IB Math AI SL — Topic 4

Statistics Toolkit

Sampling · Central Tendency & Dispersion · Outliers · Box Plots, CF Graphs & Histograms

Difficulty	EASY	MEDIUM	HARD
Questions	10	10	10
Total Marks	30	40	50

Scan me for help

Name:________________________ Date:____________

Score:__________/120

Instructions: Answer all questions. Show all working. Give answers correct to 3 significant figures unless stated otherwise. A graphic display calculator is required throughout; support GDC answers with the values entered and the feature used.

Key Formulae

- **Population mean:** $\mu = \dfrac{\sum x}{n}$ **Sample mean:** $\bar{x} = \dfrac{\sum fx}{\sum f}$
- **Grouped data mean (estimate):** $\bar{x} \approx \dfrac{\sum fm}{\sum f}$, where m is the midpoint of each class
- **Interquartile range:** $IQR = Q_3 - Q_1$
- **Outlier fences:** Lower fence $= Q_1 - 1.5 \times IQR$; Upper fence $= Q_3 + 1.5 \times IQR$
- **Linear transformation of data** $y = ax + b$: $\bar{y} = a\bar{x} + b$ and $\sigma_y = |a|\sigma_x$
- **Percentile from cumulative frequency:** locate $100 \times \dfrac{p}{n}$ on the cumulative frequency axis, then read across to the data axis
- **Five-number summary:** minimum, Q_1, median (Q_2), Q_3, maximum

GDC Tips

- **1-Var Stats (mean, median, σ, quartiles):** Enter data in List 1; frequencies (if given) in List 2. Then Statistics $\rightarrow$ Calc $\rightarrow$ 1-Var Stats, set XList=L1, Freq=L2. Read off $\bar{x}$, σ_x, Med, Q_1, Q_3.
- **Box plot from GDC:** After entering data, go to Statistics $\rightarrow$ Graph $\rightarrow$ select Box Plot type; the five-number summary is displayed when tracing.
- **Grouped data mean estimate:** Enter midpoints in List 1 and frequencies in List 2, then run 1-Var Stats as above — the GDC treats midpoints as the data values.
- **Cumulative frequency values:** Enter class upper boundaries in List 1 and cumulative frequencies in List 2; use Graph $\rightarrow$ Scatter joined with a smooth curve, or read values directly from the table.
- **Standard deviation check:** The GDC gives both σ_x (population, use for descriptive stats) and s_x (sample, use for inference). For IB AI SL descriptive questions always quote σ_x unless told otherwise.

Common Mistakes to Avoid

1. **Forgetting to use midpoints for grouped data.** When data are given in class intervals, the exact values are unknown. You must use class midpoints and the answer is an estimate — writing the word "estimate" is required for full marks.
2. **Adding b to the standard deviation after a linear transformation.** The transformation $y = ax + b$ shifts every value by b, so the mean shifts too. However, the standard deviation measures spread and is unaffected by adding a constant: $\sigma_y = |a|\sigma_x$ only. A common error is writing $\sigma_y = |a|\sigma_x + b$.
3. **Using the wrong standard deviation from the GDC.** The GDC displays both σ_x (population SD) and s_x (sample SD). For descriptive statistics questions in IB AI SL, use σ_x. Selecting s_x by mistake gives a slightly larger value and loses the accuracy mark.
4. **Misidentifying outliers without computing the fences first.** Do not eyeball a data set and declare a value an outlier. Always calculate $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ explicitly, then compare each suspected value to the appropriate fence. A "determine and justify" instruction requires the fence values to be written down.
5. **Reading the wrong axis on a cumulative frequency graph.** To find the median, locate $\frac{n}{2}$ on the vertical (cumulative frequency) axis first, draw a horizontal dashed line to the curve, then a vertical dashed line down to the horizontal (data) axis. Reversing this process — starting on the horizontal axis — answers a different question entirely.
6. **Comparing two data sets with only one statistic.** Any comparison of two distributions must include both a centre statement (mean or median, in context with units) and a spread statement (IQR or standard deviation, in context with units). A single sentence earns at most half the available marks.
7. **Not verifying the standard deviation by computation.** Do not recall or estimate a standard deviation by eye. Always enter the data (or midpoints and frequencies) into the GDC's 1-Var Stats and read σ_x

directly — an arithmetic slip in $\sum fx^2$ or in $\sum f(x-\bar{x})^2$ is the single most common source of error on this topic.

Section A — Easy **EASY** (10 questions × 3 marks = 30 marks)

Scan me for help

1. EASY Calculator **[3 marks]**

A school surveys students about their favourite sport. The school has 200 students: 80 in Year 10, 72 in Year 11, and 48 in Year 12. A stratified sample of 50 students is to be taken.
Calculate the number of Year 11 students that should be included in the sample.

2. EASY Calculator **[3 marks]**

The ages, in years, of eight volunteers at a community garden are listed below.

$$19,\ 23,\ 17,\ 31,\ 23,\ 28,\ 22,\ 25$$

Find the mean and median age of the volunteers.

3. EASY Calculator **[3 marks]**

The number of books read by each student in a class during one semester is recorded in the frequency table below.

Number of books (x)	1	2	3	4
Frequency (f)	5	9	8	3

(a) Write down the mode.

(b) Find the mean number of books read per student.

4. EASY Calculator **[3 marks]**

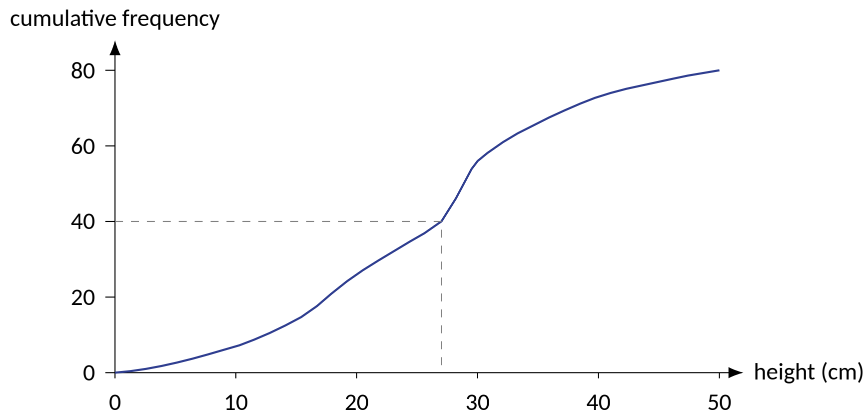
The daily high temperatures, in degrees Celsius, recorded at a weather station over ten days are:

$$14,\ 18,\ 21,\ 15,\ 19,\ 22,\ 17,\ 20,\ 16,\ 18$$

Find the range and standard deviation of these temperatures.

5. EASY Calculator **[3 marks]**

The heights of 80 seedlings are recorded. The cumulative frequency curve is shown below.



