for 78.2. If the answer is not in the form of an equation, award ***(G1)(G0)***. Accept $y = -0.619x + 78.2$.

OR

(F - 47.3 = - 0.619(S - 49.9)) ***(G1)(G1)***

Note: Award ***(G1)*** for -0.619, ***(G1)*** for the coordinates of their midpoint used. Follow through from their values in (b).

[2 marks]

B.e. line drawn on scatter diagram ***(A1)*****(ft)*****(A1)*****(ft)**

Notes: The drawn line **must** be straight for any marks to be awarded. Award ***(A1)*****(ft)** passing through their M plotted in (c). Award ***(A1)*****(ft)** for correct y-intercept. Follow through from their y-intercept found in (d).

[2 marks]

B.f. $F = -0.619 \times 44 + 78.2$ ***(M1)***

$= 51.0$ (allow 51 or 50.9) ***(A1)*****(ft)*****(G2)*****(ft)**

Note: Follow through from their equation.

OR

(M1) any indication of an acceptable graphical method. ***(M1)***

(A1)**(ft)** from their regression line. ***(A1)*****(ft)*****(G2)*****(ft)**

[2 marks]

B.g. not reliable ***(A1)***

Monique's score in Science is outside the range of scores used to create the regression line. ***(R1)***

Note: Do not award ***(A1)(R0)***.

[2 marks]

Examiners report

A.a **Part A: Chi-square test**

This question part was answered well by most candidates. The null hypothesis and degrees of freedom were mostly correct. Some candidates offered a conclusion supported by good justifications, but others still showed lack of the necessary knowledge to do that. Some responses to part d) incurred an accuracy penalty for not adhering to the required accuracy level.

A.b **Part A: Chi-square test**

This question part was answered well by most candidates. The null hypothesis and degrees of freedom were mostly correct. Some candidates offered a conclusion supported by good justifications, but others still showed lack of the necessary knowledge to do that. Some responses to part d) incurred an accuracy penalty for not adhering to the required accuracy level.

A.c **Part A: Chi-square test**

This question part was answered well by most candidates. The null hypothesis and degrees of freedom were mostly correct. Some candidates offered a conclusion supported by good justifications, but others still showed lack of the necessary knowledge to do that. Some responses to part d) incurred an accuracy penalty for not adhering to the required accuracy level.

A.d **Part A: Chi-square test**

This question part was answered well by most candidates. The null hypothesis and degrees of freedom were mostly correct. Some candidates offered a conclusion supported by good justifications, but others still showed lack of the necessary knowledge to do that. Some responses to part d) incurred an accuracy penalty for not adhering to the required accuracy level.

A.e **Part A: Chi-square test**

This question part was answered well by most candidates. The null hypothesis and degrees of freedom were mostly correct. Some candidates offered a conclusion supported by good justifications, but others still showed lack of the necessary knowledge to do that. Some responses to part d) incurred an accuracy penalty for not adhering to the required accuracy level.

B.a **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.b **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.c **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.d **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.e **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.f **Part B: Scatter plot and Regression line**

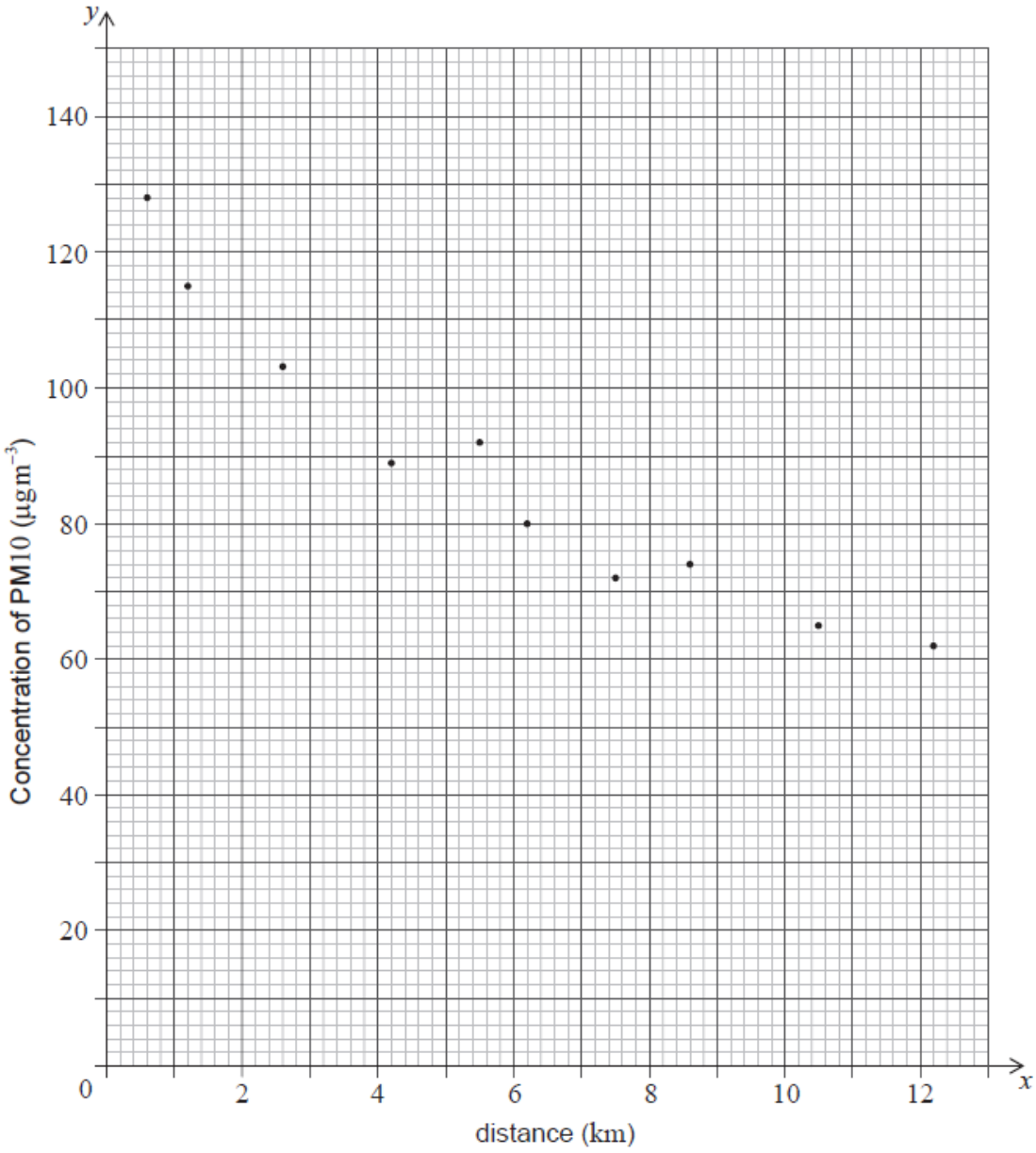
Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

B.g **Part B: Scatter plot and Regression line**

Many candidates reversed the axes in a), but the points were mostly plotted well. The values of the coefficients of the equation of the regression line $y = ax + b$ were often given not to the required 3 significant figure accuracy, and incurred a penalty. The regression line was often drawn not passing through point M and the y-intercept. The responses to the last part of the question were particularly weak, and many candidates were not able to offer a satisfactory reason to support their conclusion.

