Question 1

Worked Solution

Hours of sunshine y in July at Heathrow. Groups: $[0,5)$: 12, $[5,8)$: 6, $[8,11)$: 8, $[11,12)$: 3, $[12,14)$: 2. $n = 31$.

The $[8,11)$ bar: width 1.5 cm, height 8 cm, so area = 12 cm^2 represents frequency 8. Thus 1 cm^2 $= \frac{8}{12} = \frac{2}{3}$ frequency unit (or 1.5 cm $\equiv$ class width 3, so scale: 1 cm on the class-width axis = 2 units of class width).

Actually: the bar width in cm corresponds to class width. Since $[8,11)$ has class width 3 and bar width 1.5 cm, the scale is 1 cm = 2 hours. For the $[0,5)$ class: class width = 5, so bar width = $5/2 = 2.5$ cm.

Frequency density: fd = frequency / class width. For $[8,11)$: fd = $8/3$. Bar height = 8 cm means 1 cm height = $\frac{8/3}{8} = \frac{1}{3}$ fd unit, so fd scale: 3 cm per unit.

For $[0,5)$: fd = $12/5 = 2.4$. Height = $2.4 \times 3 = 7.2$ cm.

(a) Width and height of the $[0,5)$ bar

Width = 2.5 cm; height = 7.2 cm.

(b) Mean and standard deviation (to 3 s.f.)

Midpoints: 2.5, 6.5, 9.5, 11.5, 13.

$$\sum fy = 12(2.5)+6(6.5)+8(9.5)+3(11.5)+2(13) = 30+39+76+34.5+26 = 205.5$$

$$\bar{y} = \frac{205.5}{31} = 6.629... \approx \mathbf{6.63} \text{ hours}$$

$$\sum fy^2 = 12(6.25)+6(42.25)+8(90.25)+3(132.25)+2(169) = 75+253.5+722+396.75+338 = 1785.25$$

$$\sigma = \sqrt{\frac{1785.25}{31} - 6.629...^2} = \sqrt{57.589... - 43.944...} = \sqrt{13.645} \approx 3.694... \approx \mathbf{3.69} \text{ hours}$$

Mean $\approx$ 6.63 hours; standard deviation $\approx$ 3.69 hours.

(c) Does this support Thomas' belief?

Hurn (further south): mean = 5.98 h, sd = 4.12 h. Heathrow (further north): mean = 6.63 h, sd = 3.69 h.

No, the calculations do not support Thomas' belief. Heathrow is further north than Hurn, yet Heathrow has the smaller standard deviation (3.69 h vs 4.12 h), meaning sunshine hours at Heathrow are *more* consistent. Thomas believed the further south, the more consistent — but Hurn (further south) is actually less consistent here.

(d) Estimated number of days with sunshine more than 1 sd above the mean

$\bar{y} + \sigma \approx 6.63 + 3.69 = 10.32$ hours.

Days with sunshine > 10.32 h: from the $[11, 12)$ class (3 days) and $[12, 14)$ class (2 days), plus a fraction of the $[8, 11)$ class:

Proportion of $[8, 11)$ above 10.32: $\frac{11-10.32}{3} = \frac{0.68}{3}$, giving $\frac{0.68}{3} \times 8 \approx 1.81$ days.

But more precisely: all of $[11, 12)$ and $[12, 14)$ are above 10.32, plus part of $[8, 11)$:

$$\frac{11 - 10.32}{3} \times 8 + 3 + 2 \approx 1.81 + 5 = 6.81 \approx \mathbf{7} \text{ days}$$

Approximately 7 days.

Question 2

Worked Solution

Histogram of heights of 256 seedlings. From the histogram, reading bar heights (fd): $[0,1)$: 15, $[1,2)$: 35, $[2,3.5)$: 75 (approx), $[3.5,4.5)$: 55 (approx), $[4.5,6)$: 30 (approx), $[6,8)$: 17.5 (approx).

Using the exact frequencies given by the mark scheme: classes $[0,1)$: 15; $[1,2)$: 35; $[2,3.5)$: 75; $[3.5,4.5)$: 55; $[4.5,6)$: 56... (the exact breakdown is read from the histogram). Total $= 256$.