(a) Write down the total number of seedlings with a height of 30 cm or less.

(b) Use the graph to estimate the median height of the seedlings.

6. EASY No Calculator **[3 marks]**

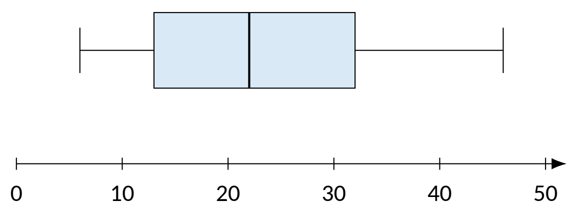
A set of data has a mean of 12 and a standard deviation of 3. Each value in the data set is transformed using the rule $y = 2x + 5$, where x is the original value and y is the transformed value.

(a) Find the mean of the transformed data.

(b) Find the standard deviation of the transformed data.

7. **EASY** **Calculator** **[3 marks]**

The box and whisker diagram below shows the distribution of distances, in kilometres, cycled by 30 participants in a charity ride.

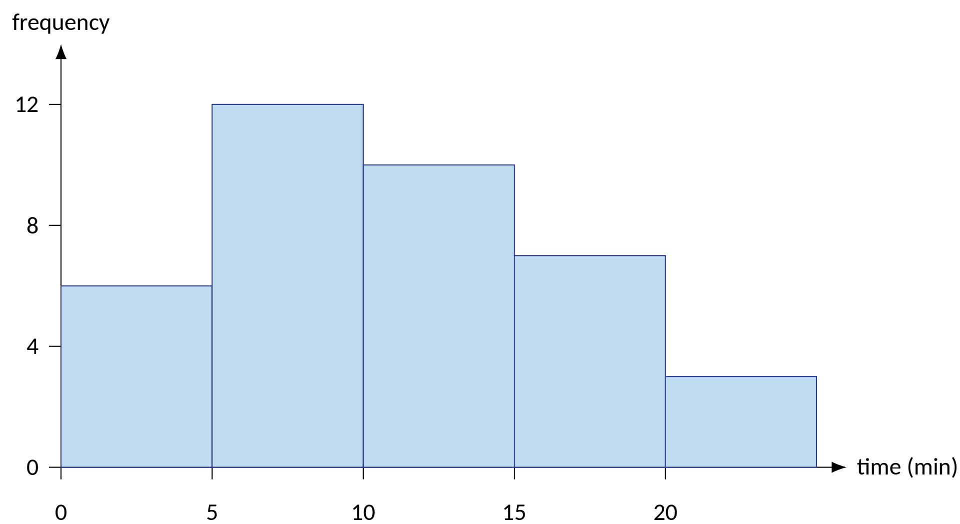


(a) Write down the median distance.

(b) Find the interquartile range.

8. **EASY** **Calculator** **[3 marks]**

The histogram below shows the time, in minutes, that 38 customers spent waiting at a bank.



(a) Write down the modal class.

(b) Find the number of customers who waited for less than 10 minutes.

9. EASY Calculator **[3 marks]**

The test scores (out of 50) of a group of students are:

$$32,\ 45,\ 38,\ 27,\ 41,\ 36,\ 29,\ 50,\ 33,\ 44$$

(a) Find the lower quartile (Q_1) and upper quartile (Q_3).

(b) Find the interquartile range.

10. EASY Calculator **[3 marks]**

A travel agency records the number of holidays booked per day over 5 days:

$$4,\ 7,\ 5,\ 9,\ 5$$

A new booking system is introduced. Each daily booking count increases by 3.

(a) Write down the new mean number of bookings per day.

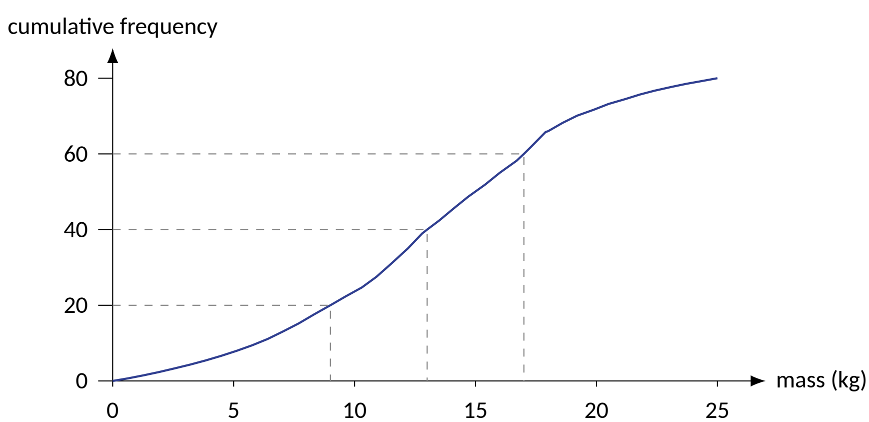
(b) Write down the new standard deviation.

Section B — Medium **MEDIUM** (10 questions × 4 marks = 40 marks)

Scan me for help

11. **MEDIUM** **Calculator** **[4 marks]**

The masses, in kilograms, of 80 parcels delivered by a courier company are recorded in the cumulative frequency graph below.



(a) Write down the median mass. [1]

(b) Find the interquartile range of the masses. [2]

(c) Find the number of parcels with a mass greater than 18 kg. [1]

12. MEDIUM Calculator **[4 marks]**

The table below shows the distribution of the number of text messages sent per day by a group of 40 students.

Number of messages	10	20	30	40
Number of students	6	14	12	8

(a) Find the mean number of messages sent per day. [2]

(b) Each student sends exactly 5 additional messages per day during a school holiday. Write down the mean number of messages sent per day during the holiday. [1]

(c) Write down the standard deviation of the number of messages sent per day during the holiday. [1]

13. MEDIUM Calculator **[4 marks]**

A swimming club records the times, in seconds, of its 30 members completing a 50 m sprint. The results are summarised below.

