2.2 Statistics: Data Presentation and Interpretation

EDEXCEL 9MA0 · Calculator allowed · about 225 minutes · Write-in edition

Total Marks

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

1 House prices, in thousands of pounds, were recorded for two streets. For 20 houses in Elm Street: sum(x) = 4200, sum(x^2) = 920000. For 15 houses in Birch Street: sum(x) = 3300, sum(x^2) = 746000.

(a) Calculate the mean and the standard deviation of house prices in Elm Street. **(4)**

(b) Calculate the mean and the standard deviation of house prices in Birch Street. **(4)**

(c) Using your answers to parts (a) and (b), compare house prices in Elm Street and Birch Street. **(2)**

(Total for Question 1 is 10 marks)

---

2 The masses, in kg, of a sample of lambs have lower quartile 34, median 42, upper quartile 49, smallest value 29 and largest value 94.

(a) Find the interquartile range. **(2)**

(b) A value is an outlier if it is more than 1.5 x IQR above the upper quartile or below the lower quartile. Determine whether the smallest value and the largest value are outliers. **(3)**

(c) Describe the skewness of the distribution, justifying your answer using the quartiles. **(2)**

(Total for Question 2 is 7 marks)

3 A histogram is drawn for the lengths of time, in minutes, that customers wait. The class widths and frequencies are:

Class: 10 - 15, 15 - 17, 17 - 22, 22 - 24

Frequency: 30, 7, 8, 24

(a) Find the frequency density for each class. **(2)**

(b) Estimate the number of customers who waited longer than 16 minutes but not more than 17 minutes. **(2)**

(c) The height of the tallest bar is 6 cm. Find the height of the bar for the class with the smallest frequency density, in cm to 2 significant figures. **(2)**

(Total for Question 3 is 6 marks)

4 The table summarises the times, in minutes, taken by 42 people to complete a puzzle.
Time: 10 - 15, 15 - 20, 20 - 25, 25 - 30, 30 - 35
Frequency: 2, 15, 10, 12, 3

Time (minutes)	Frequency
10 - 15	2
15 - 20	15
20 - 25	10
25 - 30	12
30 - 35	3

(a) Estimate the mean time. **(3)**

(b) Estimate the standard deviation of the times, to 3 significant figures. **(3)**

(c) Use linear interpolation to estimate the median time, to 3 significant figures. **(3)**

(Total for Question 4 is 9 marks)

5 The masses, in kg, of a sample of lambs have lower quartile 39, median 48, upper quartile 55, smallest value 30 and largest value 76.

(a) Find the interquartile range. **(2)**

(b) A value is an outlier if it is more than 1.5 x IQR above the upper quartile or below the lower quartile. Determine whether the smallest value and the largest value are outliers. **(3)**

(c) Describe the skewness of the distribution, justifying your answer using the quartiles. **(2)**

(Total for Question 5 is 7 marks)

6 The masses, in kg, of a sample of lambs have lower quartile 33, median 41, upper quartile 52, smallest value 28 and largest value 89.

(a) Find the interquartile range. **(2)**

(b) A value is an outlier if it is more than 1.5 x IQR above the upper quartile or below the lower quartile. Determine whether the smallest value and the largest value are outliers. **(3)**

(c) Describe the skewness of the distribution, justifying your answer using the quartiles. **(2)**

(Total for Question 6 is 7 marks)

7 The table summarises the times, in minutes, taken by 30 people to complete a puzzle.

Time: 20 - 30, 30 - 40, 40 - 50, 50 - 60, 60 - 70

Frequency: 6, 8, 9, 2, 5

Time (minutes)	Frequency
20 - 30	6
30 - 40	8
40 - 50	9
50 - 60	2
60 - 70	5

(a) Estimate the mean time. **(3)**

(b) Estimate the standard deviation of the times, to 3 significant figures. **(3)**

(c) Use linear interpolation to estimate the median time, to 3 significant figures. **(3)**

(Total for Question 7 is 9 marks)

8 A histogram is drawn for the lengths of time, in minutes, that customers wait. The class widths and frequencies are:

Class: 0 - 2, 2 - 7, 7 - 11, 11 - 13

Frequency: 24, 14, 23, 29

(a) Find the frequency density for each class. **(2)**

(b) Estimate the number of customers who waited longer than 4.5 minutes but not more than 7 minutes. **(2)**

