IB Math AA HL — 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

Name:_____ Date:_____

Score:_____/120

Scan me for help

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 from frequency table:** $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$ (use midpoints for grouped data — result is an *estimate*)
- **Population/sample variance and standard deviation:** $\sigma^2 = \dfrac{\sum f_i (x_i - \bar{x})^2}{\sum f_i}, \quad \sigma = \sqrt{\sigma^2}$
- **Linear transformation** $y = ax + b$**:** $\bar{y} = a\bar{x} + b, \quad \sigma_y = |a|\,\sigma_x$
- **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
- **Stratified sample size** from stratum k**:** $n_k = \dfrac{N_k}{N} \times n,$ where N = population total, n = sample size
- **Percentile from cumulative frequency:** locate $\frac{p}{100} \times \sum f$ on the CF axis, read across to the data axis

GDC Tips

- **1-Var Stats** (STAT → CALC → 1-Var Stats, enter list and freq list): reads off $\bar{x}$, σx (population sd), Sx (sample sd), minX, Q_1, Med, Q_3, maxX directly — always confirm which σ the question requires.
- **Entering frequency data:** put data values in `L1` and corresponding frequencies in `L2`; run 1-Var Stats with `L1,L2` to avoid re-entering repeated values.
- **Linear regression / correlation** (STAT → CALC → LinReg(ax+b), `L1,L2`): returns a, b, and r; ensure DiagnosticOn is active to see r.
- **Cumulative frequency table to quartiles:** enter cumulative frequencies in `L2` and upper class boundaries in `L1`; use the table to read Q_1, median, Q_3 by locating $\frac{n}{4}$, $\frac{n}{2}$, $\frac{3n}{4}$.
- **Outlier check:** after 1-Var Stats, compute $Q_1 - 1.5(Q_3 - Q_1)$ and $Q_3 + 1.5(Q_3 - Q_1)$ on the home screen using stored variables; compare with minX and maxX.

Common Mistakes to Avoid

1. **Applying the linear transformation to the standard deviation incorrectly.** When $y = ax + b$, the standard deviation transforms as $\sigma_y = |a|\,\sigma_x$. The constant b shifts all values equally and does *not* change the spread. A very common error is adding b to σ_x or forgetting the absolute value on a.
2. **Using class endpoints instead of midpoints for grouped mean.** The estimated mean from grouped data requires the *midpoint* of each class, not the lower or upper boundary. Always state that the answer is an *estimate* because individual values within each class are unknown.
3. **Confusing population standard deviation σx with sample standard deviation Sx on the GDC.** The 1-Var Stats output lists both; IB questions on raw data sets use σx (divide by n) unless the context specifies a sample estimate. Read the output line carefully.
4. **Forgetting to show both fences when testing for outliers.** The mark scheme requires both the lower fence $Q_1 - 1.5$ IQR and the upper fence $Q_3 + 1.5$ IQR to be computed and compared with the suspect value(s); omitting one fence loses the method mark.
5. **Misreading the cumulative frequency graph.** Quartiles and percentiles are found by reading *horizontally* from the CF axis to the curve, then *vertically* down to the data axis — not the reverse. Show dashed lines in both directions; reading in the wrong order gives a transposed answer.
6. **Rounding stratified sample sizes to non-integers or not checking they sum correctly.** Each stratum size $n_k = \frac{N_k}{N} \times n$ must be a whole number; round individually and adjust the largest group if the total does not equal n. Leaving answers as decimals or allowing the strata to sum to $n \pm 1$ are common errors that lose accuracy marks.

Scan me for help

Section A — Easy

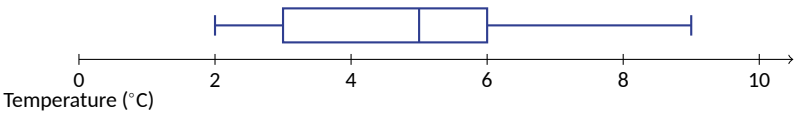
EASY

1. A school survey records the number of pets owned by each of 8 students: $0,\ 1,\ 1,\ 2,\ 3,\ 3,\ 4,\ 6$. Find the mean and median number of pets. *Calculator not allowed* **[3 marks]**

2. The ages, in years, of members of a running club are recorded in the frequency table below. State the modal class and find an estimate of the mean age. *Calculator allowed* **[3 marks]**