Time (s)	$30 \leq t < 35$	$35 \leq t < 40$	$40 \leq t < 45$	$45 \leq t < 50$	$50 \leq t < 55$
Frequency	3	8	11	6	2

(a) Write down the modal class. [1]

(b) Estimate the mean time. Justify why your answer is an estimate. [3]

14. MEDIUM Calculator **[4 marks]**

A dataset of seven daily temperatures (in °C) recorded at a weather station is:

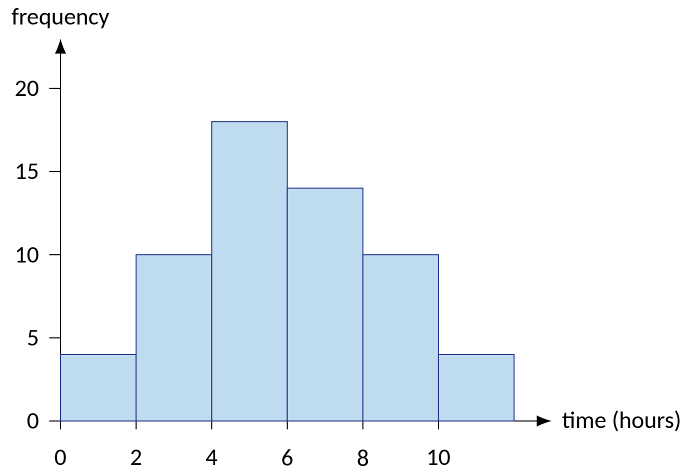
$$12,\ 15,\ 14,\ 13,\ 18,\ 16,\ 35$$

(a) Find the value of Q_1, Q_3, and the interquartile range (IQR). [2]

(b) Determine, with justification, whether the value 35 is an outlier. [2]

15. **MEDIUM** **Calculator** **[4 marks]**

The histogram below shows the number of hours per week spent on exercise by 60 adults in a fitness study. Each class has equal width of 2 hours.



(a) Write down the modal class. [1]

(b) Find the number of adults who exercise for 6 or more hours per week. [1]

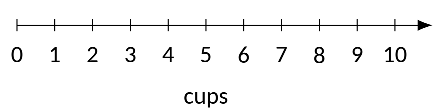
(c) Estimate the mean number of hours spent on exercise per week. [2]

16. **MEDIUM** **Calculator** **[4 marks]**

A researcher asks a sample of workers how many cups of coffee they drink per day. The five-number summary of the data is:

$$\text{Min} = 0, \quad Q_1 = 2, \quad \text{Median} = 3, \quad Q_3 = 5, \quad \text{Max} = 9$$

(a) On the grid below, draw a box and whisker diagram for this data. [2]



(b) State one feature of the distribution that can be seen from the box plot. [1]

(c) A second group of workers has median 4 cups and IQR of 1 cup. Compare the two distributions. [1]

17. MEDIUM No Calculator **[4 marks]**

At an international school, students are selected using systematic sampling from a list of 240 students. A sample of 20 students is required.

(a) Calculate the sampling interval. [1]

(b) Describe how the systematic sample would be carried out, including how the starting point is chosen. [2]

(c) State one disadvantage of systematic sampling compared to simple random sampling. [1]

18. MEDIUM Calculator **[4 marks]**

The ages, in years, of 9 members of a book club are:

$$24,\ 27,\ 31,\ 34,\ 38,\ 41,\ 45,\ 52,\ 68$$

(a) Find the mean age. [1]

(b) Find the standard deviation of the ages. [1]

(c) A tenth member joins the club. The new mean age is 41 years. Find the age of the tenth member. [2]

19. MEDIUM Calculator **[4 marks]**

A stratified sample of 50 employees is to be selected from a company with three departments: Operations (120 employees), Sales (80 employees), and Finance (50 employees).

(a) Calculate the total number of employees in the company. [1]

(b) Calculate the number of employees that should be selected from each department. [3]

20. **MEDIUM** **Calculator** **[4 marks]**

The scores of 24 students on a geography test are shown in the frequency table below.

Score	4	5	6	7	8	9
Frequency	2	3	7	6	4	2

(a) Find the mean score. [2]

(b) Find the median score. [1]

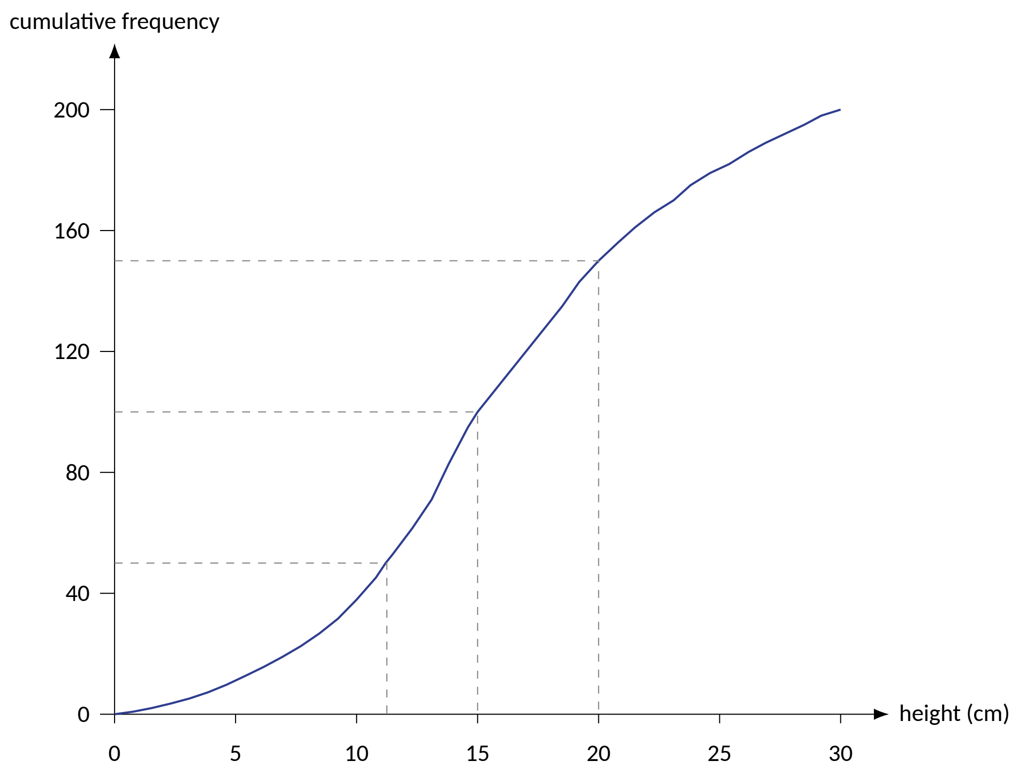
(c) Write down the mode. [1]

Section C — Hard **HARD** (10 questions $\times$ 5 marks = 50 marks)

Scan me for help

21. **HARD** **Calculator** **[5 marks]**

The heights, in cm, of 200 seedlings grown in a greenhouse are recorded. The cumulative frequency curve is shown below.



(a) Write down the median height of the seedlings. [1]

(b) Find the interquartile range of the heights. [2]

(c) A seedling is classified as an outlier if its height is more than $1.5 \times IQR$ above Q_3. Determine whether a seedling of height 27.5 cm is an outlier. Justify your answer. [2]

22. HARD Calculator **[5 marks]**

A factory produces bolts whose diameters, d mm, are normally distributed. A quality inspector measures a sample of bolts and records the following grouped frequency table.