(c) The height of the tallest bar is 6 cm. Find the height of the bar for the class with the smallest frequency density, in cm to 2 significant figures. **(2)**

(Total for Question 8 is 6 marks)

9 A sample of 9 values of x has been coded using $y = (x - 62)/10$. The coded data give Sum(y) = 18 and Sum(y^2) = 124.

(a) Find the mean and the standard deviation of the coded values y. **(4)**

(b) Hence find the mean and the standard deviation of x. **(4)**

(c) Find Sum(x). **(2)**

(Total for Question 9 is 10 marks)

10 A scientist records data on height above sea level, h hundred metres, and the air temperature, T degrees Celsius, for several days. The regression line of T on h is T = 37 - 2.5h. The product moment correlation coefficient is r = -0.33.

(a) Interpret the value of the gradient of the regression line in context. **(1)**

(b) Use the regression line to estimate T when h = 27, and comment on the reliability of this estimate. **(3)**

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.33. **(2)**

(Total for Question 10 is 6 marks)

11 A sample of 30 values of x has been coded using $y = (x - 61)/5$. The coded data give $\text{Sum}(y) = 87$ and $\text{Sum}(y^2) = 581$.

(a) Find the mean and the standard deviation of the coded values y. **(4)**

(b) Hence find the mean and the standard deviation of x. **(4)**

(c) Find Sum(x). **(2)**

(Total for Question 11 is 10 marks)

12 A scientist records data on height above sea level, h hundred metres, and the air temperature, T degrees Celsius, for several days. The regression line of T on h is T = 33 - 1.5h. The product moment correlation coefficient is r = -0.78.

(a) Interpret the value of the gradient of the regression line in context. **(1)**

(b) Use the regression line to estimate T when h = 16, and comment on the reliability of this estimate. **(3)**

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.78. **(2)**

(Total for Question 12 is 6 marks)

13 A histogram is drawn for the lengths of time, in minutes, that customers wait. The class widths and frequencies are:

Class: 10 - 15, 15 - 17, 17 - 21, 21 - 23

Frequency: 11, 11, 28, 10

(a) Find the frequency density for each class. **(2)**

(b) Estimate the number of customers who waited longer than 16 minutes but not more than 17 minutes. **(2)**

(c) The height of the tallest bar is 6 cm. Find the height of the bar for the class with the smallest frequency density, in cm to 2 significant figures. **(2)**

(Total for Question 13 is 6 marks)

14 The table summarises the times, in minutes, taken by 53 people to complete a puzzle.
Time: 30 - 40, 40 - 50, 50 - 60, 60 - 70, 70 - 80, 80 - 90
Frequency: 4, 8, 14, 13, 9, 5

Time (minutes)	Frequency
30 - 40	4
40 - 50	8
50 - 60	14
60 - 70	13
70 - 80	9
80 - 90	5

(a) Estimate the mean time. **(3)**

(b) Estimate the standard deviation of the times, to 3 significant figures. **(3)**

(c) Use linear interpolation to estimate the median time, to 3 significant figures. **(3)**

(Total for Question 14 is 9 marks)

15 The typing speeds, in words per minute, of 11 trainees on a data entry course are: 42, 45, 47, 48, 50, 51, 53, 55, 58, 60, 90.

(a) Calculate the mean and the standard deviation of these 11 typing speeds. **(4)**

(b) Show that the value 90 is an outlier, using the rule that a value is an outlier if it lies more than 2 standard deviations from the mean. **(3)**

(c) Given that, without the outlier, the sum of the remaining 10 typing speeds is 509 and the sum of their squares is 26201, calculate the new mean and new standard deviation for the remaining 10 trainees. **(4)**

(d) Comment on the effect that removing the outlier has had on the mean and on the standard deviation of the typing speeds. **(2)**

(Total for Question 15 is 13 marks)

---

16 In a mathematics exam, marks x are coded using $y = (x - 30)/5$. For 15 students, the coded marks y satisfy: sum of $y = 24$, sum of $y^2 = 140$.

(a) Calculate the mean of the coded marks, y. **(1)**

.....

(b) Hence calculate the mean exam mark, x. **(2)**

.....

(c) Calculate the standard deviation of the coded marks, y, and hence calculate the standard deviation of the original exam marks, x. **(4)**

.....