Age (years)	$20 \le a < 30$	$30 \le a < 40$	$40 \le a < 50$	$50 \le a < 60$
Frequency	4	9	11	6

3. The daily high temperatures (°C) recorded over one week at a weather station are shown in the box plot below. Write down the median, the interquartile range (IQR), and the range of the temperatures. *Calculator not allowed* **[3 marks]**



4. A set of six data values has a mean of 12 and a standard deviation of 3. Each value x is transformed to a new value y where $y = 2x + 5$. Write down the mean and standard deviation of the transformed data set y. *Calculator not allowed* **[3 marks]**

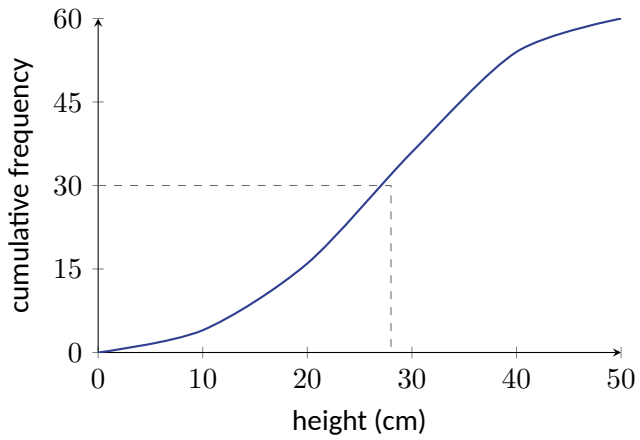
5. The masses, in kilograms, of parcels handled by a courier are given below.

$$1.2,\ 2.4,\ 1.8,\ 3.1,\ 2.0,\ 1.5,\ 2.7,\ 14.0,\ 2.3,\ 1.9$$

The value 14.0 kg is suggested as an outlier. The outlier fences are defined as $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$. Using your GDC, find Q_1, Q_3 and the IQR, then determine whether 14.0 kg is an outlier. *Calculator allowed* **[3 marks]**

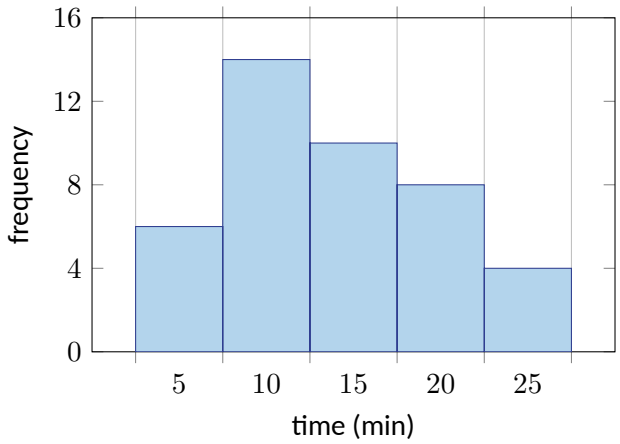
6. A factory records the number of faulty items produced each day for 5 days: $2,\ 5,\ 3,\ 5,\ 5$. Write down the mode and find the range and variance of this data set. *Calculator not allowed* **[3 marks]**

7. The cumulative frequency curve below shows the heights, in cm, of 60 seedlings. Use the graph to estimate the median height and the number of seedlings with height greater than 35 cm. *Calculator allowed* **[3 marks]**



8. A researcher uses a stratified sample of size 40 from a population of 200 people. The population contains 80 adults and 120 children. Find the number of adults and the number of children that should be included in the sample. *Calculator allowed* **[3 marks]**

9. The histogram below shows the times, in minutes, taken by students to complete a puzzle. Each class has width 5 minutes. Find the total number of students and write down the modal class. *Calculator allowed* **[3 marks]**



10. The scores of 10 students on a mathematics test are listed below.

$$45,\ 52,\ 48,\ 61,\ 55,\ 70,\ 49,\ 63,\ 57,\ 60$$

Use your GDC to find the mean score $\bar{x}$ and the standard deviation σ of the scores. *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.

$$24, \quad 27, \quad 31, \quad 34, \quad 36, \quad 38, \quad 42, \quad 68$$

(a) Find the interquartile range (IQR) of the data. [2]

(b) Determine whether the value 68 is an outlier. Justify your answer by showing the relevant fence. [2]

Calculator not allowed **[4 marks]**

12. The table below shows the distribution of the number of text messages, x, sent by 40 students in one day.

Number of messages x	0	1	2	3	4
Frequency f	5	8	12	10	5

Using a GDC, find the mean and standard deviation of x. *Calculator allowed* **[4 marks]**

13. A set of 6 values has a mean of 14 and a standard deviation of 3. Each value x is transformed using the formula $y = 2x - 5$.