Diameter (mm)	Frequency
$9.80 \leq d < 9.85$	3
$9.85 \leq d < 9.90$	11
$9.90 \leq d < 9.95$	24
$9.95 \leq d < 10.00$	31
$10.00 \leq d < 10.05$	18
$10.05 \leq d < 10.10$	9
$10.10 \leq d < 10.15$	4

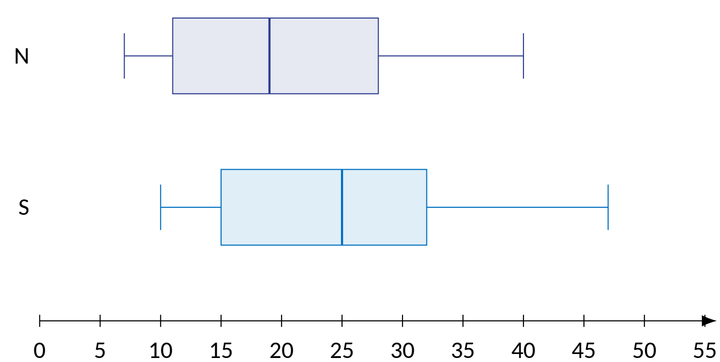
(a) Write down the modal class. [1]

(b) Show that an estimate of the mean diameter is 9.9715 mm. [2]

(c) Each bolt diameter is converted to inches using the formula $I = 0.03937\,d$. Find the mean and standard deviation of the diameters in inches, giving your answers to 3 significant figures. [2]

23. HARD Calculator **[5 marks]**

The ages, in years, of members of two chess clubs, Northside and Southside, are displayed in the box plots below.



Northside has 40 members; Southside has 60 members.

(a) Write down the median age for each club. [1]

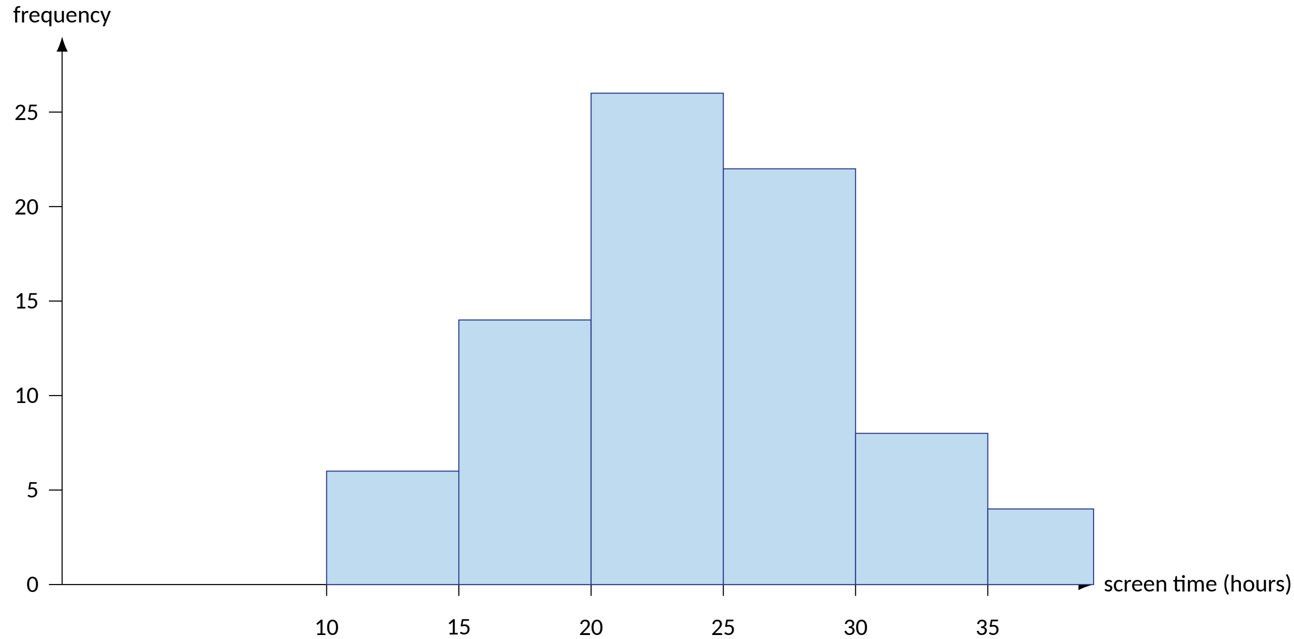
(b) Calculate the interquartile range for each club. [1]

(c) Compare the two distributions. Your answer must include one statement about centre and one statement about spread, both in context. [2]

(d) A member is selected at random from Northside. Find the probability that their age is less than 19 years, given that the minimum age in Northside is 7 years and $Q_1 = 11$ years. Justify any assumptions you make. [1]

24. **HARD** **Calculator** **[5 marks]**

A survey records the weekly screen time, t hours, of 80 teenagers. The data are summarised in the histogram below (all class widths are 5 hours).



(a) Verify that the total frequency is 80. [1]

(b) Estimate the mean weekly screen time. [2]

(c) The survey organiser states: "The mean weekly screen time for all teenagers is 24.5 hours." Explain why this statement may not be valid, giving two reasons. [2]

25. HARD Calculator **[5 marks]**

The distances, in km, cycled per week by members of a cycling club are listed below.

$$38,\ 45,\ 52,\ 60,\ 63,\ 67,\ 71,\ 74,\ 78,\ 82,\ 89,\ 95,\ 112$$

(a) Find the median, Q_1, and Q_3 of these distances. [1]

(b) Show that 112 km is **not** an outlier using the fence $Q_3 + 1.5 \times IQR$. [2]

(c) A club coordinator converts each distance to miles using $M = 0.621\,d$, where d is the distance in km. Find the mean and standard deviation of the distances in miles. [2]

26. HARD Calculator **[5 marks]**

A market researcher records the number of items purchased per customer, x, in a supermarket over one afternoon. The results are given below.

Items purchased (x)	1	2	3	4	5	6	7
Number of customers (f)	4	9	14	20	17	10	6

(a) Find the mean and standard deviation of the number of items purchased. [2]

(b) Each customer is charged a fixed entry fee of $1.50 plus $2.80 per item purchased. Let C be the total charge in dollars for a randomly selected customer. Find the mean and standard deviation of C. [2]

(c) Determine whether a customer purchasing 7 items is an outlier, using the rule: an outlier is a value more than 2 standard deviations above the mean. Justify your answer. [1]

27. HARD Calculator **[5 marks]**

A statistician is analysing the lengths, in mm, of a species of beetle. A sample of 120 beetles is taken and the following information is known:

$$Q_1 = 18.4 \text{ mm}, \quad \text{Median} = 21.7 \text{ mm}, \quad Q_3 = 25.1 \text{ mm}, \quad \text{Min} = 12.3 \text{ mm}, \quad \text{Max} = 33.8 \text{ mm}$$

(a) Calculate the interquartile range and hence find both outlier fences. [2]

(b) Determine whether the minimum value and the maximum value are outliers. Justify each answer. [2]

(c) A second sample of 120 beetles of a different subspecies has a median of 19.4 mm and an IQR of 9.6 mm. Write one conclusion comparing the two samples with respect to the typical length and the consistency of lengths. [1]

28. HARD Calculator **[5 marks]**

A marine biologist records the mass, m kg, of fish caught in a lake. The data for 100 fish are summarised in the grouped frequency table.