(Total for Question 16 is 7 marks)

17 The heights, in cm, of 50 sunflower plants grown in a trial were recorded in the grouped frequency table below.

Height (cm)	150-160	160-170	170-180	180-190	190-200
Frequency	5	12	20	9	4

(a) Using the midpoint of each class, calculate an estimate of the mean height. **(3)**

(b) Using linear interpolation, estimate the median height. **(3)**

(c) Using your answer to part (a), calculate an estimate of the variance of the heights. **(3)**

(Total for Question 17 is 9 marks)

18 The time, t minutes, taken by 100 children to complete a jigsaw puzzle was recorded in the grouped frequency table below.

Time t (minutes)	0≤t<10	10≤t<20	20≤t<35	35≤t<50	50≤t<70
Frequency	12	28	f	15	10

(a) Given that a total of 100 children completed the puzzle, calculate the value of f. **(2)**

(b) On a histogram of this data, calculate the frequency density that should be used for the bar representing 20≤t<35. **(2)**

(c) Using linear interpolation, estimate the median time taken to complete the puzzle. **(3)**

(d) Assuming the times are evenly spread within the class 50≤t<70, estimate the number of children who took more than 60 minutes to complete the puzzle. **(3)**

(Total for Question 18 is 10 marks)

19 The weekly hours worked by 13 part-time staff at a supermarket, listed in ascending order, are:
12, 15, 15, 18, 20, 21, 22, 24, 25, 27, 29, 32, 55.

(a) Find the median, the lower quartile (Q1) and the upper quartile (Q3) of these weekly hours. **(3)**

(b) Show that 55 hours is an outlier, using the rule that a value is an outlier if it lies more than 1.5 x IQR beyond the nearer quartile. **(3)**

(c) State the five figures needed to draw a box plot for this data, and explain how the outlier at 55 should be represented on the diagram. **(3)**

(d) Using your values for the median, Q1 and Q3, comment on the skewness of the distribution of weekly hours. **(2)**

(Total for Question 19 is 11 marks)

20 For 8 days last spring, a shopkeeper recorded the number of hours of sunshine, x, and the shop's ice cream sales, y, in hundreds of pounds.

Hours of sunshine, x: 2, 3, 5, 6, 7, 8, 9, 10

Ice cream sales, y (GBP hundreds): 1.2, 1.8, 2.6, 3.0, 3.4, 4.0, 4.5, 4.9

You may use: $\text{sum}(x) = 50$, $\text{sum}(y) = 25.4$, $\text{sum}(xy) = 184.1$, $\text{sum}(x^2) = 368$, $\text{sum}(y^2) = 92.26$

(a) Calculate the value of the product moment correlation coefficient (PMCC), r, for these data. **(3)**

(b) Interpret the value of r found in part (a), in context. **(1)**

(c) Find the equation of the regression line of y on x in the form $y = a + bx$, giving a and b to 3 significant figures. **(4)**

(d) Use your regression line to estimate the ice cream sales on a day with 6.5 hours of sunshine, and comment on the reliability of this estimate. **(3)**

(Total for Question 20 is 11 marks)

---

21 A distribution of delivery times has mean = 45, median = 50, mode = 58 and standard deviation = 12. It is also known that Q1 = 38, Q2 = 50 and Q3 = 54.

(a) Calculate Pearson's coefficient of skewness, given by (mean - mode)/standard deviation. **(2)**

(b) Calculate the quartile coefficient of skewness, given by (Q3 + Q1 - 2Q2)/(Q3 - Q1). **(2)**

(c) Interpret the values found in parts (a) and (b) in terms of the skewness of the distribution. **(2)**

(d) State one advantage of using the quartile coefficient of skewness, found in part (b), rather than Pearson's coefficient of skewness, found in part (a). **(1)**

(Total for Question 21 is 7 marks)

