2.6 Statistics: Data Presentation and Interpretation Depth

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

Total Marks

Name: _____ Date: ____ / ____ / _____

Answer ALL questions. Show all your working.

A-Level Statistics: Data Presentation and Interpretation - Depth Practice

Original material written for Revision Library, styled on Edexcel AS/A-level Mathematics (9MA0) Statistics content.

This pack extends core data presentation and interpretation skills to full A-level depth: measures of location and spread (including coding), identifying outliers by the quartile rule and the standard deviation rule, box plots and skewness, histograms and frequency density, linear interpolation, correlation and regression (including coding and residuals), and evaluating statistical models in context, in the style of the large data set used across the qualification. All contexts and figures are original.

1 The times, in minutes, taken by 15 students to complete an S6 statistics quiz are recorded and arranged in ascending order:

8, 10, 11, 12, 12, 13, 14, 14, 15, 16, 16, 17, 18, 19, 31

(a) Find the lower quartile, the median and the upper quartile of this data set. **(3)**

(b) A value is defined as an outlier if it lies more than 1.5 x IQR beyond the nearer quartile. Using your answers to part (a), determine whether the value 31 is an outlier, showing your working clearly. **(3)**

(Total for Question 1 is 6 marks)

2 The masses, in kg, of a sample of lambs have lower quartile 35, median 43, upper quartile 53, smallest value 25 and largest value 92.

(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 scientist records data on the age of a car, x years, and its value, V thousand pounds, for several days. The regression line of V on x is V = 6 - 1.5x. The product moment correlation coefficient is r = -0.85.

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

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

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.85. **(2)**

(Total for Question 3 is 6 marks)

4 The table summarises the times, in minutes, taken by 57 people to complete a puzzle.

Time	Frequency
30 - 40	9
40 - 50	14
50 - 60	14
60 - 70	2
70 - 80	7
80 - 90	11

(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 34, median 39, upper quartile 42, smallest value 31 and largest value 78.

(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 A scientist records data on the age of a car, x years, and its value, V thousand pounds, for several days. The regression line of V on x is V = 39 - 1.5x. The product moment correlation coefficient is r = -0.92.

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

(b) Use the regression line to estimate V when x = 15, and comment on the reliability of this estimate. **(3)**

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.92. **(2)**

(Total for Question 6 is 6 marks)

7 The masses, in kg, of a sample of lambs have lower quartile 31, median 35, upper quartile 37, smallest value 26 and largest value 55.

(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 7 is 7 marks)

8 A sample of 23 values of x has been coded using $y = (x - 44)/10$. The coded data give $\text{Sum}(y) = 62$ and $\text{Sum}(y^2) = 440$.

(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 8 is 10 marks)

9 The masses, in kg, of a sample of lambs have lower quartile 39, median 47, upper quartile 56, smallest value 30 and largest value 101.

(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 9 is 7 marks)

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

Class: 0 - 5, 5 - 10, 10 - 15, 15 - 17

Frequency: 15, 18, 23, 10

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

(b) Estimate the number of customers who waited longer than 7.5 minutes but not more than 10 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 10 is 6 marks)

11 The table summarises the times, in minutes, taken by 71 people to complete a puzzle.

Time: 0 - 5, 5 - 10, 10 - 15, 15 - 20, 20 - 25, 25 - 30

Frequency: 11, 13, 15, 9, 13, 10

(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 11 is 9 marks)

12 A sample of 22 values of x has been coded using $y = (x - 72)/5$. The coded data give $\text{Sum}(y) = 30$ and $\text{Sum}(y^2) = 230$.

(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 12 is 10 marks)

13 A scientist records data on the age of a car, x years, and its value, V thousand pounds, for several days. The regression line of V on x is V = 35 - 0.8x. The product moment correlation coefficient is r = -0.6.