---

a. For an ecological study, Ernesto measured the average concentration (y) of the fine dust, $\mathrm{PM10}$, in the air at different distances (x) from a power plant. His data are represented on the following scatter diagram. The concentration of $\mathrm{PM10}$ is measured in micrograms per cubic metre and the distance is measured in kilometres. [2]



His data are also listed in the following table.

Distance (x)	0.6	1.2	2.6	a	5.5	6.2	7.5	8.6	10.5	12.2
Concentration of PM10 (y)	128	115	103	89	92	80	72	b	65	62

Use the scatter diagram to find the value of a and of b in the table.

b. Calculate [4]

i) $\bar{x}$, the mean distance from the power plant;

ii) $\bar{y}$, the mean concentration of $\mathrm{PM10}$;

iii) r , the Pearson's product–moment correlation coefficient.

c. Write down the equation of the regression line y on x . [2]

d. Ernesto's school is located $14\,\mathrm{km}$ from the power plant. He uses the equation of the regression line to estimate the concentration of $\mathrm{PM10}$ in the air at his school. [4]

i) Calculate the value of Ernesto's estimate.

ii) State whether Ernesto's estimate is reliable. Justify your answer.

Markscheme

a. $a = 4.2\,;\ b = 74$ ***(A1)(A1)***

b. i) $5.91\,(\text{km})$ ***(A1)*****(ft)**

ii) 88 (micrograms per cubic metre) ***(A1)*****(ft)**

Note: Follow through from part (a) irrespective of working seen.

iii) $-0.956\ \ (-0.955528...)$ ***(G2)*****(ft)**

Note: Follow through from part (a) irrespective of working seen.

c. $y = -5.39x + 120\ \ (y = -5.38955...x + 119.852...)$ ***(A1)*****(ft)*****(A1)*****(ft)**

Note: Award ***(A1)*****(ft)** for -5.39. Award ***(A1)*****(ft)** for 120. If answer is not an equation award at most ***(A1)*****(ft)*****(A0)***. Follow through from part (a) irrespective of working seen.

d. i) $-5.38955... \times 14 + 119.852...$ ***(M1)***

Note: Award ***(M1)*** for correct substitution into their regression line.

$= 44.4\ (44.3984...)$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (c). Accept $44.5\ (44.54)$ from use of 3 significant figure values.

ii) Ernesto's estimate is not reliable ***(A1)***

this is extrapolation ***(R1)***

OR

$14\,\text{km}$ is not within the range (outside the domain) of distances given ***(R1)***

Note: Do not accept "14 is too high" or "14 is an outlier" or "result not valid/not reliable" if explanation not given. Do not award ***(A1)(R0)***. Do not accept reasoning based on the strength of r.

Examiners report

a. Question 1: Reading scatter diagram, mean, correlation and regression line.

The majority of the candidates scored very well on this question. There were only a few candidates who read the diagram incorrectly. The most common mistake in parts (b), (c) and (d)(i) were rounding errors, sometimes resulting in candidates losing follow-through marks when working was not presented. Part (d)(ii) was answered incorrectly by most candidates. The most common incorrect answer was based on strong correlation. Some commented on the trend of decreasing PM10 values for increasing distances, showing lack of understanding about extrapolation.

b. Question 1: Reading scatter diagram, mean, correlation and regression line.

The majority of the candidates scored very well on this question. There were only a few candidates who read the diagram incorrectly. The most common mistake in parts (b), (c) and (d)(i) were rounding errors, sometimes resulting in candidates losing follow-through marks when working was not presented. Part (d)(ii) was answered incorrectly by most candidates. The most common incorrect answer was based on strong correlation. Some commented on the trend of decreasing PM10 values for increasing distances, showing lack of understanding about extrapolation.

c. Question 1: Reading scatter diagram, mean, correlation and regression line.

The majority of the candidates scored very well on this question. There were only a few candidates who read the diagram incorrectly. The most common mistake in parts (b), (c) and (d)(i) were rounding errors, sometimes resulting in candidates losing follow-through marks when working was not presented. Part (d)(ii) was answered incorrectly by most candidates. The most common incorrect answer was based on strong correlation. Some commented on the trend of decreasing PM10 values for increasing distances, showing lack of understanding about extrapolation.

d. Question 1: Reading scatter diagram, mean, correlation and regression line.

The majority of the candidates scored very well on this question. There were only a few candidates who read the diagram incorrectly. The most common mistake in parts (b), (c) and (d)(i) were rounding errors, sometimes resulting in candidates losing follow-through marks when working was not presented. Part (d)(ii) was answered incorrectly by most candidates. The most common incorrect answer was based on strong correlation. Some commented on the trend of decreasing PM10 values for increasing distances, showing lack of understanding about extrapolation.

A manufacturer claims that fertilizer has an effect on the height of rice plants. He measures the height of fertilized and unfertilized plants. The results are given in the following table.

Plant height	Fertilized plants	Unfertilized plants
>75 cm	115	80
$50-75$ cm	45	65
<50 cm	20	35

A chi-squared test is performed to decide if the manufacturer's claim is justified at the **1 %** level of significance.

The population of fleas on a dog after t days, is modelled by

$$N = 4 \times (2)^{\frac{t}{4}},\ t \geqslant 0$$

Some values of N are shown in the table below.

t	0	4	8	12	16	20
N	p	8	16	32	q	128

i, a. Write down the null and alternative hypotheses for this test. [2]

i, b. For the number of fertilized plants with height greater than 75 cm, show that the expected value is 97.5. [3]

i, c. Write down the value of χ^2_{calc}. [2]

i, d. Write down the number of degrees of freedom. [1]

i, f. Is the manufacturer's claim justified? Give a reason for your answer. [2]

ii, a. Write down the value of p. [1]

ii, a. Write down the value of q. [2]

ii, b. Using the values in the table above, draw the graph of N for $0 \leq t \leq 20$. Use 1 cm to represent 2 days on the horizontal axis and 1 cm to represent 10 fleas on the vertical axis. [6]

ii, c. **Use your graph** to estimate the number of days for the population of fleas to reach 55. [2]

Markscheme

i, a. H_0: The height of the rice plants is independent of the use of a fertilizer. ***(A1)***

Notes: For independent accept "not associated", can accept "the use of a fertilizer has no effect on the height of the plants".

Do not accept "not correlated".

H_1: The height of the rice plants is not independent (dependent) of the use of fertilizer. ***(A1)*****(ft)**

Note: If H_0 and H_1 are reversed award ***(A0)(A1)*****(ft)**.

[2 marks]

i, b. $\frac{180 \times 195}{360}$ **or** $\frac{180}{360} \times \frac{195}{360} \times 360$ ***(A1)(A1)(M1)***

= 97.5 ***(AG)***

Notes: Award ***(A1)*** for numerator, ***(A1)*** for denominator ***(M1)*** for division.

If final 97.5 is not seen award at most ***(A1)(A0)(M1)***.

[3 marks]

i, c. $\chi^2_{calc} = 14.01 (14.0, 14)$ ***(G2)***

OR

If worked out by hand award ***(M1)*** for correct substituted formula with correct values, ***(A1)*** for correct answer. ***(M1)(A1)***

[2 marks]

i, d. 2 ***(A1)***

[1 mark]

i, f. $\chi^2_{calc} > \chi^2_{crit}$ ***(R1)***

The manufacturer's claim is justified. (or equivalent statement) ***(A1)***

Note: Do not accept ***(R0)(A1)***.