Mark scheme · 2.2 Statistics: Data Presentation and Interpretation

Question 1

(a) attempt mean = sum(x)/20	**M1**
(a) cao mean = 210 (GBP thousand)	**A1**
(a) attempt $\sqrt{\text{sum}(x^2)/20 - \text{mean}^2}$	**M1**
(a) awrt 43.6 (GBP thousand)	**A1**
(a) Answer: mean = £210,000, SD = awrt £43,600	
(b) attempt mean = sum(x)/15	**M1**
(b) cao mean = 220 (GBP thousand)	**A1**
(b) attempt $\sqrt{\text{sum}(x^2)/15 - \text{mean}^2}$	**M1**
(b) awrt 36.5 (GBP thousand)	**A1**
(b) Answer: mean = £220,000, SD = awrt £36,500	
(c) Birch Street has the higher mean price (£220,000 vs £210,000), so houses there tend to be more expensive, oe	**B1**
(c) Elm Street has the higher standard deviation (£43,600 vs £36,500), so house prices there are more variable/spread out, oe	**B1**
(c) Answer: Houses in Birch Street are, on average, more expensive (higher mean), but house prices in Elm Street are more variable (higher standard deviation).	

Question 2

(a) 49 - 34	**M1**
(a) 15	**A1**
(a) Answer: 15	
(b) lower limit 34 - 1.5 x 15 = 11.5 and upper limit 49 + 1.5 x 15 = 71.5	**M1**
(b) smallest value 29 is not below 11.5, so is not an outlier	**A1**
(b) largest value 94 is above 71.5, so is an outlier	**A1**
(b) Answer: smallest is not an outlier; largest is an outlier	
(c) compares median - Q1 = 8 and Q3 - median = 7	**M1**
(c) negative skew (as the larger spread is on that side)	**A1**
(c) Answer: negative skew	

Question 3

(a) frequency density = frequency / class width	**M1**
(a) values 6, 7/2, 8/5, 12	**A1**
(a) Answer: 6, 7/2, 8/5, 12	
(b) proportion of the class 15 - 17 lying in the required interval = (17 - 16)/2	**M1**
(b) 3.5	**A1**
(b) Answer: 3.5	
(c) height proportional to frequency density: scale = 6/largest fd	**M1**
(c) 0.80 cm	**A1**
(c) Answer: 0.80 cm	

Question 4

(a) uses the midpoint of each class	**M1**
(a) sum of f x midpoint = 940	**M1**
(a) mean = 940/42 = 22.38 minutes	**A1**
(a) Answer: 22.38 minutes	
(b) sum f x midpoint^2 = 22212.5	**M1**
(b) variance = 22212.5/42 - (22.3810)^2	**M1**
(b) standard deviation = 5.29 minutes	**A1**
(b) Answer: 5.29 minutes	
(c) identifies the median class 20 - 25 (cumulative frequency reaches 27 at its upper bound)	**M1**
(c) interpolates: 20 + ((21 - 17)/10) x 5	**M1**
(c) median = 22.0 minutes	**A1**
(c) Answer: 22.0 minutes	

Question 5

(a) 55 - 39	**M1**
(a) 16	**A1**
(a) Answer: 16	
(b) lower limit 39 - 1.5 x 16 = 15 and upper limit 55 + 1.5 x 16 = 79	**M1**
(b) smallest value 30 is not below 15, so is not an outlier	**A1**
(b) largest value 76 is not above 79, so is not an outlier	**A1**
(b) Answer: smallest is not an outlier; largest is not an outlier	
(c) compares median - Q1 = 9 and Q3 - median = 7	**M1**
(c) negative skew (as the larger spread is on that side)	**A1**
(c) Answer: negative skew	

Question 6

(a) 52 - 33	**M1**
(a) 19	**A1**
(a) Answer: 19	
(b) lower limit 33 - 1.5 x 19 = 4.5 and upper limit 52 + 1.5 x 19 = 80.5	**M1**
(b) smallest value 28 is not below 4.5, so is not an outlier	**A1**
(b) largest value 89 is above 80.5, so is an outlier	**A1**
(b) Answer: smallest is not an outlier; largest is an outlier	
(c) compares median - Q1 = 8 and Q3 - median = 11	**M1**
(c) positive skew	**A1**
(c) Answer: positive skew	

Question 7

(a) uses the midpoint of each class **M1**

(a) sum of f x midpoint = 1270 **M1**

(a) mean = 1270/30 = 42.33 minutes **A1**

(a) Answer: 42.33 minutes

(b) sum f x midpoint^2 = 58950 **M1**

(b) variance = 58950/30 - (42.3333)^2 **M1**

(b) standard deviation = 13.1 minutes **A1**

(b) Answer: 13.1 minutes

(c) identifies the median class 40 - 50 (cumulative frequency reaches 23 at its upper bound) **M1**

