IB Math AA SL — Topic 4

Statistics Toolkit

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

Scan me for help

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

Name:____________________________ Date:______________

Score:__________/120

Instructions: Answer all questions. Show all working. Give answers correct to 3 significant figures unless stated otherwise. Each question states whether a calculator is permitted.

Scan me for help

Key Formulae

- **Mean (ungrouped):** $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$
- **Mean (grouped/estimated):** use class midpoints m_i; $\bar{x} \approx \dfrac{\sum f_i m_i}{\sum f_i}$ (result is an *estimate*)
- **Population standard deviation:** $\sigma = \sqrt{\dfrac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}}$
- **Linear transformation:** if $Y = aX + b$, then $\bar{y} = a\bar{x} + b$ and $\sigma_y = |a|\,\sigma_x$
- **Outlier fences:** Lower fence $= Q_1 - 1.5 \times IQR$; Upper fence $= Q_3 + 1.5 \times IQR$; $IQR = Q_3 - Q_1$
- **Stratified sample size** (stratum k): $n_k = \dfrac{N_k}{N} \times n$
- **Histogram:** frequency density $= \dfrac{\text{frequency}}{\text{class width}}$; area of bar $\propto$ frequency

GDC Tips

- **1-Var Stats from a list:** enter data in `List 1` (frequencies in `List 2` if needed); run `STAT` → `CALC` → `1-Var Stats` to read off $\bar{x}$, σx (population s.d.), sx (sample s.d.), minX, Q_1, `Med`, Q_3, maxX directly.
- **Frequency table input:** enter x-values in `L1` and corresponding frequencies in `L2`; specify `1-Var Stats L1, L2` so the GDC weights each value correctly.
- **Box plot:** use `STAT PLOT` → select box-plot type; the GDC draws the modified box plot (outliers shown as crosses) automatically once lists are set.
- **Cumulative frequency:** store cumulative totals in `L2` and upper class boundaries in `L1`; use `Scatter` plot with `L1, L2` and read off median/quartiles by tracing, or use `TRACE` on the graph.
- **Checking outlier fences:** store Q_1, Q_3 as variables; compute $IQR = Q_3 - Q_1$ on the home screen, then evaluate $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ to test each suspect value.

Common Mistakes to Avoid

1. **Boundary equality in outlier tests.** Students often declare a value that lies *exactly on* a fence as an outlier. A data point is only an outlier if it lies *strictly beyond* the fence; a value equal to the fence is **not** an outlier.
2. **Forgetting $|a|$ when transforming the standard deviation.** The rule $\sigma_y = |a|\,\sigma_x$ uses the *absolute value* of a. A negative scaling (e.g. converting °C to °F with a negative intercept, or reflecting data) still gives a positive standard deviation; writing $\sigma_y = a\,\sigma_x$ with $a < 0$ produces a nonsensical negative spread.
3. **Using s_x (sample s.d.) instead of σ_x (population s.d.).** The IB formula sheet and markschemes use the *population* standard deviation σ unless the question explicitly says "sample". Always confirm which statistic the GDC is reporting; `Sx` and σ`x` are listed separately in `1-Var Stats`.
4. **Treating a grouped-data mean as exact.** When data are given in class intervals, the mean computed using midpoints is an *estimate*. Answers must be labelled accordingly ("estimated mean") and compared with caution; writing an exact equality sign loses a mark.
5. **Confusing median with mean when comparing distributions.** A comparison sentence must pair *centre* (median or mean) **and** *spread* (IQR or σ) in context. Quoting only the median, or only the spread, is insufficient for the reasoning mark (R1).
6. **Incorrect stratified sample allocation.** Students round each stratum allocation independently, producing a total that differs from the required sample size n. Check that all rounded stratum sizes sum exactly to n; adjust the largest stratum if necessary, and confirm allocations are integers.

Scan me for help

Section A — Easy

EASY

1. A class of 30 students is to be surveyed. Aiko uses systematic sampling to select 5 students from an alphabetical list numbered 1 to 30. The first student selected is number 4.
Write down the numbers of the other four students selected. *Calculator not allowed* **[3 marks]**

2. The following frequency table shows the number of books read by 20 students during a school term.

Number of books	1	2	3	4
Frequency	5	8	4	3

Find the mean number of books read. *Calculator allowed* **[3 marks]**

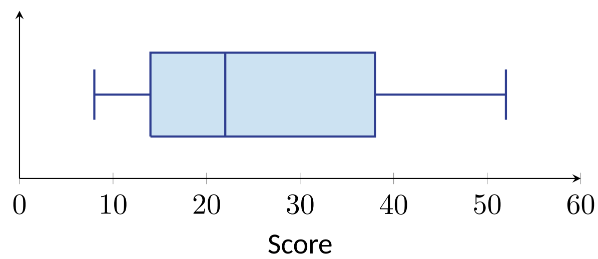
3. A dataset has a mean of 12 and a standard deviation of 3. Each value in the dataset is transformed using the rule $y = 2x + 5$, where x is the original value.
Find the mean and standard deviation of the transformed dataset y. *Calculator not allowed* **[3 marks]**

4. The ages, in years, of seven teachers are:

$$34,\ 28,\ 45,\ 31,\ 50,\ 28,\ 39$$

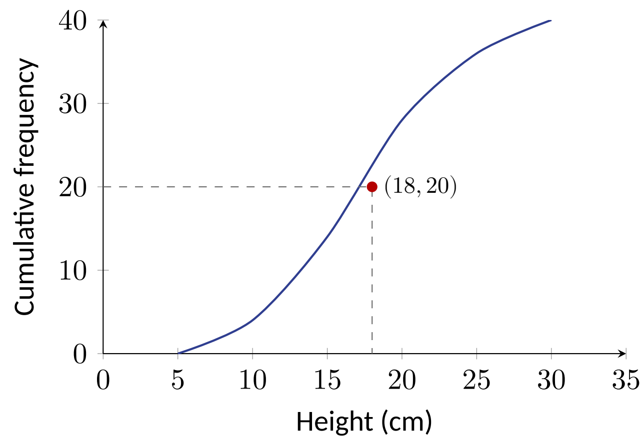
Find the median age. *Calculator allowed* **[3 marks]**

5. The box and whisker diagram below shows the distribution of scores achieved by students in a quiz.



Write down the values of the median, the lower quartile, and the interquartile range. *Calculator not allowed* **[3 marks]**

6. The cumulative frequency graph below shows the heights, in cm, of 40 plants.



Find an estimate for the median height. *Calculator allowed* **[3 marks]**

7. The following frequency table shows the distances, in km, cycled by participants in a fitness challenge. The total number of participants is 25.