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

(b) Use the regression line to estimate V when x = 30, and comment on the reliability of this estimate. **(3)**

(c) Comment on the strength and direction of the linear correlation indicated by r = -0.6. **(2)**

(Total for Question 13 is 6 marks)

14 The table summarises the times, in minutes, taken by 34 people to complete a puzzle.

Time: 30 - 40, 40 - 50, 50 - 60, 60 - 70, 70 - 80

Frequency: 5, 3, 9, 8, 9

(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 A sample of 8 values of x has been coded using $y = (x - 37)/2$. The coded data give $\text{Sum}(y) = 3$ and $\text{Sum}(y^2) = 103$.

(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 15 is 10 marks)

16 The masses, in kg, of a sample of parcels processed at a sorting depot are summarised in the grouped frequency table below. One frequency is missing.

Mass, m (kg): $0 < m \leq 10$, frequency 8
Mass, m (kg): $10 < m \leq 20$, frequency 22
Mass, m (kg): $20 < m \leq 30$, frequency f (unknown)
Mass, m (kg): $30 < m \leq 50$, frequency 18
Mass, m (kg): $50 < m \leq 80$, frequency 12

It is known that the frequency density for the class $20 < m \leq 30$ is 2.6.

(a) Find the value of f. **(2)**

(b) Calculate the frequency density for each of the other four classes, and state, with a reason, which class would be represented by the tallest bar if a histogram were drawn for this data. **(3)**

(c) Estimate the number of parcels with mass greater than 30 kg. **(1)**

(Total for Question 16 is 6 marks)

17 The table shows the distribution of the number of hours, h, spent revising per week by 60 sixth-form students during study leave.

Hours, h: $0 \leq h < 5$, midpoint 2.5, frequency 6
Hours, h: $5 \leq h < 10$, midpoint 7.5, frequency 14
Hours, h: $10 \leq h < 15$, midpoint 12.5, frequency 20
Hours, h: $15 \leq h < 20$, midpoint 17.5, frequency 12
Hours, h: $20 \leq h < 25$, midpoint 22.5, frequency 8

A coding $y = (x - 12.5)/5$ is used, where x is the class midpoint.

(a) Complete the coded midpoint values, y, for each class. **(2)**

(b) Using the coded data, calculate an estimate for the mean number of hours spent revising per week. **(3)**

(c) Calculate an estimate for the standard deviation of the number of hours spent revising per week. **(4)**

(Total for Question 17 is 9 marks)

18 For a sample of 12 UK secondary schools, x represents the average class size and y represents the mean GCSE point score per pupil. The following summary statistics were calculated:

n = 12, Sxx = 248, Syy = 196, Sxy = -152

(a) Calculate the product moment correlation coefficient, r, between x and y. **(3)**

(b) Interpret the value of r found in part (a) in context. **(2)**

(c) The headteacher of one school claims that 'smaller class sizes cause higher GCSE point scores'. Comment on the validity of this claim, referring to the value of r found in part (a). **(2)**

(Total for Question 18 is 7 marks)

19 The number of texts sent in a day by 9 students is recorded:

22, 25, 19, 30, 27, 21, 26, 23, 95

(a) Calculate the mean and standard deviation of all 9 values. **(4)**

(b) The value 95 is identified as a recording error and removed. Calculate the new mean and standard deviation for the remaining 8 values. **(4)**

(c) Comment on the effect that removing this value has had on the mean and on the standard deviation. **(2)**

(Total for Question 19 is 10 marks)

20 The table below shows the distribution of the ages, in complete years, of 200 members of a gym. Ages are recorded to the nearest whole year, so the class 16-25 has boundaries 15.5 and 25.5, and so on.

Age (years): 16-25, frequency 30
Age (years): 26-35, frequency 54
Age (years): 36-45, frequency f (unknown)
Age (years): 46-55, frequency 38
Age (years): 56-70, frequency 20

(a) Show that the frequency, f, for the class 36-45 is 58. **(1)**

(b) Calculate the frequency density for each class (using class widths based on the true class boundaries), and state, with a reason, which class would be represented by the tallest bar if a histogram were drawn for this data. **(2)**

(c) Use linear interpolation to estimate the median age of the gym members. **(3)**

(d) Estimate the number of gym members aged under 30. **(3)**

(Total for Question 20 is 9 marks)

21 As part of a study styled on the large data set, a researcher records the daily mean windspeed, in knots, at a UK weather station for the month of October (31 days). The summary statistics are sum(x) = 217 and sum(x^2) = 1859.4 (n = 31). The five smallest values in the data set are 3, 4, 4, 5, 5 knots, and the data also include one exceptionally high reading of 42 knots, recorded during a named storm.