(c) interpolates: 40 + ((15 - 14)/9) x 10 **M1**

(c) median = 41.1 minutes **A1**

(c) Answer: 41.1 minutes

Question 8

(a) frequency density = frequency / class width **M1**

(a) values 12, 14/5, 23/4, 29/2 **A1**

(a) Answer: 12, 14/5, 23/4, 29/2

(b) proportion of the class 2 - 7 lying in the required interval = (7 - 4.5)/5 **M1**

(b) 7 **A1**

(b) Answer: 7

(c) height proportional to frequency density: scale = 6/largest fd **M1**

(c) 1.2 cm **A1**

(c) Answer: 1.2 cm

Question 9

(a) mean of y = 18/9 **M1**

(a) 2.000 **A1**

(a) standard deviation = $\sqrt{124/9 - (2.0000)^2}$ **M1**

(a) 3.127 **A1**

(a) Answer: mean 2.000, sd 3.127

(b) x = 10y + 62 so mean of x = 10 x 2.0000 + 62 **M1**

(b) 82.00 **A1**

(b) standard deviation of x = 10 x 3.1269 **M1**

(b) 31.27 **A1**

(b) Answer: mean 82.00, sd 31.27

(c) n x mean of x = 9 x 82.0000 **M1**

(c) 738.00 **A1**

(c) Answer: 738.00

Question 10

(a) each extra hundred metre in h is associated with a change of -2.5 in T, so as h increases T decreases **B1**

(a) Answer: gradient -2.5: T decreases by 2.5 for each unit increase in h

(b) substitutes h = 27 **M1**

(b) T = -30.5 **A1**

(b) The weak correlation makes the prediction unreliable. The observed range is not given, so it is not possible to determine whether this is interpolation. Accept equivalent comments. **B1**

(b) Answer: -30.5 degrees Celsius. The weak correlation makes the prediction unreliable. The observed range is not given, so it is not possible to determine whether this is interpolation.

(c) weak **B1**

(c) negative correlation **B1**

(c) Answer: weak negative linear correlation

Question 11

(a) mean of y = 87/30 **M1**

(a) 2.900 **A1**

(a) standard deviation = $\sqrt{581/30 - (2.9000)^2}$ **M1**

(a) 3.310 **A1**

(a) Answer: mean 2.900, sd 3.310

(b) x = 5y + 61 so mean of x = 5 x 2.9000 + 61 **M1**

(b) 75.50 **A1**

(b) standard deviation of x = 5 x 3.3101 **M1**

(b) 16.55 **A1**

(b) Answer: mean 75.50, sd 16.55

(c) n x mean of x = 30 x 75.5000 **M1**

(c) 2265.0 **A1**

(c) Answer: 2265.0

Question 12

(a) each extra hundred metre in h is associated with a change of -1.5 in T, so as h increases T decreases **B1**

(a) Answer: gradient -1.5: T decreases by 1.5 for each unit increase in h

(b) substitutes h = 16 **M1**

(b) T = 9.00 **A1**

(b) The strong negative correlation supports the prediction, but reliability also depends on whether h = 16 lies within the observed range, which is not given. Accept equivalent comments. **B1**

(b) Answer: 9 degrees Celsius. The strong negative correlation supports the prediction, but reliability also depends on whether h = 16 lies within the observed range, which is not given.

(c) strong **B1**

(c) negative correlation **B1**

(c) Answer: strong negative linear correlation

Question 13

(a) frequency density = frequency / class width **M1**

(a) values 11/5, 11/2, 7, 5 **A1**

(a) Answer: 11/5, 11/2, 7, 5

(b) proportion of the class 15 - 17 lying in the required interval = (17 - 16)/2 **M1**

(b) 5.5 **A1**

(b) Answer: 5.5

(c) height proportional to frequency density: scale = 6/largest fd **M1**

(c) 1.9 cm **A1**

(c) Answer: 1.9 cm

Question 14

(a) uses the midpoint of each class **M1**

(a) sum of f x midpoint = 3215 **M1**

(a) mean = 3215/53 = 60.66 minutes **A1**

(a) Answer: 60.66 minutes

(b) sum f x midpoint^2 = 205125 **M1**

(b) variance = 205125/53 - (60.6604)^2 **M1**

(b) standard deviation = 13.8 minutes **A1**

(b) Answer: 13.8 minutes

(c) identifies the median class 60 - 70 (cumulative frequency reaches 39 at its upper bound) **M1**