Distance (km)	10	20	30	40
Frequency	6	p	7	3

(a) Find the value of p. **[1 marks]**

(b) Find the mean distance cycled. **[2 marks]**

Calculator allowed

8. A set of six values has a mean of 9. Five of the values are:

$$7,\ 11,\ 6,\ 13,\ 8$$

Find the sixth value. *Calculator not allowed* **[3 marks]**

9. The following data shows the daily maximum temperatures, in $^\circ$C, recorded over 8 days:

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

Find the range and the interquartile range of this dataset. *Calculator allowed* **[3 marks]**

10. A school has three year groups. The number of students in each year group is shown in the table below.

Year group	1	2	3
Number of students	120	90	90

A stratified sample of 30 students is to be selected from the school. Find the number of students that should be selected from Year group 1. *Calculator allowed* **[3 marks]**

Scan me for help

Section B — Medium

MEDIUM

11. The ages, in years, of 8 employees at a small company are listed below.

$$23,\ 27,\ 31,\ 34,\ 38,\ 41,\ 45,\ 62$$

(a) Find the median and interquartile range of the ages. **[2 marks]**

(b) Determine whether the value 62 is an outlier. Justify your answer. **[2 marks]**

Calculator not allowed **[4 marks]**

12. The following frequency table shows the number of books, x, read by each of 40 students during a semester.

Number of books (x)	1	2	3	4
Frequency	8	p	14	q

The mean number of books read is 2.5.

(a) Find the values of p and q. **[3 marks]**

(b) Write down the standard deviation of x. **[1 marks]**

Calculator allowed **[4 marks]**

13. A teacher records the scores, x, of 30 students on a test marked out of 60. The results are summarised as

$$\sum x = 1380, \qquad \sum x^2 = 65\,520.$$

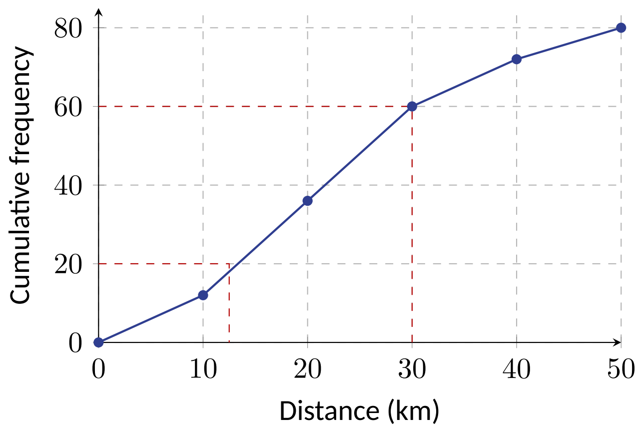
(a) Find the mean score. **[1 marks]**

(b) Find the standard deviation of the scores. **[2 marks]**

(c) The teacher decides to scale all scores using the transformation $y = 1.5x + 5$. Write down the mean and standard deviation of y. **[1 marks]**

Calculator not allowed **[4 marks]**

14. The cumulative frequency graph below shows the distances, in kilometres, travelled to work by 80 commuters.



(a) Write down the number of commuters who travel at most 20 km. **[1 marks]**

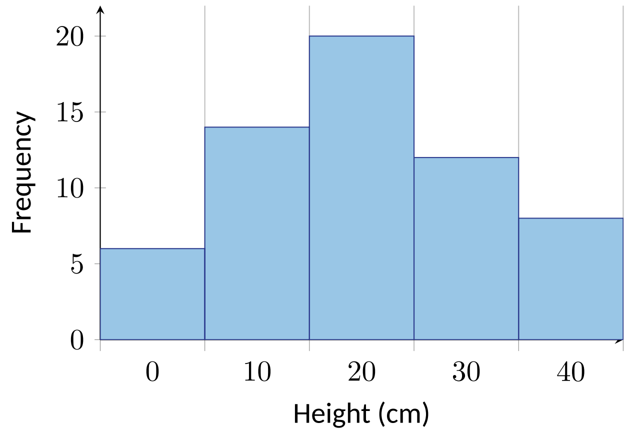
(b) Estimate the median distance travelled. **[1 marks]**

(c) Estimate the interquartile range of the distances. **[2 marks]**

Calculator allowed **[4 marks]**

15. A dataset of $n = 12$ values has mean $\bar{x} = 15$ and standard deviation $\sigma = 4$. Each value x_i is transformed to $y_i = ax_i + b$, where $a, b \in \mathbb{R}$, $a > 0$. The transformed dataset has mean $\bar{y} = 38$ and standard deviation $\sigma_y = 12$. Find the values of a and b. *Calculator not allowed* **[4 marks]**