Mass (kg)	Frequency	Cumulative Frequency
$0 \leq m < 0.5$	8	8
$0.5 \leq m < 1.0$	19	27
$1.0 \leq m < 1.5$	31	58
$1.5 \leq m < 2.0$	24	82
$2.0 \leq m < 2.5$	12	94
$2.5 \leq m < 3.0$	6	100

(a) Show that an estimate of the mean mass is 1.405 kg. [2]

(b) Estimate the standard deviation of the mass. [1]

(c) A biologist claims the median mass lies in the interval $1.0 \leq m < 1.5$. Determine whether this claim is correct and justify your answer using the cumulative frequency column. [2]

29. HARD Calculator **[5 marks]**

A school records the examination scores, out of 100, for 160 students. The scores are modelled by the grouped data below.

Score	Frequency
$20 \leq s < 35$	12
$35 \leq s < 50$	28
$50 \leq s < 65$	45
$65 \leq s < 80$	50
$80 \leq s < 95$	25

(a) Write down the modal class. [1]

(b) Estimate the mean score. [2]

(c) A student scoring above 77.5 is awarded a distinction. Using your estimated mean and the standard deviation found by GDC, determine whether a score of 88 is more than 1.5 standard deviations above the mean, and state what this implies about awarding distinctions to such students. [2]

30. HARD Calculator **[5 marks]**

A fitness tracker records the resting heart rates, h beats per minute, of 90 adults. The results are summarised below.

$$\sum f = 90, \qquad \sum fx = 5715, \qquad \sum fx^2 = 366\,489$$

where x is the midpoint of each class interval.

(a) Show that the estimated mean resting heart rate is 63.5 bpm. [1]

(b) Show that the estimated standard deviation is approximately 6.31 bpm. [2]

(c) A health guideline states that a resting heart rate is “unusually high” if it exceeds $\mu + 2\sigma$. Find the threshold for an unusually high resting heart rate and determine whether a value of 82 bpm qualifies. Justify your answer. [2]

Mark Schemes

Q1 — Stratified Sampling Allocation EASY

Number of Year 11 in sample $= \dfrac{72}{200} \times 50$ **M1**

$= 18$ students **A1**

Note: The denominator is the total school population (200), not the sum of two year groups. Award A1 for a fully correct proportional method.

Q2 — Mean and Median from Raw Data EASY

Mean $\bar{x} = \dfrac{19+23+17+31+23+28+22+25}{8} = \dfrac{188}{8} = 23.5$ years **A1**

Ordered data: $17, 19, 22, 23, 23, 25, 28, 31$; with 8 values the median is the mean of the 4th and 5th values **M1**

Median $= \dfrac{23+23}{2} = 23$ years **A1**

Note: With an even number of values the median is the mean of the two middle values, not either one alone.

Q3 — Mode and Mean from Frequency Table EASY

(a) Mode $= 2$ books (highest frequency, 9) **A1**

(b) $\bar{x} = \dfrac{(1)(5)+(2)(9)+(3)(8)+(4)(3)}{5+9+8+3} = \dfrac{5+18+24+12}{25} = \dfrac{59}{25}$ **M1**

$= 2.36$ books **A1**

Note: Divide by the total frequency (25), not by the number of distinct values (4).

Q4 — Range and Standard Deviation EASY

Range $= 22 - 14 = 8\,^\circ$C **A1**

GDC: enter the ten temperatures in List 1; Statistics $\to$ 1-Var Stats; read off σ_x **M1**

$\sigma_x = 2.45\,^\circ$C (unrounded: $2.4495\ldots$) **A1**

Note: Use the population standard deviation σ_x (σ_n on some GDCs), not the sample standard deviation s (σ_{n-1}), since this is the complete ten-day record.

Q5 — Reading a Cumulative Frequency Curve EASY

(a) Reading directly from the curve at $x = 30$ cm, the cumulative frequency is 56.
56 seedlings have a height of 30 cm or less. **A1**

(b) The median corresponds to the $\frac{80}{2} = 40$th value on the cumulative frequency axis. Draw a horizontal dashed line from 40 on the y-axis to the curve, then read the corresponding x-value. **M1**

The median height is estimated as 27 cm. **A1**

Note: The median is found by reading across from the 40th value (half of 80), not from the midpoint of the horizontal axis.

Q6 — Linear Transformation of Data EASY

(a) $\bar{y} = a\bar{x} + b$ with $a = 2$, $b = 5$, $\bar{x} = 12$ **M1**

$\bar{y} = 2(12) + 5 = 29$ **A1**

(b) $\sigma_y = |a|\sigma_x$; the additive constant b does not affect spread
$\sigma_y = |2| \times 3 = 6$ **A1**

Note: The constant $b = 5$ shifts every value equally and does not change the standard deviation; only the multiplier a affects the spread.

Q7 — Reading a Box and Whisker Diagram EASY

(a) The median is shown by the vertical line inside the box.
Median $= 22$ km **A1**

(b) $IQR = Q_3 - Q_1$. From the diagram, $Q_1 = 13$ km and $Q_3 = 32$ km **M1**

$IQR = 32 - 13 = 19$ km **A1**

Note: The IQR is measured from the left edge to the right edge of the box (i.e. Q_1 to Q_3), not from the whisker tips.

Q8 — Reading a Histogram EASY

(a) The modal class is the class interval with the tallest bar. The bar for $5 \le t < 10$ has the greatest frequency (12).
Modal class: $5 \le t < 10$ minutes **A1**

(b) Customers who waited less than 10 minutes are in the first two classes ($0 \le t < 5$ and $5 \le t < 10$) **M1**

$6 + 12 = 18$ customers **A1**

Note: Since all class widths are equal (5 minutes), frequency is read directly from the bar height; no frequency density calculation is needed.

Q9 — Quartiles and IQR from Raw Data EASY

(a) Ordered data: $27, 29, 32, 33, 36, 38, 41, 44, 45, 50$. GDC: enter the ten scores in List 1; Statistics $\to$ 1-Var Stats **M1**

$Q_1 = 32$ and $Q_3 = 44$ **A1**

(b) $IQR = Q_3 - Q_1 = 44 - 32 = 12$ **A1**

Note: Different GDC models may use slightly different quartile conventions; always check the GDC output against the ordered data for a small data set.