(a) Calculate the mean and standard deviation of the windspeed data. **(4)**

(b) Using the rule that a value is an outlier if it lies more than 2 standard deviations from the mean, determine whether the value of 42 knots should be classified as an outlier. **(3)**

(c) The researcher is deciding whether to remove the value of 42 knots before calculating summary statistics intended to describe a 'typical' October day. Discuss, with reference to the context and to your answers in parts (a) and (b), whether this value should be removed from the data set. **(6)**

(Total for Question 21 is 13 marks)

22 The daily maximum temperatures, in degrees Celsius, were recorded for a sample of 31 days in May at two UK weather stations, Aviemore and Camborne, in the style of the large data set. Summary statistics are given below.

Aviemore (n = 31): minimum = 6.1, Q1 = 9.4, median = 11.8, Q3 = 14.2, maximum = 19.5
Camborne (n = 31): minimum = 10.8, Q1 = 12.6, median = 13.9, Q3 = 15.7, maximum = 18.0

(a) Calculate the interquartile range for each location. **(2)**

(b) Using the rule that an outlier is a value more than 1.5 x IQR beyond the nearer quartile, calculate the upper outlier boundary for Aviemore, and state whether the maximum value of 19.5 degrees Celsius is an outlier. **(3)**

(c) Compare and contrast the distributions of maximum daily temperature at the two locations, referring to the median and the interquartile range. **(3)**

(d) Explain one reason, in context, why Aviemore might show a greater spread of daily maximum temperatures than Camborne. **(1)**

(Total for Question 22 is 9 marks)

Mark scheme · 2.6 Statistics: Data Presentation and Interpretation Depth

Question 1

(a) lower quartile Q1 = 12 **B1**

(a) median = 14 **B1**

(a) upper quartile Q3 = 17 **B1**

(a) Answer: Q1 = 12 minutes, median = 14 minutes, Q3 = 17 minutes

(b) IQR = Q3 - Q1 = 5 found **M1**

(b) correct boundary calculated: Q3 + 1.5 x IQR = 17 + 7.5 = 24.5 **M1**

(b) correct conclusion with comparison shown: since 31 > 24.5, 31 is an outlier **A1**

(b) Answer: 31 minutes is an outlier, since it exceeds the upper boundary of 24.5 minutes

Question 2

(a) 53 - 35 **M1**

(a) 18 **A1**

(a) Answer: 18

(b) lower limit 35 - 1.5 x 18 = 8 and upper limit 53 + 1.5 x 18 = 80 **M1**

(b) smallest value 25 is not below 8, so is not an outlier **A1**

(b) largest value 92 is above 80, so is an outlier **A1**

(b) Answer: smallest is not an outlier; largest is an outlier

(c) compares median - Q1 = 8 and Q3 - median = 10 **M1**

(c) positive skew **A1**

(c) Answer: positive skew

Question 3

(a) each extra year in x is associated with a change of -1.5 in V, so as x increases V decreases **B1**

(a) Answer: gradient -1.5: V decreases by 1.5 for each unit increase in x

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

(b) V = -18.0 **A1**

(b) reliable only if 16 is within the range of the data (interpolation); extrapolation is unreliable **B1**

(b) Answer: -18.0

(c) strong **B1**

(c) negative correlation **B1**

(c) Answer: strong negative linear correlation

Question 4

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

(a) sum of f x midpoint = 3305 **M1**

(a) mean = 3305/57 = 57.98 minutes **A1**

(a) Answer: 57.98 minutes

(b) sum f x midpoint^2 = 209025 **M1**

(b) variance = 209025/57 - (57.9825)^2 **M1**

(b) standard deviation = 17.5 minutes **A1**

(b) Answer: 17.5 minutes

(c) identifies the median class 50 - 60 (cumulative frequency reaches 37 at its upper bound) **M1**