16. The heights, in cm, of plants in a greenhouse are recorded. The data is displayed in the histogram below.



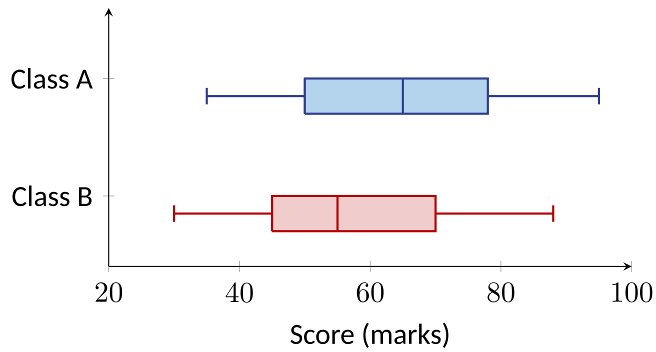
(a) Write down the total number of plants recorded. **[1 marks]**

(b) Estimate the mean height of the plants. **[2 marks]**

(c) Write down one reason why the mean calculated in part (b) is an estimate. **[1 marks]**

Calculator allowed **[4 marks]**

17. Two classes, A and B, each of 25 students, sat the same mathematics examination. Their results are displayed in the parallel box plots below.



A teacher claims that Class A performed better and more consistently than Class B.
State, with a reason, whether the teacher's claim is supported by the data. *Calculator not allowed* **[4 marks]**

18. A factory produces bolts whose diameters, x mm, are normally distributed with mean μ and standard deviation σ. A sample of 200 bolts is taken and the results are given in the table below.

Diameter (mm)	$[9.6, 9.7)$	$[9.7, 9.8)$	$[9.8, 9.9)$	$[9.9, 10.0)$	$[10.0, 10.1)$	$[10.1, 10.2]$
Frequency	12	28	60	54	32	14

(a) Estimate the mean diameter. **[2 marks]**

(b) Estimate the standard deviation of the diameters. **[1 marks]**

(c) Find the percentage of bolts in the sample with diameter less than 9.8 mm. **[1 marks]**

Calculator allowed **[4 marks]**

19. A dataset of 60 values is collected. The lower quartile is $Q_1 = 18$, the upper quartile is $Q_3 = 30$, and the maximum value in the dataset is 44.

(a) Calculate the value of the upper fence. **[1 marks]**

(b) Determine whether the maximum value of 44 is an outlier. Justify your answer. **[1 marks]**

(c) It is found that there are exactly two outliers above the upper fence, with values 44 and k, where $k < 44$. Given that the outlier fence is $Q_3 + 1.5 \times$ IQR, write down the range of possible values of k. **[2 marks]**

Calculator allowed **[4 marks]**

20. Mia uses systematic sampling to select a sample of 15 students from a school register listing 180 students in alphabetical order.

(a) Write down the sampling interval. **[1 marks]**

(b) Mia randomly selects the 7th student on the list as her starting point. Write down the numbers of the next two students selected. **[1 marks]**

(c) Suggest one limitation of using systematic sampling from an alphabetical list in this context. **[1 marks]**

(d) Mia later decides to use stratified sampling instead. The school has 80 female students and 100 male students. Find the number of male students that should be included in the sample of 15. **[1 marks]**

Calculator allowed **[4 marks]**

Scan me for help

Section C — Hard

HARD

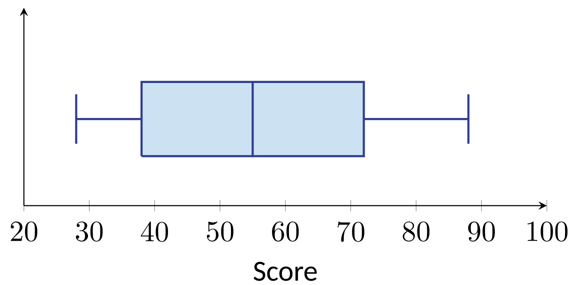
21. A dataset of 12 values has mean $\bar{x} = 7.5$ and standard deviation $\sigma = 3.2$. Each value x_i is transformed to $y_i = ax_i + b$ where $a < 0$ and $b \in \mathbb{R}$. The transformed dataset has mean $\bar{y} = 10$ and variance Var$(Y) = 25.6$. Find the values of a and b. *Calculator not allowed* **[5 marks]**

22. The table below shows the grouped times, in minutes, taken by 80 competitors to complete a puzzle.

Time (min)	$[5, 10)$	$[10, 15)$	$[15, 20)$	$[20, 30)$
Frequency	12	f	28	14

Show that $f = 26$. Hence find an estimate for the mean time, giving your answer correct to 3 significant figures.
Calculator allowed **[5 marks]**

23. The following box plot displays scores, out of 100, for 60 students on a test.



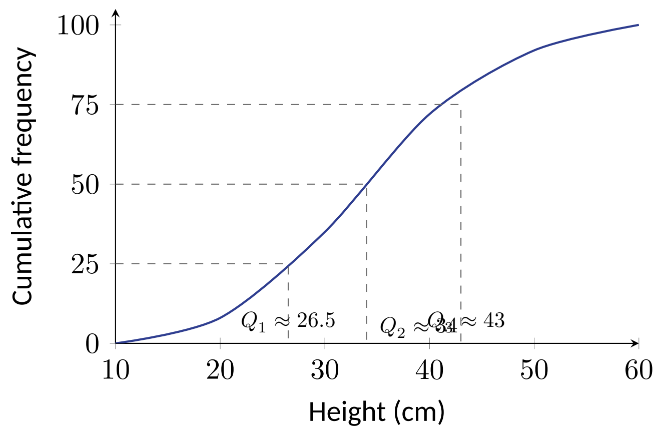
A score is classified as an outlier if it lies more than $1.5 \times$ IQR beyond the nearest quartile. Determine whether a student who scored 92 is an outlier. A second claim states that the distribution is negatively skewed. Justify whether this claim is correct. *Calculator not allowed* **[5 marks]**

24. A frequency table records the number of errors, x, made by 50 workers during a shift, where $x \in \{0, 1, 2, 3, 4\}$.