[2 marks]

ii, a(i) $= 4$ ***(G1)***

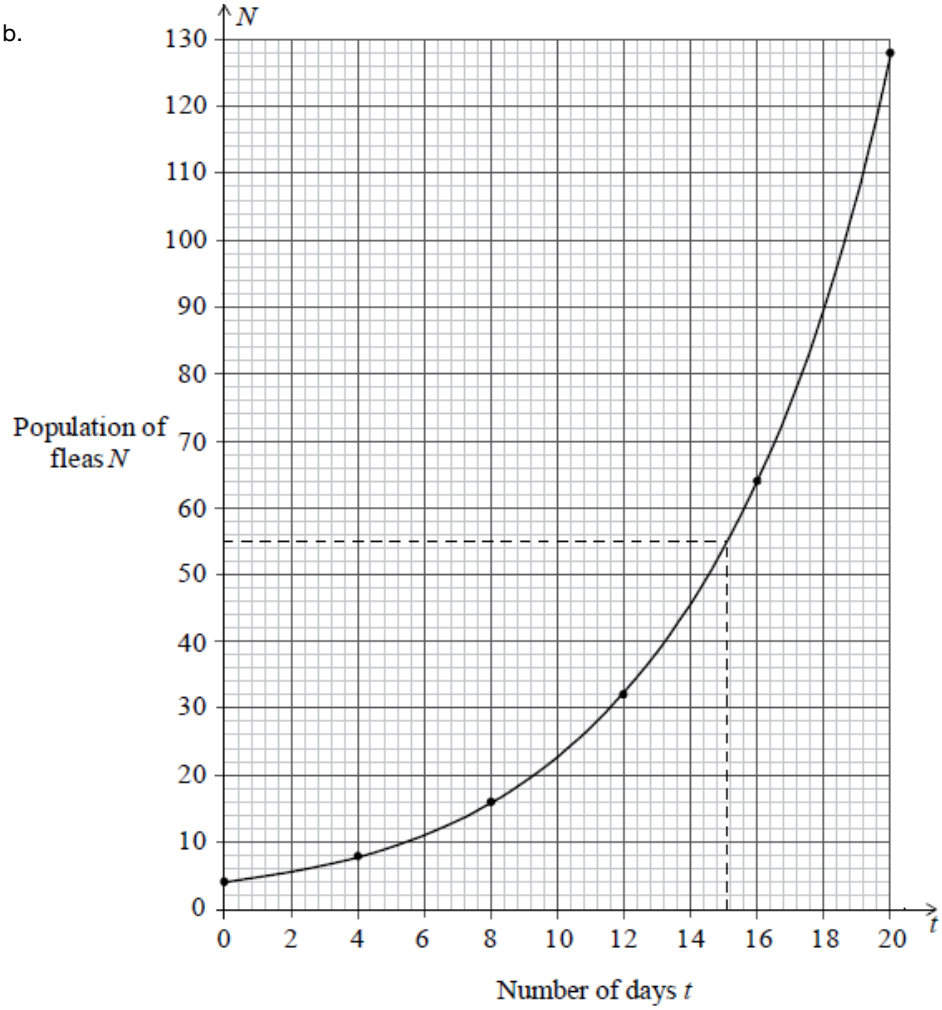
[1 mark]

ii, a(ii) $= 4(2)^{\frac{16}{4}}$ ***(M1)***

$= 64$ ***(A1)(G2)***

[2 marks]

ii, b.



(A1)(A1)(A1) ***(A3)***

Notes: Award ***(A1)*** for *x* axis with correct scale and label, ***(A1)*** for *y* axis with correct scale and label.

Accept *x* and *y* for labels.

If *x* and *y* axis reversed award at most ***(A0)(A1)*****(ft)**.

(A1) for smooth curve.

Award ***(A3)*** for all 6 points correct, ***(A2)*** for 4 or 5 points correct, ***(A1)*** for 2 or 3 points correct, ***(A0)*** otherwise.

[6 marks]

ii, c 15 (±0.8) ***(M1)(A1)*****(ft)*****(G2)***

Note: Award ***(M1)*** for line drawn shown on graph, ***(A1)*****(ft)** from candidate's graph.

[2 marks]

Examiners report

i, a. It was clear that the candidates who performed poorly in part (i) lacked the basic knowledge of chi-squared analysis. Some mixed up the null and alternate hypotheses and also were not able to correctly demonstrate the way of finding the expected value. There were many errors in finding the critical value of χ^2 at the 1% level of significance.

i, b. It was clear that the candidates who performed poorly in part (i) lacked the basic knowledge of chi-squared analysis. Some mixed up the null and alternate hypotheses and also were not able to correctly demonstrate the way of finding the expected value. There were many errors in finding the critical value of χ^2 at the 1% level of significance.

i, c. It was clear that the candidates who performed poorly in part (i) lacked the basic knowledge of chi-squared analysis. Some mixed up the null and alternate hypotheses and also were not able to correctly demonstrate the way of finding the expected value. There were many errors in finding the critical value of χ^2 at the 1% level of significance.

i, d. It was clear that the candidates who performed poorly in part (i) lacked the basic knowledge of chi-squared analysis. Some mixed up the null and alternate hypotheses and also were not able to correctly demonstrate the way of finding the expected value. There were many errors in finding the critical value of χ^2 at the 1% level of significance.

i, f. It was clear that the candidates who performed poorly in part (i) lacked the basic knowledge of chi-squared analysis. Some mixed up the null and alternate hypotheses and also were not able to correctly demonstrate the way of finding the expected value. There were many errors in finding the critical value of χ^2 at the 1% level of significance.

ii, a. Candidates found this part rather easy, with some making arithmetic mistakes and thus losing one or more marks. The graph was well done with a high percentage scoring full marks. Some candidates did not label the axes, others had an incorrect scale and a few lost one mark for not drawing a smooth curve.

ii, a. Candidates found this part rather easy, with some making arithmetic mistakes and thus losing one or more marks. The graph was well done with a high percentage scoring full marks. Some candidates did not label the axes, others had an incorrect scale and a few lost one mark for not drawing a smooth curve.

ii, b. Candidates found this part rather easy, with some making arithmetic mistakes and thus losing one or more marks. The graph was well done with a high percentage scoring full marks. Some candidates did not label the axes, others had an incorrect scale and a few lost one mark for not drawing a smooth curve.

ii, c. Candidates found this part rather easy, with some making arithmetic mistakes and thus losing one or more marks. The graph was well done with a high percentage scoring full marks. Some candidates did not label the axes, others had an incorrect scale and a few lost one mark for not drawing a smooth curve.

The weight, *W*, of basketball players in a tournament is found to be normally distributed with a mean of 65 kg and a standard deviation of 5 kg.

The probability that a basketball player has a weight that is within 1.5 standard deviations of the mean is q.

A basketball coach observed 60 of her players to determine whether their performance and their weight were independent of each other. Her observations were recorded as shown in the table.