(c) interpolates: 60 + ((26.5 - 26)/13) x 10 **M1**

(c) median = 60.4 minutes **A1**

(c) Answer: 60.4 minutes

Question 15

(a) attempt sum of the 11 values / 11 **M1**

(a) mean = awrt 54.5 **A1**

(a) attempt $\sqrt{\text{sum } x^2 / 11 - \text{mean}^2}$, using sum of squares of the 11 values **M1**

(a) SD = awrt 12.4 **A1**

(a) Answer: mean = awrt 54.5, SD = awrt 12.4 (words per minute)

(b) attempt mean + 2 x (their standard deviation) **M1**

(b) correct value: 54.4545 + 2(12.3683) = awrt 79.2 **M1**

(b) cso: since 90 > 79.2 (their boundary), 90 is an outlier; conclusion must follow correctly from their boundary **A1**

(b) Answer: 90 is an outlier (given)

(c) attempt new mean = 509/10 **M1**

(c) new mean = 50.9 (cao) **A1**

(c) attempt $\sqrt{26201/10 - (\text{their new mean})^2}$, dependent on a correct new mean **dM1**

(c) new SD = awrt 5.41 **A1**

(c) Answer: new mean = 50.9, new SD = awrt 5.41 (words per minute)

(d) the mean has decreased (from awrt 54.5 to 50.9), since the outlier of 90 was well above the rest of the data and was pulling the mean up, oe **B1**

(d) the standard deviation has decreased substantially (from awrt 12.4 to awrt 5.41), since the outlier was a long way from the mean and contributed a large amount to the spread, oe **B1**

(d) Answer: Both the mean and the standard deviation have decreased: the mean fell from awrt 54.5 to 50.9 because the high outlier was no longer pulling it up, and the standard deviation fell sharply from awrt 12.4 to awrt 5.41 because the outlier had been contributing a large deviation to the spread.

Question 16

(a) cao mean of y = 1.6 **B1**

(a) Answer: mean of y = 1.6

(b) correct rearrangement x = 5y + 30 applied to their mean of y **M1**

(b) cao mean of x = 38 **A1**

(b) Answer: mean of x = 38

(c) attempt $\sqrt{\text{sum } y^2 / 15 - (\text{mean of } y)^2}$ **M1**

(c) SD of y = awrt 2.60 **A1**

(c) correct scaling: SD of x = 5 x (their SD of y), since x = 5y + 30 **M1**

(c) SD of x = awrt 13.0 **A1**

(c) Answer: SD of y = awrt 2.60, SD of x = awrt 13.0

Question 17

(a) correct midpoints identified: 155, 165, 175, 185, 195 **M1**

(a) attempt sum of (frequency x midpoint) / 50 **M1**

(a) cao mean = 174 (cm) **A1**

(a) Answer: estimated mean = 174 cm

(b) correct cumulative frequencies identified (5, 17, 37, 46, 50) and median class 170-180 identified **M1**

(b) correct interpolation: 170 + ((25 - 17)/20) x 10 **M1**

(b) cao median = 174 (cm) **A1**

(b) Answer: estimated median = 174 cm

(c) attempt sum of (frequency x midpoint^2) **M1**

(c) correct method: (sum f x midpoint^2)/50 - (their mean)^2, dependent on a correct mean from part (a) **dM1**

(c) cao variance = 113 (cm^2) **A1**

(c) Answer: estimated variance = 113 cm^2

Question 18

(a) 100 - 12 - 28 - 15 - 10 **M1**

(a) cao f = 35 **A1**

(a) Answer: f = 35

(b) attempt frequency / class width, i.e. their f / 15 **M1**

(b) awrt 2.33 **A1**

(b) Answer: frequency density = awrt 2.33

(c) correct cumulative frequencies identified (12, 40, 75, 90, 100) and median class $20 \le t < 35$ identified **M1**

(c) correct interpolation: 20 + ((50 - 40)/35) x 15 **M1**

(c) awrt 24.3 (minutes) **A1**

(c) Answer: estimated median = awrt 24.3 minutes

(d) identifies that this requires the proportion of the $50 \le t < 70$ class above t = 60 **M1**