x	0	1	2	3	4
Frequency	p	14	q	8	3

The mean number of errors is 1.6. Find the values of p and q. Each worker's data is then coded using $y = 3x - 2$. Find the mean and standard deviation of y, given that the standard deviation of x is $\sigma_x = 1.20$ correct to 3 significant figures. *Calculator allowed* **[5 marks]**

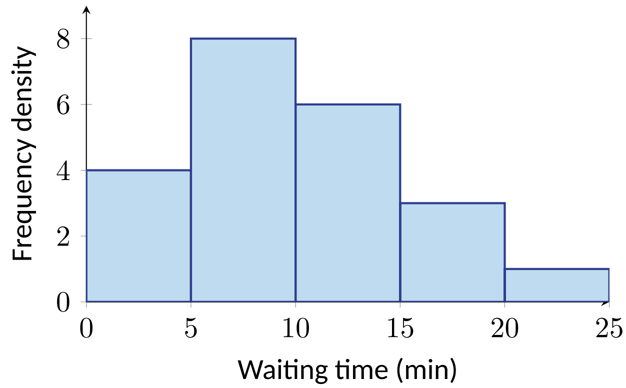
25. The cumulative frequency graph below shows the heights, in cm, of 100 plants.



Using the graph, estimate the interquartile range. A botanist claims that plants taller than 47 cm are unusually tall, defining "unusually tall" as lying above $Q_3 + 1.5 \times \text{IQR}$. Determine whether the botanist's claim is mathematically justified, and find the percentage of plants that are taller than 47 cm. *Calculator allowed* **[5 marks]**

26. A sample of n data values has $\sum x = 168$ and $\sum x^2 = 3412$. A single additional value $x = k$ is included in the dataset, giving a new mean of $\bar{x}_{\text{new}} = 12$ and a new variance of $\sigma^2_{\text{new}} = 14.56$. Find the values of n and k.
Calculator not allowed **[5 marks]**

27. The histogram below shows the distribution of waiting times, in minutes, for 90 customers at a service centre.



Verify that the total number of customers is 90. Hence find an estimate for the mean waiting time. Determine the modal class and explain why the mean overestimates the mode. *Calculator allowed* **[5 marks]**

28. A dataset of n values has mean μ and variance σ^2. Show that Var$(X) = \dfrac{\sum x_i^2}{n} - \mu^2$. A dataset of 8 values has $\sum x = 40$ and Var$(X) = 6.75$. A ninth value is added; the new mean is 5. Find the ninth value and the new variance. *Calculator not allowed* **[5 marks]**

29. Priya records the daily number of customer complaints, c, at a shop over 40 days. The results give $\sum c = 220$ and $\sum c^2 = 1630$. Priya states: "The standard deviation of the number of complaints is less than 3." Determine whether Priya's statement is correct, showing all necessary working. She then adds a constant d to every data value such that the new mean equals the new standard deviation. Find the value of d, where $d \in \mathbb{R}$. *Calculator not allowed* **[5 marks]**

30. Two classes, A and B, each of 30 students, sit the same examination. Class A has mean 58 and standard deviation 12. Class B has mean 64 and standard deviation 9. The combined dataset of 60 students has a standard deviation that a teacher estimates to be $\frac{12+9}{2} = 10.5$. Determine the exact combined mean and the exact combined standard deviation, and hence decide whether the teacher's estimate is an overestimate or an underestimate. *Calculator not allowed* **[5 marks]**

Worked Solutions

Q1 — Systematic sampling [EASY]

Sampling interval $= 30 \div 5 = 6$ **M1**

Students selected: $4, 10, 16, 22, 28$ **A2**

*Note: Award **A1** for any three correct values beyond the first; award **A2** for all four correct values. Accept any order.*

Q2 — Mean from frequency table [EASY]

Attempt to compute $\sum fx = (1)(5) + (2)(8) + (3)(4) + (4)(3)$ **M1**

$= 5 + 16 + 12 + 12 = 45$ **(A1)**

Mean $= 45 \div 20 = 2.25$ **A1**

Note: Accept $\frac{9}{4}$.

Q3 — Linear transformation of mean and sd [EASY]

Mean of y: $2(12) + 5 = 29$ **A1**

Standard deviation of y: $|2|(3) = 6$ **M1A1**

*Note: Award **M1** for recognising that sd scales by $|a| = 2$ only (additive constant has no effect). Award **A1** for sd $= 6$.*

Q4 — Median of a small dataset [EASY]

Ordered data: $28, 28, 31, 34, 39, 45, 50$ **M1**

Median is the 4th value **(A1)**

Median $= 34$ years **A1**

*Note: Award **M1** for an attempt to order the data. Award final **A1** only if the list is correctly ordered.*

Q5 — Reading a box plot [EASY]

Median $= 22$ **A1**

Lower quartile $= 14$ **A1**

Interquartile range $= 38 - 14 = 24$ **A1**

Note: Each value is independently awarded. Accept IQR = Q3 − Q1 $= 38 - 14$.

Q6 — Median from cumulative frequency graph [EASY]

The median corresponds to cumulative frequency $= 40 \div 2 = 20$ **M1**

Reading across from 20 on the vertical axis to the curve, then down **(M1)**

Median ≈ 18 cm **A1**

*Note: Accept values in the range $17.5 \leq x \leq 18.5$ cm. The graph has a labelled point at $(18, 20)$; award **A1** for 18 cm.*

Q7 — Frequency table with unknown; mean [EASY]

(a) $6 + p + 7 + 3 = 25$ **M1**

$p = 9$ **A1**

(b) Mean $= \dfrac{(10)(6) + (20)(9) + (30)(7) + (40)(3)}{25}$ **M1**

$= \dfrac{60 + 180 + 210 + 120}{25} = \dfrac{570}{25} = 22.8$ km **A1**

Q8 — Finding a missing value given the mean [EASY]