		Performance	
		Satisfactory	Excellent
Weight	Below average	6	10
	Average	7	15
	Above average	12	10

She decided to conduct a χ^2 test for independence at the 5% significance level.

a.i. Find the probability that a basketball player has a weight that is less than 61 kg. [2]

a.ii. In a training session there are 40 basketball players. [2]

Find the expected number of players with a weight less than 61 kg in this training session.

b.i. Sketch a normal curve to represent this probability. [2]

b.ii. Find the value of q. [1]

c. Given that $P(W > k) = 0.225$, find the value of k. [2]

d.i. For this test state the null hypothesis. [1]

d.ii. For this test find the p-value. [2]

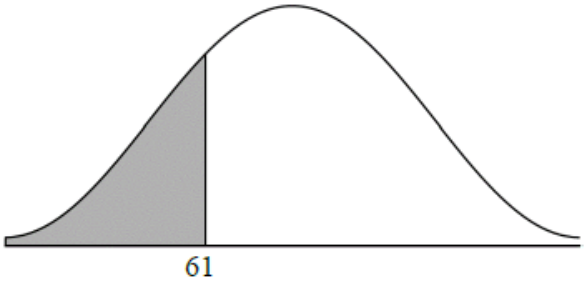
e. State a conclusion for this test. Justify your answer. [2]

Markscheme

a.i. $P(W < 61)$ ***(M1)***

Note: Award ***(M1)*** for correct probability statement.

OR



(M1)

Note: Award ***(M1)*** for correct region labelled and shaded on diagram.

= 0.212 (0.21185…, 21.2%) ***(A1)(G2)***

[2 marks]

a.ii. 40 × 0.21185… ***(M1)***

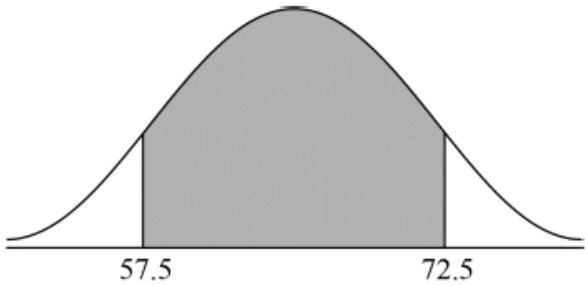
Note: Award ***(M1)*** for product of 40 and their 0.212.

= 8.47 (8.47421...) ***(A1)*****(ft)*****(G2)***

Note: Follow through from their part (a)(i) provided their answer to part (a)(i) is less than 1.

[2 marks]

b.i.



(A1)(M1)

Note: Award ***(A1)*** for two correctly labelled vertical lines in approximately correct positions. The values 57.5 and 72.5, or $\mu - 1.5\sigma$ and $\mu + 1.5\sigma$ are acceptable labels. Award ***(M1)*** for correctly shaded region marked by their two vertical lines.

[2 marks]

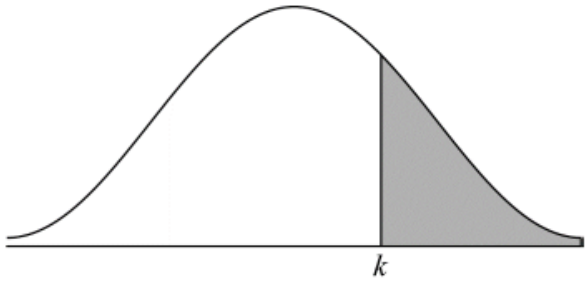
b.ii. 0.866 (0.86638…, 86.6%) ***(A1)*****(ft)**

Note: Follow through from their part (b)(i) shaded region if their values are clear.

[1 mark]

c. $P(W < k) = 0.775$ ***(M1)***

OR



(M1)

Note: Award ***(A1)*** for correct region labelled and shaded on diagram.

(k =) 68.8 (68.7770…) ***(A1)(G2)***

[2 marks]

d.i. (H_0:) performance (of players) and (their) weight are independent. ***(A1)***

Note: Accept “there is no association between performance (of players) and (their) weight”. Do not accept "not related" or "not correlated" or "not influenced".

[1 mark]

d.ii. 0.287 (0.287436…) ***(G2)***

[2 marks]

e. accept/ do not reject null hypothesis/H_0 ***(A1)*****(ft)**

OR

performance (of players) and (their) weight are independent. ***(A1)*****(ft)**

0.287 > 0.05 ***(R1)*(ft)**

Note: Accept *p*-value>significance level provided their *p*-value is seen in b(ii). Accept 28.7% > 5%. Do not award ***(A1)(R0)***. Follow through from part (d).

[2 marks]

Examiners report

a.i. [N/A]
a.ii. [N/A]
b.i. [N/A]
b.ii. [N/A]
c. [N/A]
d.i. [N/A]
d.ii. [N/A]
e. [N/A]

The heat output in thermal units from burning 1 kg of wood changes according to the wood's percentage moisture content. The moisture content and heat output of 10 blocks of the same type of wood each weighing 1 kg were measured. These are shown in the table.

Moisture content % (x)	8	15	22	30	34	45	50	60	74	82
Heat output (y)	80	77	74	69	68	61	61	55	50	45

a. Draw a scatter diagram to show the above data. Use a scale of 2 cm to represent 10% on the *x*-axis and a scale of 2 cm to represent 10 thermal units on the *y*-axis. [4]

b. Write down [2]

(i) the mean percentage moisture content, $\bar{x}$;

(ii) the mean heat output, $\bar{y}$.

c. Plot the point $(\bar{x}, \bar{y})$ on your scatter diagram and label this point M . [2]

d. Write down the product-moment correlation coefficient, r . [2]