(d) correct method: 10 x (70 - 60)/(70 - 50) **M1**

(d) cao 5 children **A1**

(d) Answer: 5 children

Question 19

(a) median = 22 (the 7th of 13 ordered values) **B1**

(a) correct method: Q1 as the median of the lower 6 values, Q3 as the median of the upper 6 values **M1**

(a) cao Q1 = 16.5 and Q3 = 28 **A1**

(a) Answer: median = 22, Q1 = 16.5, Q3 = 28 (hours)

(b) IQR = 28 - 16.5 = 11.5 **M1**

(b) upper boundary = Q3 + 1.5 x IQR = 28 + 1.5(11.5) **M1**

(b) cso: boundary = 45.25, and since 55 > 45.25, 55 is an outlier **A1**

(b) Answer: 55 is an outlier (given)

(c) minimum = 12, Q1 = 16.5, median = 22, Q3 = 28 **B1**

(c) the upper whisker is drawn to 32, the highest value that is not an outlier (not to 55) **B1**

(c) 55 is plotted separately, e.g. as a cross or a dot beyond the whisker, to show it is an outlier **B1**

(c) Answer: Five-figure summary: min = 12, Q1 = 16.5, median = 22, Q3 = 28, with the upper whisker drawn to 32 (the largest non-outlier value); 55 is marked separately as a cross beyond the whisker to indicate it is an outlier.

(d) compares Q3 - median (= 6) with median - Q1 (= 5.5): these are close in size, using their values from part (a) **B1ft**

(d) concludes the quartiles suggest a roughly symmetrical distribution, but the presence of the high outlier (55) suggests some positive skew overall, oe **B1**

(d) Answer: Q3 - median = 6 and median - Q1 = 5.5, which are close in size, so the quartiles alone suggest a fairly symmetrical distribution; however, the presence of the high outlier at 55 suggests the overall distribution has some positive skew (a long tail towards higher values).

Question 20

(a) attempt Sxy = sum(xy) - sum(x)sum(y)/8, and Sxx = sum(x^2) - sum(x)^2/8, and Syy = sum(y^2) - sum(y)^2/8 **M1**

(a) correct formula: $r = Sxy / \sqrt{Sxx \times Syy}$ **M1**

(a) awrt 0.998 **A1**

(a) Answer: r = awrt 0.998

(b) a very strong positive (linear) correlation: more hours of sunshine are associated with higher ice cream sales, oe **B1**

(b) Answer: There is a very strong positive linear correlation: days with more hours of sunshine are strongly associated with higher ice cream sales.

(c) attempt b = Sxy / Sxx **M1**

(c) b = awrt 0.457 **A1**

(c) attempt a = (mean of y) - b(mean of x), dependent on their value of b **dM1**

(c) a = awrt 0.320 **A1**

(c) Answer: y = 0.320 + 0.457x

(d) substitute x = 6.5 into their regression equation **M1**

(d) awrt 3.29 (i.e. sales of about £329) **A1**

(d) x = 6.5 lies within the range of the given data (2 to 10 hours), so this is interpolation and the estimate is reasonably reliable, oe **B1**

(d) Answer: y = awrt 3.29 (hundreds of pounds), i.e. sales of about £329; reasonably reliable since 6.5 lies within the given data range.

Question 21

(a) substitute values into (mean - mode)/SD, i.e. (45 - 58)/12 **M1**

(a) awrt -1.08 **A1**

(a) Answer: awrt -1.08

(b) substitute values into (Q3 + Q1 - 2Q2)/(Q3 - Q1), i.e. (54 + 38 - 100)/(54 - 38) **M1**

(b) cao -0.5 **A1**

(b) Answer: -0.5

(c) both coefficients are negative, indicating the distribution is negatively skewed (a longer tail towards lower values) **B1**

(c) this is consistent with mean < median < mode, the typical pattern for a negatively skewed distribution, oe **B1**

(c) Answer: Both coefficients are negative, so the distribution of delivery times is negatively skewed, with a longer tail towards lower values. This agrees with the fact that mean (45) < median (50) < mode (58), the pattern typical of negative skew.

(d) the quartile coefficient is not affected by extreme values/outliers (or does not require the mode, which may not be well defined for continuous or multi-modal data), oe **B1**

(d) Answer: The quartile coefficient is not affected by extreme values (outliers) and does not rely on the mode, which is not always well defined for continuous data.