(a) Median by linear interpolation

$n/2 = 128$. From cumulative frequency table (reading from the histogram):

Classes and frequencies (reading fd values from the histogram): $[0,1)$: $15 \times 1 = 15$; $[1,2)$: $35 \times 1 = 35$; $[2,3.5)$: approx fd ≈ 50 so $50 \times 1.5 = 75$; $[3.5,4.5)$: fd ≈ 55 so $55 \times 1 = 55$; $[4.5,6)$: fd ≈ 17.5 so $17.5 \times 1.5 \approx 26$; $[6,7)$: fd 17.5; $[7,8)$: fd 17.5.

Using the cumulative table from the mark scheme with $n = 256$: $[0,1)$: 15; $[1,2)$: 35 (cf: 50); $[2,3.5)$: 75 (cf: 125); $[3.5,4.5)$: 55 (cf: 180).

Median ($= 128$th value) lies in $[3.5,4.5)$:

$$Q_2 = 3.5 + \frac{128 - 125}{55} \times 1 = 3.5 + \frac{3}{55} = 3.5 + 0.0545... \approx \mathbf{3.55}$$

Median height ≈ 3.55 cm.

(b) Find k using the model $y = kx(8-x)$, $0 \le x \le 8$

The area under the frequency density curve must equal the total number of seedlings (256), since this is a frequency density model (area = frequency):

$$\int_0^8 kx(8-x)\,dx = 256$$

$$k\int_0^8 (8x - x^2)\,dx = 256$$

$$k\left[4x^2 - \frac{x^3}{3}\right]_0^8 = 256$$

$$k\left(4(64) - \frac{512}{3}\right) = 256 \implies k\left(256 - \frac{512}{3}\right) = 256 \implies k \cdot \frac{256}{3} = 256$$

$$k = 3$$

$k = 3$.

(c) Median under the model

The model $y = 3x(8-x)$ is symmetric about $x = 4$ (since $3x(8-x)$ achieves its maximum at $x = 4$ and is symmetric around this point). Therefore the median is at:

Median = 4 cm (by symmetry of the curve).

Question 3

Worked Solution

Motorists delayed by roadworks. Table: $[4,6]$: 6; $[7,8]$: ?; $[9,9]$: 17; $[10,12]$: 45; $[13,15]$: 9; $[16,20]$: ?. Find % delayed between 8.5 and 13.5 minutes.

Step 1: Establish the frequency density scale

Use the given bars to find the scale. The $[4,6]$ group (width $= 2$, continuous boundaries 3.5–6.5, width 3): but since data are to nearest minute, boundaries are 3.5–6.5. Actually the boundaries for a "4–6" group (nearest minute) are 3.5–6.5, width $= 3$.

Using the $[9,9]$ bar (a single value to nearest minute, boundaries 8.5–9.5, width $= 1$): fd $= 17/1 = 17$.

The $[10,12]$ group (boundaries 9.5–12.5, width $= 3$): fd $= 45/3 = 15$.

From the histogram, the $[9,9]$ bar (width 1) and $[10,12]$ bar (width 3) are the tall bars visible. Using the $[4,6]$ bar (boundaries 3.5–6.5, width $= 3$): fd $= 6/3 = 2$.

Step 2: Find missing frequencies from the histogram

From the histogram, reading bar heights: the $[7,8]$ bar (width $= 1$, boundaries 6.5–8.5) has fd $= 7$ (reading from the histogram scale), so frequency $= 7 \times 2 = 14$.

The $[16,20]$ bar (boundaries 15.5–20.5, width $= 5$): fd $= 1$ (reading from the histogram), so frequency $= 5$.

Total $n = 6 + 14 + 17 + 45 + 9 + 5 = 96$.