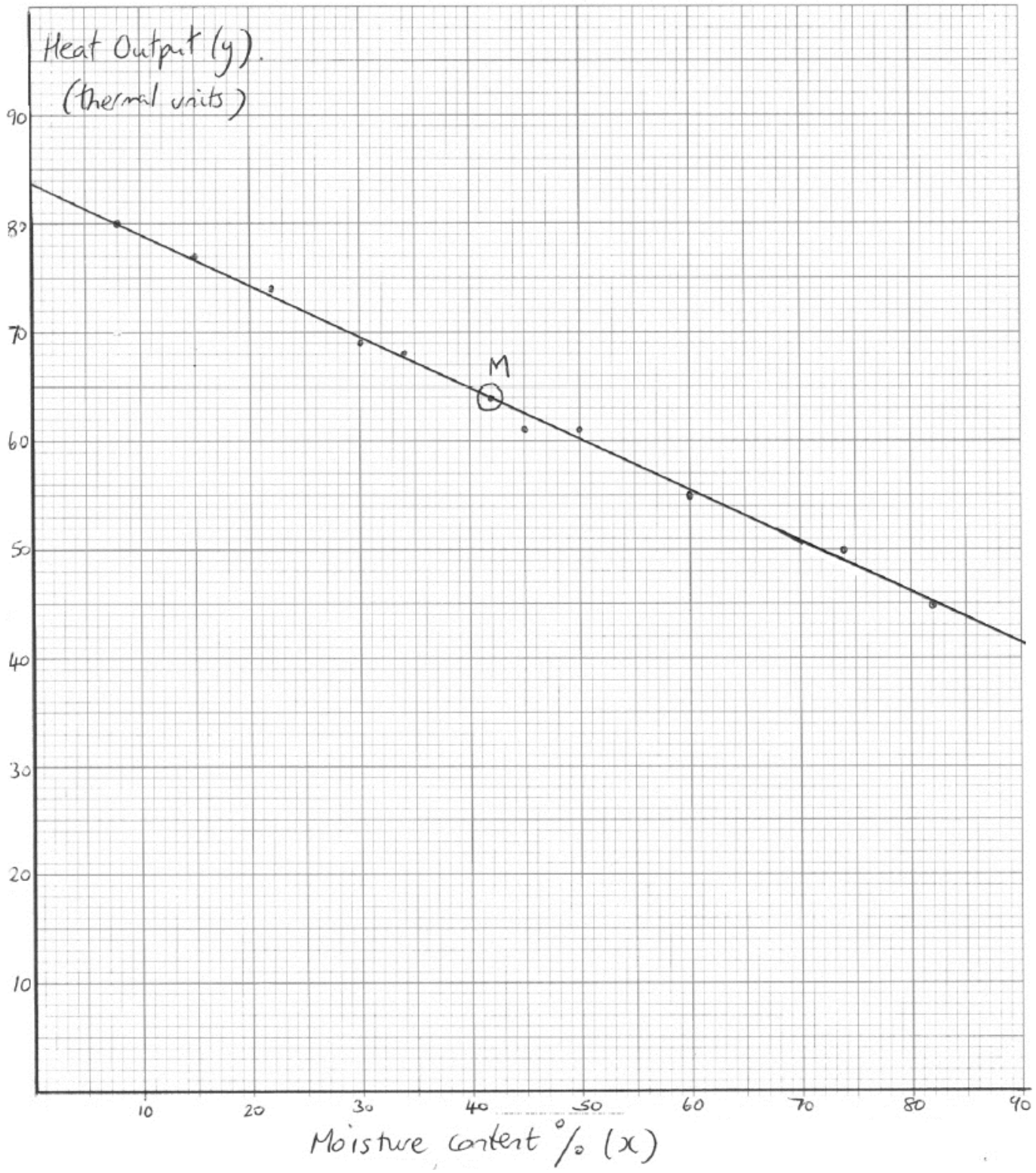
e. The equation of the regression line y on x is $y = -0.470x + 83.7$. Draw the regression line y on x on your scatter diagram. [2]

f. The equation of the regression line y on x is $y = -0.470x + 83.7$. Estimate the heat output in thermal units of a 1 kg block of wood that has 25% moisture content. [2]

g. The equation of the regression line y on x is $y = -0.470x + 83.7$. State, with a reason, whether it is appropriate to use the regression line y on x to estimate the heat output in part (f). [2]

Markscheme

a.



(A1) for correct scales and labels
(A3) for all ten points plotted correctly
(A2) for eight or nine points plotted correctly
(A1) for six or seven points plotted correctly ***(A4)***

Note: Award at most ***(A0)(A3)*** if axes reversed.

[4 marks]

b. (i) $\bar{x} = 42$ ***(A1)***

(ii) $\bar{y} = 64$ ***(A1)***

[2 marks]

c. $(\bar{x}, \bar{y})$ plotted on graph and labelled, M ***(A1)*****(ft)*****(A1)***

Note: Award ***(A1)*****(ft)** for position, ***(A1)*** for label.

[2 marks]

d. -0.998 ***(G2)***

Note: Award ***(G1)*** for correct sign, ***(G1)*** for correct absolute value.

[1 mark]

e. line on graph ***(A1)*****(ft)*****(A1)***

Notes: Award ***(A1)*****(ft)** for line through their M, **(A1)** for approximately correct intercept (allow between 83 and 85). It is not necessary that the line is seen to intersect the y-axis. The line must be straight for any mark to be awarded.

[2 marks]

f. $y = -0.470(25) + 83.7$ ***(M1)***

Note: Award ***(M1)*** for substitution into formula or some indication of method on their graph. $y = -0.470(0.25) + 83.7$ is incorrect.

$= 72.0$ (accept 71.95 and 72) ***(A1)*****(ft)*****(G2)***

Note: Follow through from graph only if they show working on their graph. Accept 72 ± 0.5 .

[2 marks]

g. Yes since 25% lies within the data set and r is close to -1 ***(R1)(A1)***

Note: Accept Yes, since r is close to -1

Note: Do not award ***(R0)(A1)***.

[2 marks]

Examiners report

a. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

b. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

c. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the

validity of using the regression line y on x .

d. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

e. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

f. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

g. The great majority of candidates found this question to be a good start to the paper. The common errors were (1) incorrect scales being used; SI units are standard in this course and candidates are expected to know the difference between centimetres and millimetres (2) the lack of r on the GDC (3) not knowing that the regression line y on x passes through the mean point and (4) not realising that the value of r determines the validity of using the regression line y on x .

In an environmental study of plant diversity around a lake, a biologist collected data about the number of different plant species (y) that were growing at different distances (x) in metres from the lake shore.

Distance (x)	2	5	8	10	13	17	23	35	40
Plant species (y)	35	34	30	29	24	19	15	13	8

a. Draw a scatter diagram to show the data. Use a scale of 2 cm to represent 10 metres on the x-axis and 2 cm to represent 10 plant species on the y-axis. [4]

b. Using your scatter diagram, describe the correlation between the number of different plant species and the distance from the lake shore. [1]

c.i. Use your graphic display calculator to write down $\bar{x}$, the mean of the distances from the lake shore. [1]

c.ii. Use your graphic display calculator to write down $\bar{y}$, the mean number of plant species. [1]

d. Plot the point $(\bar{x}, \bar{y})$ on your scatter diagram. **Label this point M.** [2]