Q10 — Effect of Adding a Constant on Mean and SD EASY

(a) Original mean $= \dfrac{4+7+5+9+5}{5} = \dfrac{30}{5} = 6$ bookings. Adding 3 to every value increases the mean by 3 **M1**

New mean $= 6 + 3 = 9$ bookings per day **A1**

(b) Adding a constant to every value shifts all data by the same amount; the spread does not change.
New standard deviation $=$ original standard deviation $= 1.79$ (unrounded: $1.7889\ldots$) **A1**

Note: Adding a constant changes the mean but does NOT change the standard deviation; only multiplication by a factor changes the standard deviation.

Q11 — Cumulative Frequency — Median; IQR; Threshold Count MEDIUM

(a) The median corresponds to the 40th value on the cumulative frequency axis. Reading across from 40 on the vertical axis to the curve, then down to the horizontal axis.
Median $= 13$ kg **A1**

(b) Q_1 corresponds to the 20th value: $Q_1 = 9$ kg. Q_3 corresponds to the 60th value: $Q_3 = 17$ kg **(M1)**

$IQR = Q_3 - Q_1 = 17 - 9 = 8$ kg **A1**

(c) Reading from the curve at $x = 18$: cumulative frequency ≈ 66. Number of parcels above 18 kg $= 80 - 66 = 14$ **A1**

Note: Always subtract the cumulative frequency at the threshold from the total (80) to find the number above that value; do not report the cumulative frequency itself.

Q12 — Frequency Table — Mean and Linear Transformation of Mean/SD MEDIUM

(a) $\bar{x} = \dfrac{10(6)+20(14)+30(12)+40(8)}{40} = \dfrac{60+280+360+320}{40} = \dfrac{1020}{40} = 25.5$ messages **M1A1**

(b) Adding 5 to every value shifts the mean by exactly 5 ($y = x + 5 \Rightarrow \bar{y} = \bar{x} + 5$)
New mean $= 25.5 + 5 = 30.5$ messages **A1**

(c) Adding a constant does not change the spread; the standard deviation is unaffected by a pure transla-

tion.
GDC gives $\sigma_x = 9.73$ messages (unrounded: $9.7340\ldots$); new standard deviation $= 9.73$ messages **A1**
Note: The standard deviation scales only by the multiplicative factor $|a|$; adding a constant (5) leaves the standard deviation unchanged.

Q13 — Grouped Data: Modal Class and Estimated Mean MEDIUM

(a) The class with the highest frequency is $40 \le t < 45$ (frequency 11).
Modal class: $40 \le t < 45$ **A1**

(b) Midpoints: $32.5, 37.5, 42.5, 47.5, 52.5$. GDC: enter midpoints in List 1, frequencies in List 2; run 1-Var Stats **M1**

$$\hat{\bar{x}} = \frac{32.5(3) + 37.5(8) + 42.5(11) + 47.5(6) + 52.5(2)}{30} = \frac{97.5 + 300 + 467.5 + 285 + 105}{30} = \frac{1255}{30}$$

$= 41.83\ldots \approx 41.8$ s **A1**

This is an estimate because the exact data values within each class are unknown; the midpoint is used to represent all values in a class. **A1**

Note: The word "estimate" is mandatory when working with grouped data and midpoints.

Q14 — Outlier Test using IQR Fences MEDIUM

(a) Arrange the data in order: $12, 13, 14, 15, 16, 18, 35$.
$Q_1 = 13$, $Q_3 = 18$, $IQR = 18 - 13 = 5$ **M1A1**

(b) Lower fence $= Q_1 - 1.5 \times IQR = 13 - 7.5 = 5.5$
Upper fence $= Q_3 + 1.5 \times IQR = 18 + 7.5 = 25.5$ **M1**

Since $35 > 25.5$, the value 35 lies beyond the upper fence **R1**

Therefore, $35\,^{\circ}$C is an outlier. *(dependent conclusion)*

Note: The reasoning mark (R1) requires an explicit comparison of the data value with the computed fence.

Q15 — Histogram — Modal Class; Threshold Count; Estimated Mean MEDIUM

(a) The tallest bar spans $4 \le t < 6$ hours (frequency 18).
Modal class: $4 \le t < 6$ hours **A1**

(b) Adults who exercise for 6 or more hours are in classes $6 \le t < 8$, $8 \le t < 10$, $10 \le t < 12$.
Number $= 14 + 10 + 4 = 28$ adults **A1**

(c) Midpoints $1, 3, 5, 7, 9, 11$ with frequencies $4, 10, 18, 14, 10, 4$. GDC: 1-Var Stats **M1**

$$\hat{\bar{x}} = \frac{1(4) + 3(10) + 5(18) + 7(14) + 9(10) + 11(4)}{60} = \frac{4 + 30 + 90 + 98 + 90 + 44}{60} = \frac{356}{60}$$

$= 5.93\ldots \approx 5.93$ hours **A1**

Note: The mean is an estimate because individual values within each class are unknown.

Q16 — Box Plot Construction and Comparison MEDIUM

(a) Plot the five-number summary accurately: a whisker from 0 to 2, a box from $Q_1 = 2$ to $Q_3 = 5$ with a median line at 3, and a whisker from 5 to 9 **M1A1**

(b) Any one valid feature stated in context, e.g. the distribution is positively skewed (the right whisker is longer than the left whisker) **A1**

(c) A comparison must include one centre statement AND one spread statement, both in context.
The second group has a higher median (4 cups vs 3 cups), suggesting workers in the second group drink more coffee on average. The second group has a smaller IQR (1 cup vs 3 cups), indicating more consistent coffee consumption. **A1**

Note: A comparison of two distributions must reference both centre and spread in context to earn the mark.

Q17 — Systematic Sampling — Interval; Procedure; Disadvantage MEDIUM

(a) Sampling interval $= \dfrac{240}{20} = 12$ **A1**

(b) Number the students on the list from 1 to 240. Select a random starting point between 1 and 12 (inclusive) **M1**

Then select every 12th student after the starting point (e.g. if start is 7: select students $7, 19, 31, \ldots, 235$) **A1**

(c) One valid disadvantage: if the list has a periodic pattern with period 12 (or a multiple), the systematic sample may be biased because every selected student falls in the same position within each period **A1**

Note: The starting point must be chosen randomly from the first interval (1 to 12); starting always at 1 makes the sample non-random.

Q18 — Mean from Raw Data; New Data Value from Updated Mean MEDIUM

(a) $\bar{x} = \dfrac{24 + 27 + 31 + 34 + 38 + 41 + 45 + 52 + 68}{9} = \dfrac{360}{9} = 40$ years **A1**

(b) GDC: from the same 1-Var Stats output, read σ_x.
$\sigma_x = 12.9$ years (unrounded: $12.910\ldots$) **A1**

(c) Let the tenth member's age be a. New mean with 10 members is 41, so new total sum is $10 \times 41 = 410$ **M1**

$a = 410 - 360 = 50$ years **A1**

Note: A common error is to use the old mean (40) instead of first finding the old total sum (360); always work with total sums when a new member is added.