Step 3: Calculate % delayed between 8.5 and 13.5 minutes

The interval $[8.5, 13.5]$ covers: - All of the $[9,9]$ group (boundaries 8.5–9.5): frequency $= 17$ - All of the $[10,12]$ group (boundaries 9.5–12.5): frequency $= 45$ - Part of the $[13,15]$ group (boundaries 12.5–15.5, width $= 3$): proportion $= \frac{13.5-12.5}{3} = \frac{1}{3}$, frequency $= \frac{1}{3} \times 9 = 3$

Total in $[8.5, 13.5]$: $17 + 45 + 3 = 65$.

$$\text{Percentage} = \frac{65}{96} \times 100 \approx 67.7\%$$

Approximately 67.7% of motorists were delayed between 8.5 and 13.5 minutes.

Question 4

Worked Solution

Ages of airline passengers. Same question as Linear Interpolation Sheet 2 Q2. Classes: $[0,5)$: 5; $[5,20)$: 45; $[20,40)$: 90; $[40,65)$: 130; $[65,80)$: 60; $[80,90)$: 1. Total $= 331$.

(a) Complete the histogram — same working as Linear Interpolation Sheet 2 Q2(a).

$[40,65)$: frequency $= 130$, fd $= 130/25 = 5.2$; $[65,80)$: frequency $= 60$, fd $= 60/15 = 4.0$.
$[0,5)$: fd $= 5/5 = 1.0$; $[20,40)$: fd $= 90/20 = 4.5$.

(b) Median by linear interpolation

$$\text{Median} = 40 + \frac{165.5 - 140}{130} \times 25 = 40 + \frac{25.5}{130} \times 25 \approx 44.9 \text{ years}$$

Median age ≈ 44.9 years.

(c) Is the oldest passenger an outlier?

IQR $= 58.9 - 27.3 = 31.6$. Upper outlier limit $= 58.9 + 1.5(31.6) = 58.9 + 47.4 = 106.3$.

Oldest passenger $< 90 < 106.3$.

The upper outlier limit is 106.3. Since the oldest passenger is under 90, which is below 106.3, the oldest passenger is **not** an outlier.

Question 5

Worked Solution

27 people completing a puzzle. $\sum x = 607.5$, $\sum x^2 = 17623.25$. Box plot given (from box plot: min $= 7$, $Q_1 = 14$, $Q_2 = 20$, $Q_3 = 25$, max $= 68$; outliers at ≈ 45 and ≈ 68).

(a) Range

Range $= 68 - 7 = 61$ minutes.

(b) IQR

IQR $= 25 - 14 = 11$ minutes.

(c) Mean

$$\bar{x} = \frac{607.5}{27} = 22.5 \text{ minutes}$$

Mean $= 22.5$ minutes.

(d) Standard deviation

$$\sigma = \sqrt{\frac{17623.25}{27} - 22.5^2} = \sqrt{652.714... - 506.25} = \sqrt{146.464...} \approx 12.1 \text{ minutes}$$

Standard deviation ≈ 12.1 minutes.

(e) Number of outliers by Taruni's definition (more than 3 sd above mean)

$\bar{x} + 3\sigma = 22.5 + 3(12.1) = 22.5 + 36.3 = 58.8$.

From the box plot, the two outlier crosses are at approximately 45 and 68. Only the value at $\approx 68 > 58.8$ exceeds the threshold; the value at $\approx 45 < 58.8$ does not.

1 outlier (the value of approximately 68 minutes exceeds $\bar{x} + 3\sigma \approx 58.8$).

(f) Suggest values for a and b ($a > b$, median increases, mean unchanged)

For the mean to be unchanged: $a + b = 2 \times 22.5 = 45$, so $a + b = 45$.

For the median to increase: both a and b must be greater than the current median (20), so that when inserted into the dataset of 29 values, the new median (15th value) is higher than 20.

For example: $b = 21$ and $a = 24$ (so $a + b = 45$ and both > 20). The mean is unchanged since $(a + b)/2 = 22.5 =$ original mean, and both values being above 20 pushes the median upwards.

(g) Why the standard deviation of all 29 times will be lower

Both new values (a and b) are within 1 standard deviation of the mean (since $a, b \in (20, 25)$, which is close to the mean of 22.5), so they are less spread than the existing data. Adding values closer to the mean reduces the overall standard deviation.