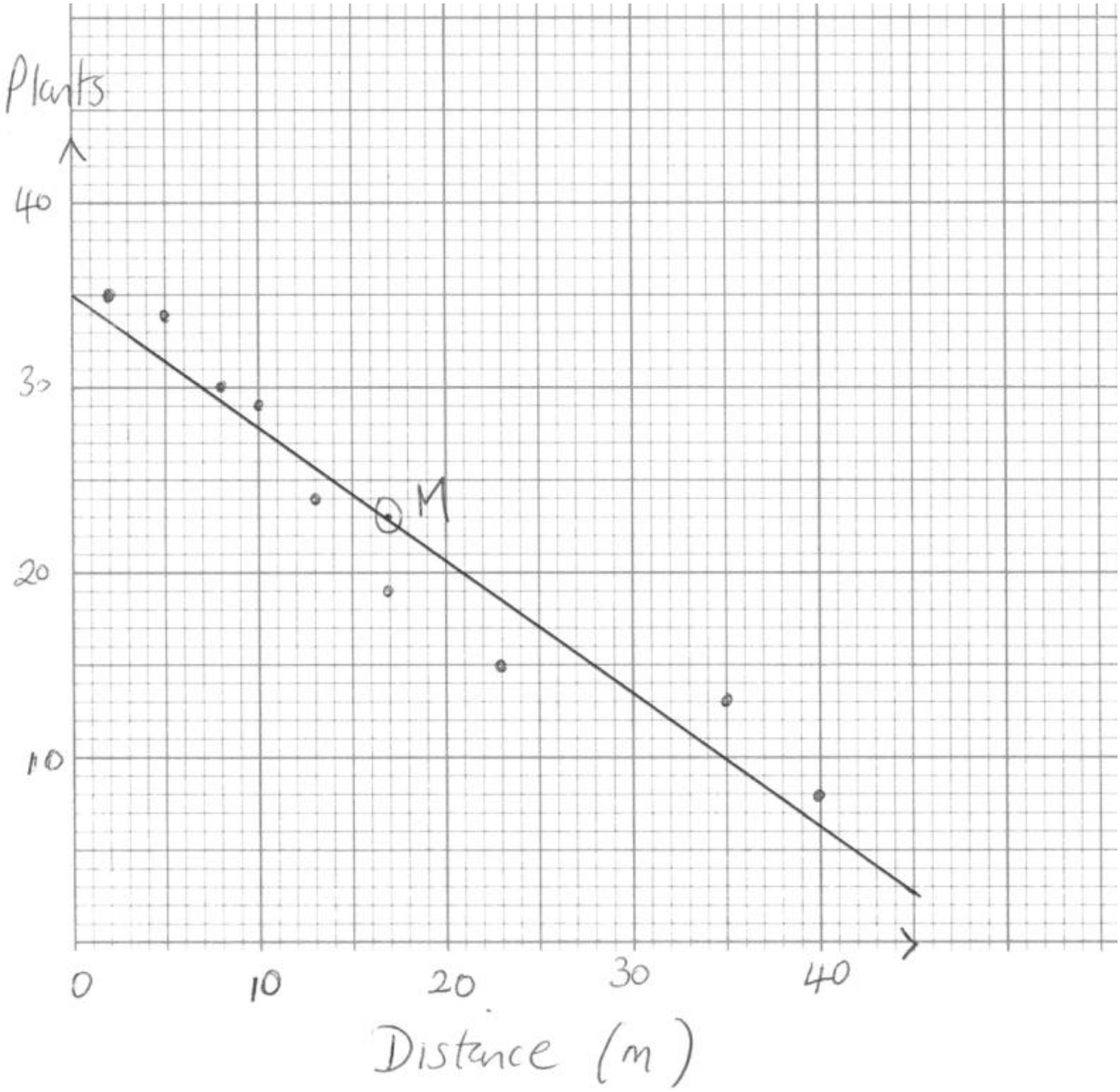
e. Write down the equation of the regression line y on x for the above data. [2]

f. Draw the regression line y on x on your scatter diagram. [2]

g. Estimate the number of plant species growing 30 metres from the lake shore. [2]

Markscheme

a.



(A1)(A3)

Notes: Award ***(A1)*** for scales and labels (accept *x*/*y*).

Award ***(A3)*** for all points correct.

Award ***(A2)*** for 7 or 8 points correct.

Award ***(A1)*** for 5 or 6 points correct.

Award at most ***(A1)(A2)*** if points are joined up.

If axes are reversed award at most ***(A0)(A3)*****(ft)**.

[4 marks]

b. Negative ***(A1)***

[1 mark]

c.i. 17 ***(G1)***

[1 mark]

c.ii. 23 ***(G1)***

[1 mark]

d. Point correctly placed and labelled M ***(A1)*****(ft)*****(A1)***

Note: Accept an error of ±0.5.

[2 marks]

e. $y = -0.708x + 35.0$ ***(G1)(G1)***

Note: Award at most ***(G1)(G0)*** if $y =$ not seen. Accept 35.

[2 marks]

f. Regression line drawn that passes through M and (0, 35) ***(A1)*****(ft)*****(A1)*****(ft)***

Note: Award ***(A1)*** for straight line that passes through M, ***(A1)*** for line (extrapolated if necessary) that passes through (0, 35) (accept error of ±1). If ruler not used, award a maximum of ***(A1)(A0)***.

[2 marks]

g. $y = -0.708(30) + 35.0$ ***(M1)***

$= 14$ *(Accept 13)* ***(A1)*****(ft)*****(G2)***

OR

Using graph: ***(M1)*** for some indication on graph of point, ***(A1)*****(ft)** for answers. Final answer must be consistent with their graph. ***(M1)(A1)*****(ft)*****(G2)***

Note: The final answer must be an integer.

[2 marks]

Examiners report

a. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its y intercept.

b. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

c.i. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

c.ii. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

d. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

e. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

f. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

g. This question, by far, was the most accessible to the great majority of candidates. However, far too many candidates do not (1) use the scale as required by the question, (2) use a scale at all, (3) either draw or label axes, (4) use a ruler at all (5) use the provided graph paper. Accurate plotting of points can not be assessed unless graph paper has been used; the diagram is not a graph.

Many candidates did not seem aware that the regression line must pass through the mean point. Others, though they had obtained the equation of the regression line, did not use it to identify its *y* intercept.

George leaves a cup of hot coffee to cool and measures its temperature every minute. His results are shown in the table below.

Time, t (minutes)	0	1	2	3	4	5	6
Temperature, y (°C)	94	54	34	24	k	16.5	15.25

a. Write down the decrease in the temperature of the coffee [3]

(i) during the first minute (between $t = 0$ and $t = 1$) ;

(ii) during the second minute;

(iii) during the third minute.

b. Assuming the pattern in the answers to part (a) continues, show that $k = 19$. [2]

c. Use the **seven** results in the table to draw a graph that shows how the temperature of the coffee changes during the first six minutes. [4]

Use a scale of 2 cm to represent 1 minute on the horizontal axis and 1 cm to represent 10 °C on the vertical axis.

d. The function that models the change in temperature of the coffee is $y = p\,(2^{-t}) + q$. [2]

(i) Use the values $t = 0$ and $y = 94$ to form an equation in p and q.

(ii) Use the values $t = 1$ and $y = 54$ to form a second equation in p and q.

e. Solve the equations found in part (d) to find the value of p and the value of q. [2]

f. The graph of this function has a horizontal asymptote. [2]

Write down the equation of this asymptote.

g. George decides to model the change in temperature of the coffee with a linear function using correlation and linear regression. [4]

Use the **seven** results in the table to write down

(i) the correlation coefficient;

(ii) the equation of the regression line y on t.

h. Use the equation of the regression line to estimate the temperature of the coffee at $t = 3$. [2]

i. Find the percentage error in this estimate of the temperature of the coffee at $t = 3$. [2]

Markscheme

a. (i) 40

(ii) 20

(iii) 10 ***(A3)***

Notes: Award ***(A0)(A1)*****(ft)*****(A1)*****(ft)** for –40, –20, –10.

Award ***(A1)(A0)(A1)*****(ft)** for 40, 60, 70 seen.