Q19 — Stratified Sampling: Total and Department Allocations MEDIUM

(a) Total employees $= 120 + 80 + 50 = 250$ **A1**

(b) Sampling fraction $= \dfrac{50}{250} = 0.2$ **M1**

Operations: $120 \times 0.2 = 24$; Sales: $80 \times 0.2 = 16$; Finance: $50 \times 0.2 = 10$ **A1A1**

Check: $24 + 16 + 10 = 50$ ✓

Note: Each department allocation must be a whole number of people; check the allocations sum exactly to 50.

Q20 — Frequency Table — Mean; Median; Mode MEDIUM

(a) $\bar{x} = \dfrac{4(2)+5(3)+6(7)+7(6)+8(4)+9(2)}{24} = \dfrac{8+15+42+42+32+18}{24} = \dfrac{157}{24} = 6.5416\ldots \approx 6.54$ **M1A1**

(b) With 24 values the median is the mean of the 12th and 13th values. Cumulative frequencies: score 4: 2; score 5: 5; score 6: 12; score 7: 18. The 12th value is 6, the 13th is 7.

Median $= \dfrac{6+7}{2} = 6.5$ **A1**

(c) Mode $= 6$ (highest frequency, 7) **A1**

Note: Build the cumulative frequency carefully to locate the median in a frequency table; do not simply use the middle score value.

Q21 — Cumulative Frequency: IQR and Outlier Test HARD

(a) The median corresponds to the 100th value; reading across from 100 to the curve, then down.
Median $= 15$ cm **A1**

(b) Q_1 is the value at cumulative frequency 50 (25th percentile of 200 values); Q_3 is the value at cumulative frequency 150.
$Q_1 = 11.25$ cm, $Q_3 = 20$ cm **A1A1**

$IQR = 20 - 11.25 = 8.75$ cm

(c) Upper outlier fence $= Q_3 + 1.5 \times IQR = 20 + 1.5 \times 8.75 = 20 + 13.125 = 33.125$ cm **M1**

Since $27.5 < 33.125$, the value 27.5 cm does NOT exceed the upper fence **R1**

Therefore 27.5 cm is not an outlier.

Note: Premature rounding of the IQR (e.g. to 8.8) would shift the fence and could reverse the verdict.

Q22 — Grouped Data Mean (AG) and Linear Transformation HARD

(a) Modal class: $9.95 \leq d < 10.00$ (highest frequency, 31) **A1**

(b) Midpoints: $9.825, 9.875, 9.925, 9.975, 10.025, 10.075, 10.125$ **M1**

$\sum fx = 3(9.825) + 11(9.875) + 24(9.925) + 31(9.975) + 18(10.025) + 9(10.075) + 4(10.125)$
$= 29.475 + 108.625 + 238.200 + 309.225 + 180.450 + 90.675 + 40.500 = 997.15$

$n = 3 + 11 + 24 + 31 + 18 + 9 + 4 = 100$; $\hat{\mu} = \dfrac{997.15}{100} = 9.9715$ mm **AG**

Note: This is a "show that" (AG); the unrounded value 997.15 must appear before the printed target. Using class boundaries instead of midpoints gives an incorrect $\sum fx$.

(c) $\bar{I} = 0.03937 \times 9.9715 = 0.392578\ldots \approx 0.393$ inches **M1**

GDC: $\sigma_d = 0.0679$ mm (unrounded $0.06792\ldots$); $\sigma_I = 0.03937 \times 0.0679 = 0.00267$ inches **A1**

Note: There is no additive constant in $I = 0.03937d$, so $\sigma_I = |0.03937| \times \sigma_d$ only.

Q23 — Comparing Two Box Plots — IQR and Conditional Probability HARD

(a) Northside median $= 19$ years; Southside median $= 25$ years **A1**

(b) Northside: $Q_1 = 11$, $Q_3 = 28$, $IQR_N = 17$ years. Southside: $Q_1 = 15$, $Q_3 = 32$, $IQR_S = 17$ years **A1**

(c) Centre: the typical age of Northside members (median 19) is lower than Southside (median 25), so Northside tends to attract younger players. **A1**

Spread: both clubs have the same interquartile range (17 years). Northside's maximum age (40 years) is lower than Southside's (47 years), giving Northside the smaller overall range (Northside: $40 - 7 = 33$ years; Southside: $47 - 10 = 37$ years), so Southside shows slightly more extreme ages at the top end. **A1**

Note: Comparisons must be in context and include BOTH a centre statement and a spread statement.

(d) $Q_1 = 11$ years and the median is 19 years. Since 19 is exactly the median, and for a continuous variable $P(\text{age} < \text{median}) = 0.5$ by definition of the median,
$P(\text{age} < 19) = 0.5$, assuming ages are continuously distributed. **R1**

Note: No further modelling is needed once 19 is recognised as exactly the median from part (a); do not overcomplicate with a within-quartile uniformity assumption that isn't required here.

Q24 — Histogram Mean Estimate and Critique of Generalisation HARD

(a) Sum the frequencies: $6 + 14 + 26 + 22 + 8 + 4 = 80$. ✓ **A1**

(b) Midpoints: $12.5, 17.5, 22.5, 27.5, 32.5, 37.5$ **M1**

$\sum fx = 6(12.5) + 14(17.5) + 26(22.5) + 22(27.5) + 8(32.5) + 4(37.5)$
$= 75 + 245 + 585 + 605 + 260 + 150 = 1920$

$\bar{x} = \dfrac{1920}{80} = 24$ hours **A1**

Note: Using class boundaries instead of midpoints is the most common error here.

(c) Two valid reasons (one mark each):
Reason 1: The data were collected from only one afternoon at one supermarket; extrapolating this sample mean to "all teenagers" goes beyond the data collected. **A1**

Reason 2: The mean is an estimate based on grouped data using midpoints, so claiming an exact figure for a different, wider population is unjustified. **A1**

Q25 — Outlier Test (AG) and Linear Transformation HARD

(a) The 13 values in order: $38, 45, 52, 60, 63, 67, 71, 74, 78, 82, 89, 95, 112$.

Median $=$ 7th value $=$ 71 km; Q_1 $=$ median of lower 6 $= \dfrac{52+60}{2} = 56$ km; Q_3 $=$ median of upper 6 $= \dfrac{82+89}{2} = 85.5$ km **A1**

(b) $IQR = 85.5 - 56 = 29.5$ km **M1**

Upper fence $= Q_3 + 1.5 \times IQR = 85.5 + 1.5(29.5) = 85.5 + 44.25 = 129.75$ km

Since $112 < 129.75$, 112 km does not exceed the upper fence. **AG**

*Note: This is a "show that" (AG) confirming 112 km is **not** an outlier by this fence; the unrounded fence value 129.75 must be shown.*

(c) GDC: $\bar{d} = 71.2$ km (unrounded $71.231\ldots$), $\sigma_d = 19.8$ km (unrounded $19.780\ldots$)

$\bar{M} = 0.621 \times 71.231 = 44.2\ldots \approx 44.2$ miles **M1**

$\sigma_M = 0.621 \times 19.780 = 12.3\ldots \approx 12.3$ miles **A1**

Note: There is no additive constant in $M = 0.621d$, so $\sigma_M = |0.621| \times \sigma_d$ only.