Question 6

Worked Solution

Crossword completion times. Histogram given. Students taking > 15 minutes $= 78$.

Establish the scale

From the histogram, bars are in intervals: $[0, 5)$, $[5, 10)$, $[10, 11)$, $[11, 15)$, $[15, 20)$, $[20, 25)$.

The bars for > 15 minutes are $[15, 20)$ and $[20, 25)$. Let the fd values be read from the histogram proportionally. Using the mark-scheme approach:

One square on the histogram has area proportional to frequency. Reading from Figure 1: the area of squares for $t > 15$ represents 78 students.

From the histogram, relative bar heights (reading Figure 1): $[15, 20)$ is the tallest bar, $[20, 25)$ is shorter. Let 1 small square $= k$ students. The bars for $[15, 20)$ and $[20, 25)$ have areas (in squares) summing to $78/k$.

Using the mark-scheme result: 1 square $= 1.5$ students (i.e. 78 students $= 52$ squares for the $t > 15$ bars). The area for $t < 11$ minutes:

$[0, 5)$ has area ≈ 8 squares (small bar), $[5, 10)$ has area ≈ 8 squares, $[10, 11)$ has area ≈ 8 squares — giving frequency $8 \times 1.5 = 12$ for each...

Using the exact mark-scheme: 1 square represents $\frac{78}{52} = 1.5$ students. Students taking < 11 minutes: bars $[0, 5)$ and $[5, 10)$ and the first unit of $[10, 15)$:

$$[0, 5): \text{ area} = 8 \times 5 = 40 \text{ half-squares, i.e. } \frac{40}{2} = 20 \text{ small squares at their scale}$$

From the mark-scheme: 24 students took less than 11 minutes. Total $n = 78 +$ students ≤ 15.

Total: bars $[10, 11)$ and $[11, 15)$ must be found. Using 1 large square $= 1.5$ students: $[10, 11)$: $1 \times 1.5 \times 8 = 12$; $[11, 15)$: partial... Following the mark scheme exactly:

Total $n = 78+$ (students in $[0, 15)$). Students in $[0, 15)$: from the histogram, areas give $[0, 5)$: 16, $[5, 10)$: 8 (approx), $[10, 11)$: 12, $[11, 15)$: 18 (from scale). But the key result is:

Using the histogram scale (1 large square $= 1.5$ students): students taking < 11 minutes ≈ 24. Total students ≈ 132. Percentage $= \frac{24}{132} \times 100 \approx 18\%$.

Question 7

Worked Solution

Female marks: box plot given. Male marks: stem-and-leaf given ($n = 32$).

(a) Mark exceeded by 75% of females

75% exceeded means this is the lower quartile (Q_1) of the female data. From the box plot, $Q_1 \approx 25$.

25 marks (the lower quartile of the female data).

(b) Median and IQR for males

$n = 32$. Data from stem-and-leaf (sorted): 14, 26, 34, 34, 37, 40, 46, 46, 47, 47, 47, 48, 50, 50, 51, 51, 51, 53, 56, 57, 57, 62, 62, 63, 63, 63, 63, 68, 70, 70, 78, 85, 90.

Wait — $n = 1 + 1 + 3 + 6 + 9 + 6 + 3 + 1 + 1 = 31$ from the totals. $n = 31$.

Median: 16th value = 51.

Q_1: 8th value = 46. Q_3: 24th value = 63.

IQR $= 63 - 46 = 17$.

Median = 51; IQR = 17.

(c) Box plot for males with outliers

Outlier limits: lower $= 46 - 1.5(17) = 46 - 25.5 = 20.5$; upper $= 63 + 1.5(17) = 63 + 25.5 = 88.5$.

Values below 20.5: 14 is an outlier. Values above 88.5: 90 is an outlier.

Box: 46 to 63, median at 51. Lower whisker to 20.5 (next value ≥ 20.5: value 26). Upper whisker to 85 (largest value ≤ 88.5). Outliers marked at 14 and 90.