(c) interpolates: 50 + ((28.5 - 23)/14) x 10 **M1**

(c) median = 53.9 minutes **A1**

(c) Answer: 53.9 minutes

Question 5

(a) 42 - 34 **M1**

(a) 8 **A1**

(a) Answer: 8

(b) lower limit 34 - 1.5 x 8 = 22 and upper limit 42 + 1.5 x 8 = 54 **M1**

(b) smallest value 31 is not below 22, so is not an outlier **A1**

(b) largest value 78 is above 54, so is an outlier **A1**

(b) Answer: smallest is not an outlier; largest is an outlier

(c) compares median - Q1 = 5 and Q3 - median = 3 **M1**

(c) negative skew (as the larger spread is on that side) **A1**

(c) Answer: negative skew

Question 6

(a) each extra year in x is associated with a change of -1.5 in V, so as x increases V decreases **B1**

(a) Answer: gradient -1.5: V decreases by 1.5 for each unit increase in x

(b) substitutes x = 15 **M1**

(b) V = 16.5 **A1**

(b) reliable only if 15 is within the range of the data (interpolation); extrapolation is unreliable **B1**

(b) Answer: 16.5

(c) strong **B1**

(c) negative correlation **B1**

(c) Answer: strong negative linear correlation

Question 7

(a) 37 - 31 **M1**

(a) 6 **A1**

(a) Answer: 6

(b) lower limit 31 - 1.5 x 6 = 22 and upper limit 37 + 1.5 x 6 = 46 **M1**

(b) smallest value 26 is not below 22, so is not an outlier **A1**

(b) largest value 55 is above 46, so is an outlier **A1**

(b) Answer: smallest is not an outlier; largest is an outlier

(c) compares median - Q1 = 4 and Q3 - median = 2 **M1**

(c) negative skew (as the larger spread is on that side) **A1**

(c) Answer: negative skew

Question 8

(a) mean of y = 62/23 **M1**

(a) 2.696 **A1**

(a) standard deviation = $\sqrt{440/23 - (2.6957)^2}$ **M1**

(a) 3.444 **A1**

(a) Answer: mean 2.696, sd 3.444

(b) x = 10y + 44 so mean of x = 10 x 2.6957 + 44 **M1**

(b) 70.96 **A1**

(b) standard deviation of x = 10 x 3.4444 **M1**

(b) 34.44 **A1**

(b) Answer: mean 70.96, sd 34.44

(c) n x mean of x = 23 x 70.9565 **M1**

(c) 1632.0 **A1**

(c) Answer: 1632.0

Question 9

(a) 56 - 39 **M1**

(a) 17 **A1**

(a) Answer: 17

(b) lower limit 39 - 1.5 x 17 = 13.5 and upper limit 56 + 1.5 x 17 = 81.5 **M1**

(b) smallest value 30 is not below 13.5, so is not an outlier **A1**

(b) largest value 101 is above 81.5, so is an outlier **A1**

(b) Answer: smallest is not an outlier; largest is an outlier

(c) compares median - Q1 = 8 and Q3 - median = 9 **M1**

(c) positive skew **A1**

(c) Answer: positive skew

Question 10

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

(a) values 3, 18/5, 23/5, 5 **A1**

(a) Answer: 3, 18/5, 23/5, 5

(b) proportion of the class 5 - 10 lying in the required interval = (10 - 7.5)/5 **M1**

(b) 9 **A1**

(b) Answer: 9

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

(c) 3.6 cm **A1**

(c) Answer: 3.6 cm

Question 11

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

(a) sum of f x midpoint = 1037.5 **M1**

(a) mean = 1037.5/71 = 14.61 minutes **A1**

(a) Answer: 14.61 minutes

(b) sum f x midpoint^2 = 20043.8 **M1**

(b) variance = 20043.8/71 - (14.6127)^2 **M1**

(b) standard deviation = 8.29 minutes **A1**

(b) Answer: 8.29 minutes

(c) identifies the median class 10 - 15 (cumulative frequency reaches 39 at its upper bound) **M1**