Q26 — Frequency Table Statistics and Linear Transformation HARD

(a) GDC: enter values 1–7 in List 1, frequencies $4, 9, 14, 20, 17, 10, 6$ in List 2; 1-Var Stats **M1**

$$\bar{x} = \frac{4(1)+9(2)+14(3)+20(4)+17(5)+10(6)+6(7)}{80} = \frac{4+18+42+80+85+60+42}{80} = \frac{331}{80} = 4.1375 \approx 4.14 \text{ items}$$

$\sigma_x = 1.56$ items (unrounded $1.5632\ldots$) **A1**

Total $n = 4+9+14+20+17+10+6 = 80$.

(b) $C = 2.80x + 1.50$, a linear transformation with $a = 2.80$, $b = 1.50$.
Mean of C: $\bar{C} = 2.80 \times 4.1375 + 1.50 = 11.585 + 1.50 = \$13.1$ (3 s.f.) **M1**

Standard deviation of C: $\sigma_C = |2.80| \times \sigma_x = 2.80 \times 1.5632 = \$4.38$ (3 s.f.) **A1**

Note: The additive constant \$1.50 does NOT affect the standard deviation; only the scale factor 2.80 multiplies σ.

(c) Threshold $= \bar{x} + 2\sigma_x = 4.1375 + 2(1.5632) = 4.1375 + 3.1264 = 7.264$ **M1**

Since $7 < 7.264$, a customer purchasing 7 items does NOT exceed the threshold **R1**

Therefore 7 items is not an outlier by this rule.

Q27 — Outlier Fences (Both Tails) and Comparison HARD

(a) $IQR = Q_3 - Q_1 = 25.1 - 18.4 = 6.7$ mm **M1**

Lower fence $= Q_1 - 1.5 \times IQR = 18.4 - 1.5(6.7) = 18.4 - 10.05 = 8.35$ mm
Upper fence $= Q_3 + 1.5 \times IQR = 25.1 + 1.5(6.7) = 25.1 + 10.05 = 35.15$ mm **A1**

(b) Minimum $= 12.3$ mm: since $12.3 > 8.35$, it is NOT below the lower fence **R1**

Therefore the minimum value is not an outlier. **A1**

Maximum $= 33.8$ mm: since $33.8 < 35.15$, it does not exceed the upper fence.
Therefore the maximum value is also not an outlier.

Note: Both fences must be compared against the correct value (minimum vs lower fence, maximum vs upper fence).

(c) The first sample has a higher median (21.7 mm vs 19.4 mm), so its beetles are typically longer. The second subspecies has a larger IQR (9.6 mm vs 6.7 mm), indicating its lengths are more variable (less consistent). **A1**

Q28 — Grouped Mean (AG) — Standard Deviation and Median Justification HARD

(a) Midpoints: $0.25, 0.75, 1.25, 1.75, 2.25, 2.75$ kg **M1**

$\sum fx = 8(0.25) + 19(0.75) + 31(1.25) + 24(1.75) + 12(2.25) + 6(2.75)$
$= 2.00 + 14.25 + 38.75 + 42.00 + 27.00 + 16.50 = 140.50$

$\hat{\mu} = \dfrac{140.50}{100} = 1.405$ kg **AG**

Note: This is a "show that" (AG); the unrounded value 140.50 must appear before the printed target.

(b) GDC: from the same 1-Var Stats entry, $\sigma_x = 0.647$ kg (unrounded $0.6469\ldots$) **A1**

(c) The median is the average of the 50th and 51st values. From the cumulative frequency column, 27 values lie below 1.0 kg and 58 values lie below 1.5 kg **M1**

Since $27 < 50 \leq 58$ and $27 < 51 \leq 58$, both the 50th and 51st values fall in the class $1.0 \leq m < 1.5$ **R1**

Therefore the biologist's claim is correct; the median does lie in the interval $1.0 \leq m < 1.5$.

Q29 — Modal Class — Grouped Mean Estimate and SD Threshold HARD

(a) Modal class: $65 \leq s < 80$ (highest frequency, 50) **A1**

(b) Midpoints: $27.5, 42.5, 57.5, 72.5, 87.5$ **M1**

$\sum fx = 12(27.5) + 28(42.5) + 45(57.5) + 50(72.5) + 25(87.5)$
$= 330 + 1190 + 2587.5 + 3625 + 2187.5 = 9920$

$\hat{\mu} = \dfrac{9920}{160} = 62$ (estimated mean score) **A1**

Note: The answer is an estimate; midpoints must be used, not class boundaries.

(c) GDC: $\sigma = 17.25$ (unrounded, exact) ≈ 17.3 (3 s.f.)

Threshold $= \bar{x} + 1.5\sigma = 62 + 1.5(17.25) = 62 + 25.875 = 87.875$ **M1**

Since $88 > 87.875$, a score of 88 does exceed $\mu + 1.5\sigma$ **R1**

Therefore a student scoring 88 is more than 1.5 standard deviations above the estimated mean; such students could reasonably be awarded a distinction.

Note: The margin here is thin (88 vs 87.875) — always carry the unrounded $\sigma = 17.25$ through the threshold calculation rather than a prematurely rounded value, to avoid reversing a borderline verdict.

Q30 — Summary Statistics (AG x2) and Threshold Decision HARD

(a) $\hat{\mu} = \dfrac{\sum fx}{\sum f} = \dfrac{5715}{90} = 63.5$ bpm **AG**

Note: This is a "show that"; the unrounded division $5715 \div 90 = 63.5$ must be written explicitly.

(b) $\sigma^2 = \dfrac{\sum fx^2}{\sum f} - \bar{x}^2$ **M1**

$\sigma^2 = \dfrac{366489}{90} - 63.5^2 = 4072.10 - 4032.25 = 39.85$

$\sigma = \sqrt{39.85} = 6.31$ bpm **AG**

Note: This is a "show that"; the variance formula requires the unrounded mean. The standard deviation is NOT $\sqrt{\frac{\sum fx^2 - (\sum fx)^2}{\sum f}}$ without the correct grouping.

(c) Threshold $= \mu + 2\sigma = 63.5 + 2(6.313) = 63.5 + 12.626 = 76.1$ bpm **M1**

Since $82 > 76.1$, the value 82 bpm exceeds $\mu + 2\sigma$ **R1**

Therefore a resting heart rate of 82 bpm is classified as "unusually high" according to the guideline.