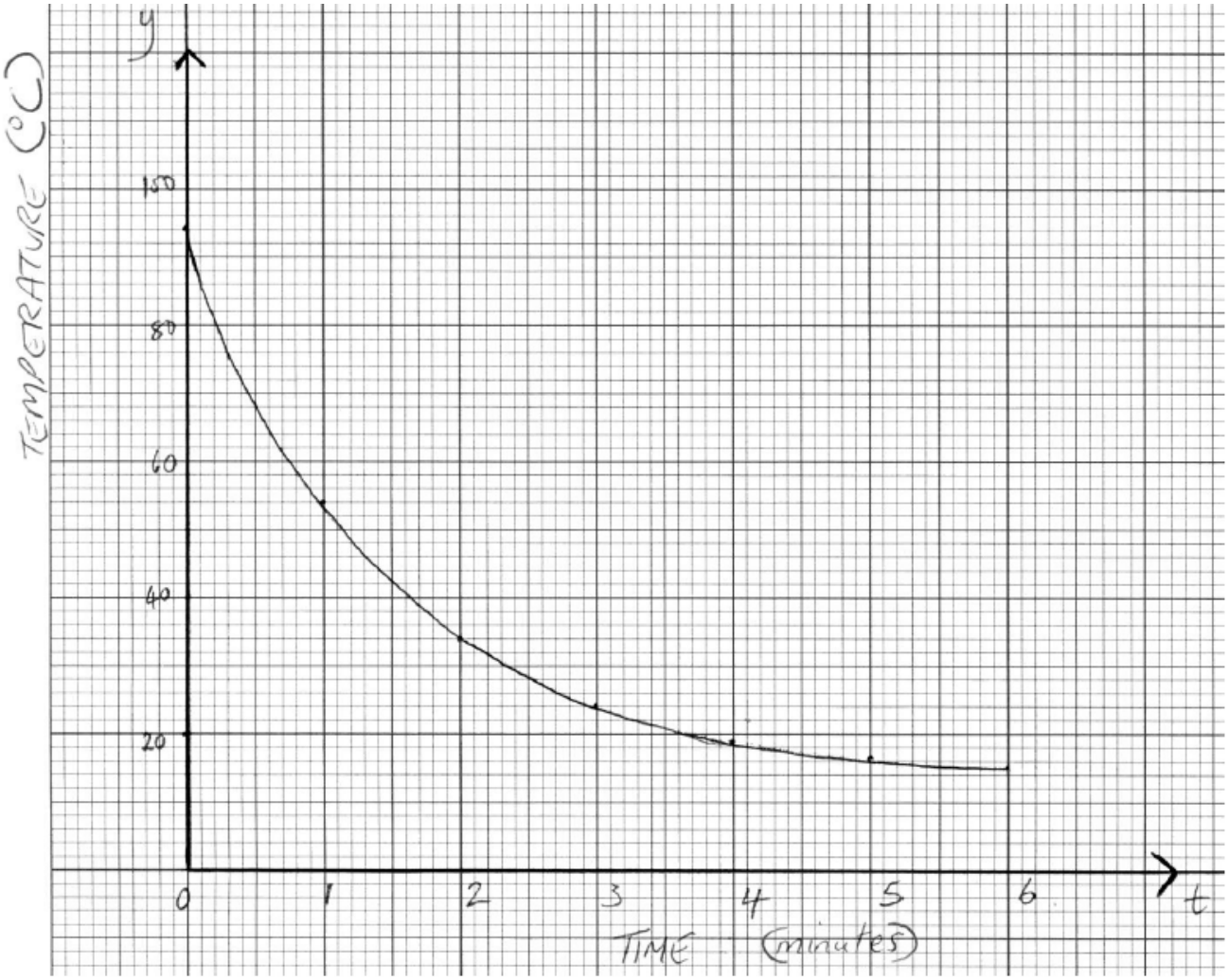
Award ***(A0)(A0)(A1)*****(ft)** for –40, –60, –70 seen.

b. $24 - k = 5$ or equivalent ***(A1)(M1)***

Note: Award ***(A1)*** for 5 seen, ***(M1)*** for difference from 24 indicated.

$k = 19$ ***(AG)***

Note: If 19 is not seen award at most ***(A1)(M0)***.

c. 

(A1)(A1)(A1)(A1)

Note: Award ***(A1)*** for scales and labelled axes (*t* or "time" and *y* or "temperature").

Accept the use of *x* on the horizontal axis only if "time" is also seen as the label.

Award ***(A2)*** for all seven points accurately plotted, award ***(A1)*** for **5 or 6** points accurately plotted, award ***(A0)*** for 4 points or fewer accurately plotted.

Award ***(A1)*** for smooth curve that passes through all points on domain [0, 6].

If graph paper is not used or one or more scales is missing, award a maximum of ***(A0)(A0)(A0)(A1)***.

d. (i) $94 = p + q$ ***(A1)***

(ii) $54 = 0.5p + q$ ***(A1)***

Note: The equations need not be simplified; accept, for example $94 = p(2^{-0}) + q$.

e. $p = 80$, $q = 14$ ***(G1)(G1)*****(ft)**

Note: If the equations have been incorrectly simplified, follow through even if no working is shown.

f. $y = 14$ ***(A1)(A1)*****(ft)**

Note: Award ***(A1)*** for y = a constant, ***(A1)*** for their 14. Follow through from part (e) only if their q lies between 0 and 15.25 inclusive.

g. (i) –0.878 (–0.87787...) ***(G2)***

Note: Award ***(G1)*** if –0.877 seen only. If negative sign omitted award a maximum of ***(A1)(A0)***.

(ii) $y = -11.7t + 71.6$ ($y = -11.6517...t + 71.6336...$) ***(G1)(G1)***

Note: Award ***(G1)*** for $-11.7t$, ***(G1)*** for 71.6.

If $y =$ is omitted award at most ***(G0)(G1)***.

If the use of x in part (c) has **not** been penalized (the axis has been labelled "time") then award at most ***(G0)(G1)***.

h. $-11.6517...(3) + 71.6339...$ ***(M1)***

Note: Award ***(M1)*** for correct substitution in their part (g)(ii).

= 36.7 (36.6785...) ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (g). Accept 36.5 for use of the 3sf answers from part (g).

i. $\frac{36.6785...-24}{24} \times 100$ ***(M1)***

Note: Award ***(M1)*** for their correct substitution in percentage error formula.

= 52.8% (52.82738...) ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (h). Accept 52.1% for use of 36.5.

Accept 52.9 % for use of 36.7. If partial working ($\times 100$ omitted) is followed by their correct answer award ***(M1)(A1)***. If partial working is followed by an incorrect answer award ***(M0)(A0)***. The percentage sign is not required.

Examiners report

a. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

b. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

c. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

d. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

e. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

f. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

g. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

h. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

i. Almost all candidates were able to score on the first parts of this question; errors occurring only when insufficient care was taken in reading what the question was asking for. The graph was usually well drawn, other than for those who have no idea what centimetres are.

The majority were able to determine the simultaneous equations, if only in unsimplified form; there was less success in solving these – though this is easily done via the GDC (the preferred approach) and the equation of the asymptote proved a discriminating task. The final parts, involving correlation and regression were largely independent of the previous parts and were accessible to most. Hopefully, contrasting the large percentage error with the value of the correlation coefficient will be valuable in class discussions. Given the many scripts that gave the value of the coefficient of determination as that of r , it seems better that the former is simply not taught.

A biologist is studying the relationship between the number of chirps of the Snowy Tree cricket and the air temperature. He records the chirp rate, x, of a cricket, and the corresponding air temperature, T, in degrees Celsius.

The following table gives the recorded values.

Cricket's chirp rate, x, (chirps per minute)	20	40	60	80	100	120
Temperature, T (°C)	8.0	12.8	15.0	18.2	20.0	21.1

a. Draw the scatter diagram for the above data. Use a scale of 2 cm for 20 chirps on the horizontal axis and 2 cm for 4°C on the vertical axis. [4]

b. Use your graphic display calculator to write down the Pearson's product–moment correlation coefficient, r, between x and T. [2]

c. Interpret the relationship between x and T using your value of r. [2]

d. Use your graphic display calculator to write down the equation of the regression line T on x. Give the equation in the form $T = ax + b$. [2]

e. Calculate the air temperature when the cricket's chirp rate is 70. [2]

f. Given that $\bar{x} = 70$, draw the regression line T on x on your scatter diagram. [2]

g. A forest ranger uses her own formula for estimating the air temperature. She counts the number of chirps in 15 seconds, z, multiplies this number by 0.45 and then she adds 10. [1]

Write down the formula that the forest ranger uses for estimating the temperature, T.