(c) interpolates: 10 + ((35.5 - 24)/15) x 5 **M1**

(c) median = 13.8 minutes **A1**

(c) Answer: 13.8 minutes

Question 12

(a) mean of y = 30/22 **M1**

(a) 1.364 **A1**

(a) standard deviation = $\sqrt{230/22 - (1.3636)^2}$ **M1**

(a) 2.932 **A1**

(a) Answer: mean 1.364, sd 2.932

(b) x = 5y + 72 so mean of x = 5 x 1.3636 + 72 **M1**

(b) 78.82 **A1**

(b) standard deviation of x = 5 x 2.9317 **M1**

(b) 14.66 **A1**

(b) Answer: mean 78.82, sd 14.66

(c) n x mean of x = 22 x 78.8182 **M1**

(c) 1734.0 **A1**

(c) Answer: 1734.0

Question 13

(a) each extra year in x is associated with a change of -0.8 in V, so as x increases V decreases **B1**

(a) Answer: gradient -0.8: V decreases by 0.8 for each unit increase in x

(b) substitutes x = 30 **M1**

(b) V = 11.0 **A1**

(b) reliable only if 30 is within the range of the data (interpolation); extrapolation is unreliable **B1**

(b) Answer: 11.0

(c) moderate **B1**

(c) negative correlation **B1**

(c) Answer: moderate negative linear correlation

Question 14

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

(a) sum of f x midpoint = 2000 **M1**

(a) mean = 2000/34 = 58.82 minutes **A1**

(a) Answer: 58.82 minutes

(b) sum f x midpoint^2 = 123850 **M1**

(b) variance = 123850/34 - (58.8235)^2 **M1**

(b) standard deviation = 13.5 minutes **A1**

(b) Answer: 13.5 minutes

(c) identifies the median class 50 - 60 (cumulative frequency reaches 17 at its upper bound) **M1**

(c) interpolates: 50 + ((17 - 8)/9) x 10 **M1**

(c) median = 60.0 minutes **A1**

(c) Answer: 60.0 minutes

Question 15

(a) mean of y = 3/8 **M1**

(a) 0.3750 **A1**

(a) standard deviation = $\sqrt{103/8 - (0.3750)^2}$ **M1**

(a) 3.569 **A1**

(a) Answer: mean 0.3750, sd 3.569

(b) x = 2y + 37 so mean of x = 2 x 0.3750 + 37 **M1**

(b) 37.75 **A1**

(b) standard deviation of x = 2 x 3.5685 **M1**

(b) 7.137 **A1**

(b) Answer: mean 37.75, sd 7.137

(c) n x mean of x = 8 x 37.7500 **M1**

(c) 302.00 **A1**

(c) Answer: 302.00

Question 16

(a) f = 2.6 x 10 (class width 10) **M1**

(a) f = 26 **A1**

(a) Answer: f = 26

(b) attempt at frequency density = frequency/class width for at least two classes **M1**

(b) all four correct: 0.8, 2.2, 0.9, 0.4 **A1**

(b) correctly identifies $20 < m \leq 30$ as the tallest bar, since it has the highest frequency density (2.6, higher than all of 0.8, 2.2, 0.9, 0.4) **B1**

(b) Answer: Frequency densities: 0.8, 2.2, 0.9, 0.4; the tallest bar is for $20 < m \leq 30$ (frequency density 2.6)

(c) 18 + 12 = 30 parcels **B1**

(c) Answer: 30 parcels

Question 17

(a) three of the five coded values correct **B1**

(a) all five coded values correct: -2, -1, 0, 1, 2 **B1**

(a) Answer: y values: -2, -1, 0, 1, 2

(b) sum(fy) = 2 found (using -12-14+0+12+16) **M1**

(b) mean of y = 2/60 used, then mean of x = 12.5 + 5 x mean of y **M1**

(b) mean = awrt 12.7 hours (exact 38/3) **A1**

(b) Answer: Mean = 12.7 hours (3sf)