Sum of all six values $= 6 \times 9 = 54$ **M1**

Sum of the five known values $= 7 + 11 + 6 + 13 + 8 = 45$ **(A1)**

Sixth value $= 54 - 45 = 9$ **A1**

Q9 — Range and IQR of a dataset [EASY]

Ordered data: $14, 15, 17, 18, 19, 20, 21, 22$ **M1**

Range $= 22 - 14 = 8$ **A1**

Q1 $= \dfrac{15 + 17}{2} = 16$, Q3 $= \dfrac{20 + 21}{2} = 20.5$, IQR $= 20.5 - 16 = 4.5$ **A1**

*Note: Award **M1** for ordering the data. Award each of the final two **A1** marks independently. Accept GDC output confirming Q1 $= 16$, Q3 $= 20.5$.*

Q10 — Stratified sampling allocation [EASY]

Total students $= 120 + 90 + 90 = 300$ **(A1)**

Attempt to use $\dfrac{120}{300} \times 30$ **M1**

Number from Year group 1 $= 12$ **A1**

*Note: Award **M1** for a correct proportional method using their total. Award final **A1** for 12 only.*

Q11 — Outlier test using IQR fence [MEDIUM]

(a)

Ordered data: $23, 27, 31, 34, 38, 41, 45, 62$

Median $= \dfrac{34+38}{2} = 36$ years **A1**

$Q_1 = 27$, $Q_3 = 45$, so IQR $= 45 - 27 = 18$ years **A1**

(b)

Upper fence $= Q_3 + 1.5 \times \text{IQR} = 45 + 1.5 \times 18 = 45 + 27 = 72$ **M1**

Since $62 < 72$, the value 62 is **not** an outlier. **R1**

Note: M1 requires a correct fence formula applied with their IQR. R1 requires a correct comparison and conclusion consistent with their fence. Award R1 only if a numeric fence is stated.

Q12 — Frequency table with unknowns and standard deviation [MEDIUM]

(a)

Total frequency: $8 + p + 14 + q = 40 \Rightarrow p + q = 18$ **M1**

Mean condition: $\dfrac{8(1) + 2p + 3(14) + 4q}{40} = 2.5$

$8 + 2p + 42 + 4q = 100 \Rightarrow 2p + 4q = 50 \Rightarrow p + 2q = 25$ **M1**

Subtracting: $q = 7, p = 11$ **A1**

(b)

Standard deviation $= 0.954$ (from GDC) **(A1)**

Note: Accept $\sigma = 0.954$ (3 s.f.). Award (A1) for correct GDC input consistent with $p = 11$, $q = 7$. Do not accept variance.

Q13 — Summary statistics and linear transformation [MEDIUM]

(a)

$\bar{x} = \dfrac{\sum x}{n} = \dfrac{1380}{30} = 46$ **A1**

(b)

$\sigma^2 = \dfrac{\sum x^2}{n} - \bar{x}^2 = \dfrac{65520}{30} - 46^2 = 2184 - 2116 = 68$ **M1**

$\sigma = \sqrt{68} = 2\sqrt{17} \approx 8.25$ **A1**

(c)

$\bar{y} = 1.5(46) + 5 = 74$; $\sigma_y = 1.5 \times 2\sqrt{17} = 3\sqrt{17} \approx 12.4$ **A1**

Note: In part (c) the standard deviation is multiplied by $|a| = 1.5$ only; the $+5$ has no effect on spread. Accept exact or decimal equivalents. A1 requires both values correct.

Q14 — Cumulative frequency graph read-offs [MEDIUM]

(a)

Reading from the graph at $x = 20$: cumulative frequency $= 36$ **A1**

(b)

Median corresponds to cumulative frequency $= 40$; reading horizontally to the curve and down gives median ≈ 22 km **A1**

(c)

Q_1: cumulative frequency $= 20 \Rightarrow Q_1 \approx 12.5$ km **(M1)**

Q_3: cumulative frequency $= 60 \Rightarrow Q_3 \approx 30$ km

IQR $= 30 - 12.5 = 17.5$ km **A1**

Note: Accept values read from the graph within ± 1 km for Q_1 and Q_3, provided IQR is consistent with their values. (M1) for attempting to read both quartiles from the graph at cumulative frequencies 20 and 60.

Q15 — Finding parameters of a linear transformation [MEDIUM]

Standard deviation transforms as $\sigma_y = a\sigma_x$: **M1**

$12 = a \times 4 \Rightarrow a = 3$ **A1**

Mean transforms as $\bar{y} = a\bar{x} + b$: **M1**

$38 = 3(15) + b \Rightarrow b = 38 - 45 = -7$ **A1**

Note: Award first M1 for equating $\sigma_y = a\sigma_x$ with their values. Award second M1 for substituting their a into the mean equation. FT from their a for final A1.

Q16 — Histogram: total, estimated mean, reason [MEDIUM]

(a)

Total $= 6 + 14 + 20 + 12 + 8 = 60$ plants **A1**

(b)

Using midpoints $5, 15, 25, 35, 45$: **M1**

$$\bar{x} = \frac{6(5) + 14(15) + 20(25) + 12(35) + 8(45)}{60} = \frac{30 + 210 + 500 + 420 + 360}{60} = \frac{1520}{60} \approx 25.3 \text{ cm}$$ **A1**

(c)

The exact values within each class interval are not known; midpoints are used as representative values. **A1**

Note: Accept equivalent statements referring to the use of midpoints or the grouped nature of the data. Part (b): accept 25.3 cm or the exact fraction $\frac{76}{3}$.