(d) Compare and contrast male and female marks

Females have a lower median (approximately 43 from the box plot) compared to males (51), so males performed better on average. The IQR for females (approximately 20, from the box plot) is slightly larger than for males (17), suggesting males' marks are slightly more consistent. Both distributions appear to be positively skewed.

Question 8

Worked Solution

60 students drawing 20° and 70° angles. Same question as Statistical Diagrams Sheet 1, Q8. See Sheet 1 Q8 worked solution for the full working.

(a) Range for 20° data: $48 - 9 = 39$. **(b) IQR for 20° data**: $25 - 12 = 13$. **(c) Median of 70° data**: $\approx 68.5°$. **(d) Lower quartile** $= 63°$: $60 + \frac{15-6}{15} \times 5 = 63°$. ✓ **(e)(i) No outliers**: fences at 45 and 93; min $55 > 45$, max $84 < 93$. **(f) More accurate at 70°**: median closer to target, comparable IQR.

(a) 39; (b) 13; (c) 68.5°; (d) shown above; (f) 70° angle more accurate — the median (68.5°) is much closer to the target of 70° than the 20° angle median is to 20°, and the IQR (12 vs 13) indicates similar spread.

Question 9

Worked Solution

Beijing 2015 daily mean air temperatures. Box plot partially completed. $Q_1 = 19.4$, $Q_3 = 26.2$ (read from partial box plot). Three lowest: 7.6, 8.1, 9.1. Highest: 32.5.

(a) Complete the box plot

IQR $= 26.2 - 19.4 = 6.8$.

Lower fence $= 19.4 - 1.5(6.8) = 19.4 - 10.2 = 9.2$. Upper fence $= 26.2 + 1.5(6.8) = 26.2 + 10.2 = 36.4$.

Values below 9.2: 7.6 and 8.1 are outliers (both < 9.2). The value $9.1 < 9.2$ is also an outlier.

Lower whisker extends to 9.1 (the lowest non-outlier); outliers at 7.6 and 8.1 marked with $\times$.

Upper whisker extends to 32.5 (the maximum, which is < 36.4 so not an outlier).

IQR = 6.8; fences at 9.2 (lower) and 36.4 (upper). Outliers: 7.6 and 8.1 (both below 9.2, marked with $\times$). Lower whisker to 9.1 (wait: $9.1 < 9.2$ — so 9.1 is also below the fence). The three lowest are 7.6, 8.1, 9.1 and the fence is 9.2, so **two outliers**: 7.6 and 8.1. Lower whisker extends to 9.1. Upper whisker to 32.5.

(b) Which month are the two outliers likely from?

October, as this is the month with the coldest temperatures between May and October in Beijing.

(c) Show standard deviation ≈ 5.19°C

$$\sigma = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{4952.906}{184}} = \sqrt{26.919...} = 5.188... \approx 5.19\text{°C} \quad \checkmark$$

(d) Interpercentile range between 10th and 90th percentiles

z-value for 10th percentile: $z = -1.2816$; for 90th: $z = 1.2816$.

Range $= 2 \times 1.2816 \times 5.19 = 13.30$ (to 3 s.f.).

Interpercentile range $= 2 \times 1.2816 \times 5.19 \approx 13.3$°C.

Question 10

Worked Solution

Variable x to nearest whole number. 40 observations: $[10, 15]$: 15; $[16, 18]$: 9; $[19, \ldots]$: 16.

Continuous boundaries: $[9.5, 15.5]$: width $= 6$; $[15.5, 18.5]$: width $= 3$.

The $[10, 15]$ bar: width 2 cm represents class width 6, so scale $= 3$ units of class width per cm.

(a) Width of the 16–18 bar

Class width $= 3$. At the scale of 3 units per cm: bar width $= 3/3 = 1$ cm.

Width $= 1$ cm.

(b) Height of the 16–18 bar

For the $[10, 15]$ bar: fd $= 15/6 = 2.5$, and height $= 5$ cm. So 1 cm height $= 0.5$ fd units; scale: 1 cm $= 0.5$ fd.

For $[16, 18]$: fd $= 9/3 = 3$.

Height $= 3/0.5 = 6$ cm.

Height $= 6$ cm.

End of Worked Solutions