Give the equation in the form $T = mz + n$.

h. A cricket makes 20 chirps in **15** seconds. [6]

For this chirp rate

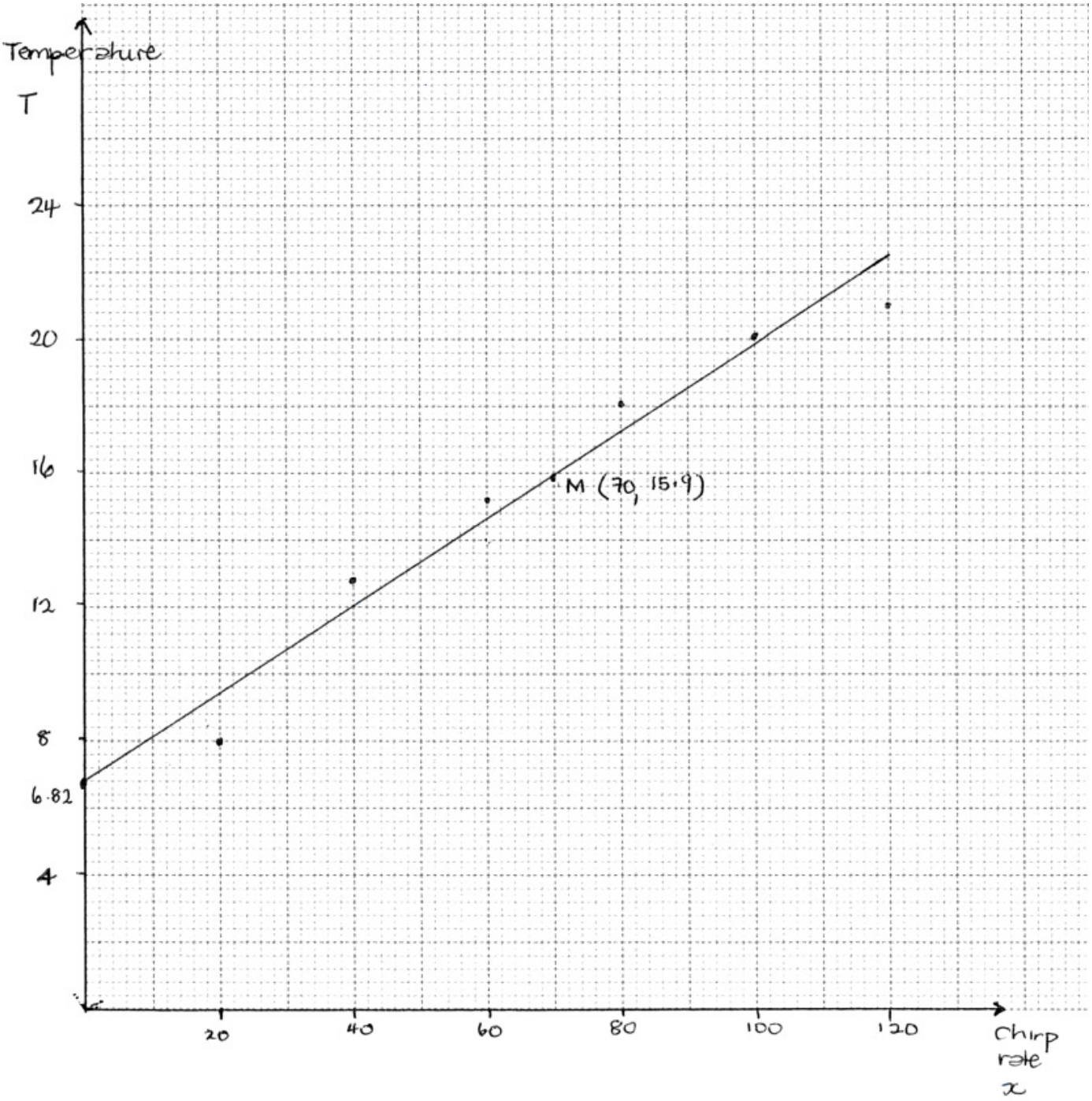
(i) calculate an estimate for the temperature, T, **using the forest ranger's formula**;

(ii) determine the actual temperature recorded by the biologist, **using the table above**;

(iii) calculate the percentage error in the forest ranger's estimate for the temperature, compared to the actual temperature recorded by the biologist.

Markscheme

a.



(A4)

Notes: Award ***(A1)*** for correct scales and labels.

Award ***(A3)*** for all six points correctly plotted,

(A2) for four or five points correctly plotted,

(A1) for two or three points correctly plotted.

Award at most ***(A0)(A3)*** if axes reversed.

Accept tolerance for T-axis.

b. $0.977 \quad (0.977324\ldots)$ ***(G2)***

Notes: Award ***(G1)*** for 0.97.

c. (Very) strong positive correlation ***(A1)*****(ft)*****(A1)*****(ft)**

Notes: Award ***(A1)*** for (very) strong, ***(A1)*** for positive.

Follow through from part (b).

d. $T = 0.129x + 6.82$ ***(G2)***

Notes: Award ***(G1)*** for $0.129x$, ***(G1)*** for $+6.82$.

Award a maximum of ***(G0)(G1)*** if the answer is not an equation.

e. $0.129 \times 70 + 6.82$ ***(M1)***

Note: Award ***(M1)*** for substitution of 70 into their equation of regression line.

OR

$\frac{8+12.8+\ldots+21.1}{6}$ ***(M1)***

$= 15.9\ (15.85)$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (d) without working.

f. regression line through $(70,\ 15.9)$ ***(A1)*****(ft)**

Note: Accept 15.9 ± 0.2.

Follow through from part (e).

with T-intercept, 6.82 ***(A1)*****(ft)**

Note: Follow through from part (d). Accept 6.82 ± 0.2.

In case the regression line is not straight (ruler not used), award ***(A0)(A1)*****(ft)** if line passes through both their $(70,\ 15.9)$ and $(0,\ 6.82)$, otherwise award ***(A0)(A0)***.

Do not penalize if line does not intersect the T-axis.

g. $T = 0.45z + 10$ ***(A1)***

h. (i) $0.45(20) + 10$ ***(M1)***

Note: Award ***(M1)*** for correct substitution of 20 into their formula from part (g).

$= 19 \quad (^\circ\text{C})$ ***(A1)*****(ft)*****(G2)***

Note: Follow through from part (g).

(ii) $= 18.2 \quad (^\circ\text{C})$ ***(A1)***

(iii) $\left|\frac{19-18.2}{18.2}\right| \times 100\%$ ***(M1)(A1)*****(ft)**

Note: Award ***(M1)*** for substitution in the percentage error formula, ***(A1)*** for correct substitution.

$4.40\% \quad (4.39560\ldots)$ ***(A1)*****(ft)*****(G2)***

Notes: Follow through from parts (h)(i) and (h)(ii).

Examiners report

a. [N/A]
[N/A]

b. [N/A]
d. [N/A]
e. [N/A]
f. [N/A]
g. [N/A]
h. [N/A]