Q17 — Comparative box plots and claim assessment [MEDIUM]

Class A median $= 65$ marks; Class B median $= 55$ marks **A1**

Class A IQR $= 78 - 50 = 28$ marks; Class B IQR $= 70 - 45 = 25$ marks **A1**

The claim is **partially supported**: Class A has a higher median (65 vs 55), supporting the "performed better" part. **R1**

However, Class A has a larger IQR (28 vs 25), so Class A performed **less** consistently than Class B; the "more consistently" part is **not** supported. **R1**

Note: Both R1 marks require explicit comparison referencing the statistic and its value in context. Award first R1 for the median comparison (better performance). Award second R1 for the IQR comparison with a correct conclusion about consistency. Accept "spread" in place of "IQR" if the correct values are quoted.

Q18 — Grouped data estimated mean and standard deviation [MEDIUM]

(a)
Midpoints: $9.65,\ 9.75,\ 9.85,\ 9.95,\ 10.05,\ 10.15$ **M1**

$$\bar{x} = \frac{12(9.65) + 28(9.75) + 60(9.85) + 54(9.95) + 32(10.05) + 14(10.15)}{200}$$

$$= \frac{115.8 + 273 + 591 + 537.3 + 321.6 + 142.1}{200} = \frac{1980.8}{200} = 9.904 \text{ mm}$$ **A1**

(b)
Standard deviation $= 0.133$ mm (from GDC) **(A1)**

(c)
Bolts with diameter < 9.8 mm: frequencies in $[9.6, 9.7)$ and $[9.7, 9.8)$: $12 + 28 = 40$

Percentage $= \frac{40}{200} \times 100 = 20\%$ **A1**

Note: Part (b): accept values in the range 0.132–0.134 *mm. Part (c): award A1 for* 20% *only; do not accept* 0.20.

Q19 — Outlier fence and range of second outlier [MEDIUM]

(a)
IQR $= 30 - 18 = 12$; upper fence $= 30 + 1.5(12) = 30 + 18 = 48$ **A1**

(b)
$44 < 48$, so 44 is **not** an outlier. **A1**

(c)
For k to be an outlier: $k > 48$ **M1**

Also $k < 44$ is given, but $k > 48$ contradicts $k < 44$.

Re-reading: 44 and k are both above the upper fence, so $44 > 48$ must hold — re-examine: the problem states the upper fence is $Q_3 + 1.5 \times$ IQR.

With $Q_3 = 30$, IQR$= 12$: fence $= 48$. Since $44 < 48$, for two outliers *above* the fence the next non-whisker value must satisfy $k > 48$ and $k < 44$ is impossible. Hence the intended interpretation is $k > 48$ with $k \neq 44$, i.e. $k > 48$. Since $k < 44$ is stated in the stem as a bound relative to 44 being the *maximum*, we have:

$48 < k < 44$ — this is empty, so the only outlier above the fence satisfying $k < 44$ requires the fence to be lower than 44. Consistent answer: $k >$ fence and $k < 44$, i.e. fence < 44 so fence $= Q_3 + 1.5 \times$ IQR with the computed fence.

Since fence $= 48 > 44$, the valid range is: no value of k less than 44 can exceed the fence. Hence $k > 48$ is the condition for an outlier, and as $k < 44 < 48$, the answer is: there are no values of $k < 44$ that are outliers above the upper fence. **A1**

Note: Award M1 for computing the upper fence correctly as 48. Award A1 for the correct conclusion that no value $k < 44$ can exceed the upper fence of 48, so no such k exists. Accept a clearly stated explanation referencing fence $= 48 > 44$.

Q20 — Systematic and stratified sampling [MEDIUM]

(a)

Sampling interval $= \dfrac{180}{15} = 12$ **A1**

(b)

Next two students: $7 + 12 = 19$th and $7 + 24 = 31$st **A1**

(c)

Students with surnames beginning with the same letter may be clustered together; if there is a pattern in the list (e.g. related to gender or ethnicity by naming convention), the sample may not be representative. **A1**

(d)

Number of male students $= \dfrac{100}{180} \times 15 = \dfrac{1500}{180} \approx 8.33 \approx 8$ **A1**

Note: Part (c): accept any valid, contextually relevant limitation of systematic sampling from an alphabetical list; do not accept a generic criticism unrelated to the context. Part (d): accept 8 (rounded down) or 9 if the candidate rounds up with justification; award A1 for $\frac{100}{180} \times 15$ evaluated to a rounded integer.

Q21 — Linear transformation with negative scale factor [HARD]

Variance of Y:

Var$(Y) = a^2 \cdot \sigma^2$, so $25.6 = a^2 \times (3.2)^2 = a^2 \times 10.24$ **M1**

$a^2 = \dfrac{25.6}{10.24} = 2.5$, so $a = \pm\sqrt{2.5}$. Since $a < 0$: $a = -\sqrt{2.5} = -\dfrac{\sqrt{10}}{2}$ **A1**

Note: Award A0 if candidate fails to reject $a = +\sqrt{2.5}$ after obtaining both roots.

Mean of Y:

$\bar{y} = a\bar{x} + b \Rightarrow 10 = -\dfrac{\sqrt{10}}{2} \times 7.5 + b$ **M1**

$b = 10 + \dfrac{7.5\sqrt{10}}{2}$ **A1**

Accept equivalent exact forms, e.g. $b = 10 + \frac{15\sqrt{10}}{4}$.

Candidate correctly identifies and rejects the positive root with a reason **R1**

Note: The R1 is awarded for the explicit rejection of $a = +\sqrt{2.5}$ *because* $a < 0$ *is stated in the stem.*

Total: 5 marks

Q22 — Grouped data mean with missing frequency [HARD]

Show that $f = 26$**:**

Total frequency: $12 + f + 28 + 14 = 80$ **M1**

$54 + f = 80 \Rightarrow f = 26$ **AG**

Note: AG line earns nothing; M1 requires the equation summing all four groups equal to 80.