(c) sum(fy^2) = 82 found **M1**

(c) variance of y = sum(fy^2)/60 - (mean of y)^2 used **M1**

(c) variance of x = 25 x variance of y (or equivalent unscaling of the coding) **M1**

(c) standard deviation = awrt 5.84 hours **A1**

(c) Answer: Standard deviation = 5.84 hours (3sf)

Question 18

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

(a) $\sqrt{\text{Sxx x Syy}}$ = $\sqrt{48608}$ = awrt 220 (or 220.5) found **M1**

(a) r = awrt -0.689 cao **A1**

(a) Answer: r = -0.689 (3sf)

(b) recognises a fairly strong negative linear correlation (r close to -1, not exactly -1) **B1**

(b) correct context: as average class size increases, mean GCSE point score per pupil tends to decrease (or vice versa), across these 12 schools **B1**

(b) Answer: There is a fairly strong negative linear correlation: as average class size increases, mean GCSE point score tends to decrease

(c) correlation does not imply causation; r only measures the strength of a linear association between the two variables, not a cause and effect relationship **B1**

(c) valid contextual reasoning, e.g. other factors such as school funding, socio-economic background or prior attainment could explain the association, so the claim is not fully justified by r alone **B1**

(c) Answer: The claim is not fully justified; r shows association, not causation, and other factors could explain the relationship

Question 19

(a) sum(x) = 288 found **M1**

(a) mean = 32 **A1**

(a) sum(x^2) = 13770 found and Sxx = 13770 - $288^2/9$ = 4554 used **M1**

(a) standard deviation = awrt 22.5 **A1**

(a) Answer: Mean = 32, standard deviation = 22.5 (3sf)

(b) new sum(x) = 288 - 95 = 193 found **M1**

(b) new mean = 24.125 (or 24.1) **A1**

(b) new sum(x^2) = 13770 - 9025 = 4745 found and new Sxx = 4745 - $193^2/8$ = 88.875 used **M1**

(b) new standard deviation = awrt 3.33 **A1**

(b) Answer: New mean = 24.1 (24.125), new standard deviation = 3.33 (3sf)

(c) the mean decreases noticeably, from 32 to 24.1 **B1**

(c) the standard deviation decreases much more dramatically (in relative terms), from 22.5 to 3.33, showing that the original spread was dominated almost entirely by the single outlying value **B1**

(c) Answer: Both the mean and standard deviation decrease; the standard deviation falls far more dramatically, showing the outlier dominated the original spread

Question 20

(a) cso: 200 - 30 - 54 - 38 - 20 = 58 **B1**

(a) Answer: f = 58

(b) attempt at frequency density = frequency / class width for at least three classes, using correct widths (10, 10, 10, 10, 15) **M1**

(b) all five correct: 3.0, 5.4, 5.8, 3.8, 1.33 (3sf) and correctly identifies 36-45 as the tallest bar (highest frequency density, 5.8) **A1**

(b) Answer: Frequency densities: 3.0, 5.4, 5.8, 3.8, 1.33; the tallest bar is for the 36-45 class

(c) identifies the median class as 35.5-45.5 using cumulative frequencies (30, 84, 142, ...) and n/2 = 100 **M1**

(c) correct interpolation set up: median = 35.5 + [(100-84)/58] x 10 **M1**

(c) median = awrt 38.3 years **A1**

(c) Answer: Median = 38.3 years (3sf)

(d) interpolates within the 25.5-35.5 class: 54 x (4.5/10) attempted **M1**

(d) adds the full first class: 30 + 54 x (4.5/10) **M1**

(d) total = awrt 54 members (54.3) **A1**

(d) Answer: Approximately 54 members (54.3) are aged under 30

Question 21

(a) mean = sum(x)/n = 217/31 attempted **M1**

(a) mean = 7.00 (exact) **A1**

(a) variance = sum(x^2)/n - mean^2 = 1859.4/31 - 7^2 used **M1**

(a) standard deviation = awrt 3.31 **A1**

(a) Answer: Mean = 7.00 knots, standard deviation = 3.31 knots (3sf)