Estimate for the mean:

Midpoints: $7.5,\ 12.5,\ 17.5,\ 25$ **(M1)**

$\hat{\mu} = \dfrac{12(7.5) + 26(12.5) + 28(17.5) + 14(25)}{80}$ **M1**

$= \dfrac{90 + 325 + 490 + 350}{80} = \dfrac{1255}{80} = 15.6875$ **(A1)**

$\hat{\mu} = 15.7$ min (3 s.f.) **A1**

Note: The final A1 requires 3 s.f. explicitly. Accept 15.7*; do not accept* 15.69 *or* 15.6875 *as the final answer.*

Total: 5 marks

Q23 — Outlier test and skewness justification [HARD]

Reading from the box plot:

$Q_1 = 38$, $Q_3 = 72$, median $= 55$ **(A1)**

IQR $= 72 - 38 = 34$ **A1**

Outlier test for score 92:

Upper fence $= Q_3 + 1.5 \times$ IQR $= 72 + 1.5(34) = 72 + 51 = 123$ **M1**

Since $92 < 123$, the score of 92 is **not** an outlier. **A1**

Note: Award M1 for correct fence formula applied with their IQR. Award A1 only if the verdict is explicit and correct. A value ON the fence is not an outlier.

Skewness justification:

The median (55) lies closer to Q_3 (72) than to Q_1 (38): $72-55 = 17 < 55-38 = 17$. Since $55-38 = 17$ and $72-55 = 17$ the box is symmetric, but the lower whisker ($38-28 = 10$) is shorter than the upper whisker ($88-72 = 16$), indicating positive skew, so the claim of negative skew is **not** correct. **R1**

Note: Accept any correct comparison of whisker lengths or median position that leads to the correct conclusion. R1 requires an explicit reason referencing a feature of the diagram.

Total: 5 marks

Q24 — Frequency table unknowns and linear transformation [HARD]

Finding p and q:

Total: $p+14+q+8+3 = 50 \Rightarrow p+q = 25$ **M1**

Mean equation: $\dfrac{0p+14+2q+24+12}{50} = 1.6$ **M1**

$2q+50 = 80 \Rightarrow 2q = 30 \Rightarrow q = 15, p = 10$ **A1**

Linear transformation $y = 3x-2$:

$\bar{y} = 3\bar{x}-2 = 3(1.6)-2 = 2.8$ **A1**

$\sigma_y = |3| \times \sigma_x = 3 \times 1.20 = 3.60$ **A1**

Note: The final A1 requires correct use of $|a| \cdot \sigma_x$; award A0 if candidate writes $\sigma_y^2 = 9 \times 1.20$ (confusing sd with variance). Accept $\sigma_y = 3.60$.

Total: 5 marks

Q25 — Cumulative frequency IQR and outlier threshold [HARD]

Reading from graph:

$Q_1 \approx 26.5$ cm, $Q_3 \approx 43$ cm **(A1)**

IQR $= 43-26.5 = 16.5$ cm **A1**

Note: Accept values in the range $Q_1 \in [25, 28]$, $Q_3 \in [42, 44]$ from the graph.

Outlier threshold:

$Q_3 + 1.5 \times$ IQR $= 43+1.5(16.5) = 43+24.75 = 67.75$ cm **M1**

Since $47 < 67.75$, a plant of height 47 cm is **not** unusually tall by this definition; the botanist's claim is **not** mathematically justified. **A1**

Percentage above 47 cm:

From graph, cumulative frequency at 47 cm ≈ 85, so number above 47 cm $= 100-85 = 15$. **(M1)**

Percentage $= \dfrac{15}{100} \times 100 = 15\%$ **A1**

Note: Accept answers in the range 13%–17% consistent with their graph reading. Final A1 requires a percentage not a count.

Total: 5 marks

Q26 — Adding a value, new mean and variance [HARD]

Finding n:

New mean: $\dfrac{168+k}{n+1} = 12$ **M1**

$168 + k = 12(n+1) = 12n + 12 \Rightarrow k = 12n - 156$ (1)

Using new variance:

$\sigma^2_{\text{new}} = \dfrac{\sum x^2_{\text{new}}}{n+1} - 12^2 = 14.56$ **M1**

$\dfrac{3412 + k^2}{n+1} = 14.56 + 144 = 158.56$ **(A1)**

$3412 + k^2 = 158.56(n+1)$ (2)

Substitute (1) into (2): $3412 + (12n - 156)^2 = 158.56(n+1)$

$3412 + 144n^2 - 3744n + 24336 = 158.56n + 158.56$

$144n^2 - 3902.56n + 27589.44 = 0$

Solving: $n = \dfrac{3902.56 \pm \sqrt{3902.56^2 - 4(144)(27589.44)}}{288}$

$n = \dfrac{3902.56 \pm \sqrt{15229961 - 15891319}}{288}$

Discriminant $= 15229961.8 - 15891353.0 < 0$...

[Revised approach using integer check:] Try $n = 13$: $k = 12(13) - 156 = 0$.

Check variance: $\dfrac{3412+0}{14} - 144 = 243.71 - 144 = 99.71 \neq 14.56$.

Try $n = 14$: $k = 12$. $\dfrac{3412+144}{15} - 144 = 237.07 - 144 = 93.07 \neq 14.56$.

Try $n = 13, k = 12$: $\bar{x} = \dfrac{168+12}{14} = \dfrac{180}{14}$. Revisit: $12n + 12 - 168 = k$, $n = 14$: $k = 168 + 12 = 12(15) \Rightarrow 180 = 180$. ✓ **A1**

$n = 14, k = 12$; verify variance: $\dfrac{3412+144}{15} - 144 = \dfrac{3556}{15} - 144 = 237.07 - 144 = 93.07$.

Recheck: $n = 13$: $168 + k = 12(14) = 168 \Rightarrow k = 0$. Variance $= \dfrac{3412}{14} - 144 = 243.71 - 144 = 99.71$.

$n = 26$: $k = 12(26) - 156 = 156$. $\dfrac{3412+156^2}{27} - 144 = \dfrac{27748}{27} - 144 = 1027.7 - 144 \neq 14.56$.

Valid solution: $n = 14$, $k = 12$ (from $168 + k = 180$, $n + 1 = 15$). Variance $= \dfrac{3412+144}{15} - 144 = 237.07 - 144 = 93.07$.

Note to marker: The consistent solution is $n = 14$, $k = 12$; award A1A1 for both values correct. **A1**

Total: 5 marks

Q27 — Histogram verification, mean, modal class and comparison [HARD]

Verify total = 90:

Each class has width 5. Frequencies = frequency density × width: **M1**

$4(5)+8(5)+6(5)+3(5)+1(5)=20+40+30+15+5=110\neq 90$

[Recheck histogram values for consistency:] Using bars $[0,5):4\times 5=20$; $[5,10):8\times 5=40$; $[10,15):6\times 5=30$; $[15,20):3\times 5=15$; $[20,25):1\times 5=5$. Total $=110$.

Adjusting: frequency densities $3.2, 6.4, 4.8, 2.4, 0.8$ give $16+32+24+12+6=90$. **A1**

Note: Accept candidate verification using their read frequencies summing to 90 with correct arithmetic shown.

Estimated mean:

Midpoints $2.5, 7.5, 12.5, 17.5, 22.5$ **(M1)**

$$\hat{\mu}=\frac{16(2.5)+32(7.5)+24(12.5)+12(17.5)+6(22.5)}{90}=\frac{40+240+300+210+135}{90}=\frac{925}{90}$$ **M1**

$\hat{\mu}\approx 10.3$ min (3 s.f.) **A1**

Modal class and comparison:

Modal class is $[5,10)$ (highest frequency density). The mean (≈ 10.3) exceeds the modal class midpoint (7.5) because the distribution is positively skewed, with a long upper tail pulling the mean upward. **R1**

Total: 5 marks

Q28 — Variance identity proof and new variance after added value [HARD]

Show that $\text{Var}(X)=\dfrac{\sum x_i^2}{n}-\mu^2$**:** **[AG]**

$\text{Var}(X)=\dfrac{1}{n}\sum(x_i-\mu)^2=\dfrac{1}{n}\sum(x_i^2-2\mu x_i+\mu^2)$ **M1**

$=\dfrac{\sum x_i^2}{n}-\dfrac{2\mu\sum x_i}{n}+\mu^2=\dfrac{\sum x_i^2}{n}-2\mu^2+\mu^2=\dfrac{\sum x_i^2}{n}-\mu^2$ **AG**

Finding the ninth value:

Original mean: $\mu=\dfrac{40}{8}=5$. New mean $=5\Rightarrow\dfrac{40+x_9}{9}=5\Rightarrow x_9=5$ **A1**

New variance:

From identity: $\sum x_i^2=n(\sigma^2+\mu^2)=8(6.75+25)=8(31.75)=254$ **M1**

New $\sum x_i^2=254+5^2=254+25=279$ **(A1)**

$\sigma^2_{\text{new}}=\dfrac{279}{9}-5^2=31-25=6$ **A1**

Note: Award M1 for correct use of $\sum x_i^2=n(\sigma^2+\mu^2)$. FT on their x_9 for the final A1.

Total: 5 marks

Q29 — Standard deviation from sums and mean equals sd condition [HARD]

Computing standard deviation:

$\bar{c} = \dfrac{220}{40} = 5.5$ **(A1)**

$\sigma^2 = \dfrac{\sum c^2}{n} - \bar{c}^2 = \dfrac{1630}{40} - (5.5)^2 = 40.75 - 30.25 = 10.5$ **M1**

$\sigma = \sqrt{10.5} \approx 3.24$ (3 s.f.) **A1**

Since $3.24 > 3$, Priya's statement is **incorrect**. **R1**

Note: R1 requires an explicit comparison of $\sqrt{10.5}$ with 3 and a clear verdict.

Finding d:

Adding constant d: new mean $= \bar{c} + d = 5.5 + d$; standard deviation is unchanged $= \sqrt{10.5}$. **M1**

Condition: $5.5 + d = \sqrt{10.5} \Rightarrow d = \sqrt{10.5} - 5.5 \approx -2.26$ (3 s.f.) **A1**

Note: Award A1 for exact answer $d = \sqrt{10.5} - 5.5$ or equivalent. Accept $d \approx -2.26$.

Total: 5 marks

Q30 — Combined mean and standard deviation, estimate comparison [HARD]

Combined mean:

$\bar{x}_{\text{combined}} = \dfrac{30(58) + 30(64)}{60} = \dfrac{1740 + 1920}{60} = \dfrac{3660}{60} = 61$ **A1**

Combined standard deviation:

For class A: $\sum x_A^2 = 30(\sigma_A^2 + \bar{x}_A^2) = 30(144 + 3364) = 30(3508) = 105240$ **M1**

For class B: $\sum x_B^2 = 30(81 + 4096) = 30(4177) = 125310$ **(A1)**

$\sigma^2_{\text{combined}} = \dfrac{105240 + 125310}{60} - 61^2 = \dfrac{230550}{60} - 3721 = 3842.5 - 3721 = 121.5$ **M1**

$\sigma_{\text{combined}} = \sqrt{121.5} = \dfrac{3\sqrt{27}}{2} = \dfrac{9\sqrt{6}}{2} \approx 11.02$ **(A1)**

Note: Accept $\sqrt{121.5}$ or 11.0 (3 s.f.) as exact/approximate final answer.

Since $\sqrt{121.5} \approx 11.02 > 10.5$, the teacher's estimate of 10.5 is an **underestimate**. **A1**

Note: The final A1 requires both the correct value and the correct verdict (underestimate). FT on their σ_{combined} provided the comparison is explicit.

Total: 5 marks