(b) upper boundary = mean + 2 x SD = 7.00 + 2(3.3137) attempted **M1**

(b) upper boundary = awrt 13.6 (ft from (a)) **A1**

(b) correct conclusion with comparison: since 42 > 13.6, the value of 42 knots is classified as an outlier **A1**

(b) Answer: Upper boundary = 13.6 knots; 42 knots is an outlier

(c) **L1** A basic, one-sided comment is made (e.g. simply states the value should or should not be removed), with little or no reference to the statistical outlier test from part (b) or to the storm context. **1-2**

(c) **L2** Some valid discussion is given, referencing either the numerical outlier classification from part (b) or the contextual, genuine nature of the storm reading, but the discussion is not fully developed or does not weigh up both sides. **3-4**

(c) **L3** A full, balanced discussion is given: correctly identifies 42 as a statistical outlier (using the result from part (b)), recognises that it is a genuine measurement linked to an identifiable cause (a storm) rather than a recording error, and reaches a reasoned conclusion (e.g. that it may be reasonable to exclude it only when describing a 'typical' day, while retaining it in the full data set for other purposes such as studying extreme weather), showing clear understanding that the decision depends on the purpose of the analysis. **5-6**

(c) Indicative content:

- 42 knots is confirmed as a statistical outlier by the 2 standard deviation rule from part (b), since it exceeds the upper boundary of approximately 13.6 knots by a large margin
- However, the value is a genuine recorded measurement, not a data entry or recording error, since it is linked to an identifiable and explainable cause, a named storm
- There is therefore no statistical justification for removing it on the grounds that it is invalid data
- If the aim is specifically to describe a 'typical' October day, including the storm value would inflate both the mean and, especially, the standard deviation (as seen in part (a)), giving a misleading impression of a typical day
- This gives some justification for excluding it for that specific purpose, provided this is stated clearly and transparently
- Removing genuine data can understate the true variability and risk of extreme wind events at this location, so it should be retained in any analysis concerned with extremes or risk, and both sets of summary statistics could reasonably be quoted
- Overall, the decision should depend on the purpose of the analysis rather than being made automatically because the value is a statistical outlier

(c) Answer: 42 knots is a genuine, explainable outlier (a storm); it could reasonably be excluded only when describing a 'typical' day, but should be retained for analyses of extreme weather, since the decision depends on the purpose of the analysis

Question 22

(a) IQR(Aviemore) = 4.8 **B1**

(a) IQR(Camborne) = 3.1 **B1**

(a) Answer: IQR(Aviemore) = 4.8 degrees Celsius, IQR(Camborne) = 3.1 degrees Celsius

(b) upper boundary = Q3 + 1.5 x IQR = 14.2 + 1.5(4.8) attempted **M1**

(b) upper boundary = 21.4 **A1**

(b) correct conclusion with comparison: since 19.5 < 21.4, the maximum value is not an outlier **A1**

(b) Answer: Upper boundary = 21.4 degrees Celsius; 19.5 is not an outlier

(c) median comparison: Camborne has the higher median (13.9 vs 11.8), so typical maximum temperatures are higher at Camborne **B1**

(c) spread comparison: Aviemore has the larger IQR (4.8 vs 3.1), so maximum temperatures are more variable (less consistent) at Aviemore **B1**

(c) valid contextual link, e.g. Camborne's coastal position gives a milder, more consistent climate, while Aviemore's inland/upland position leads to more variable daily maximum temperatures **B1**

(c) Answer: Camborne is typically warmer (higher median) and more consistent (smaller IQR); Aviemore is cooler on average and more variable

(d) valid contextual reason, e.g. Aviemore is an inland/upland location in the Scottish Highlands, so lacks the moderating maritime influence that keeps Camborne's coastal temperatures more stable, oe **B1**

(d) Answer: Aviemore's inland/upland location lacks the moderating coastal influence that keeps Camborne's temperatures more consistent