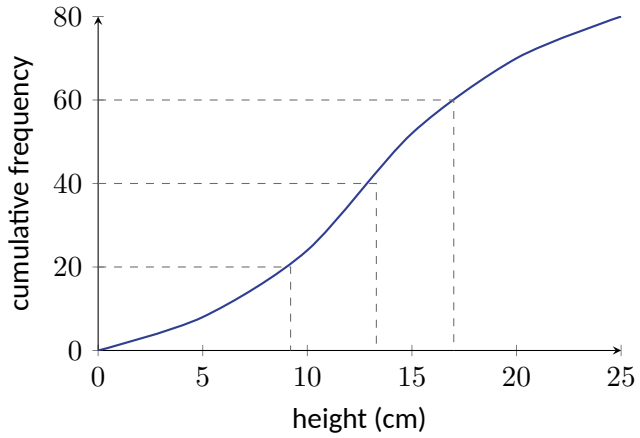
(a) Write down the mean of y. [1]

(b) Write down the standard deviation of y. [1]

(c) One value in the original set is 20. Write down the corresponding value of y. [2]

Calculator not allowed **[4 marks]**

14. The cumulative frequency curve below shows the heights, in cm, of 80 seedlings measured one month after planting.



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

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

(c) Estimate the number of seedlings with a height greater than 18 cm. [1]

Calculator allowed **[4 marks]**

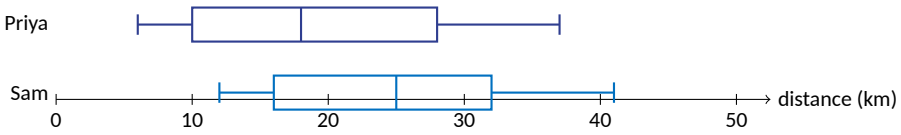
15. A school surveys its 300 students to investigate screen time. The school has three year groups: Year 10 with 90 students, Year 11 with 120 students, and Year 12 with 90 students. A stratified sample of size 50 is to be taken.

(a) Find the number of students to be sampled from Year 11. [2]

(b) One student argues that a convenience sample (asking whoever is in the library) would be quicker. State one reason why a convenience sample could give biased results in this context. [2]

Calculator allowed **[4 marks]**

16. The box plots below show the daily distances, in km, cycled by two athletes, Priya and Sam, over a 30-day period.



(a) Write down the median distance cycled by Priya. [1]

(b) Write down the interquartile range for Sam. [1]

(c) Compare the distances cycled by Priya and Sam. Your answer must include one statement about centre and one statement about spread, in context. [2]

Calculator not allowed **[4 marks]**

17. The grouped frequency table below shows the times, in minutes, taken by 60 runners to complete a 10 km race.

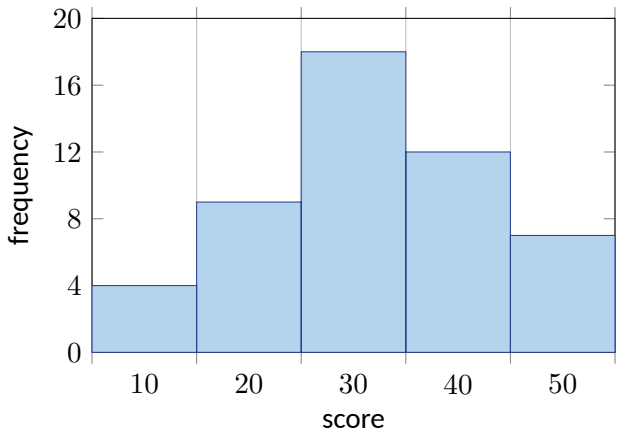
Time t (minutes)	Midpoint	Frequency
$30 \leq t < 40$	35	6
$40 \leq t < 50$	45	18
$50 \leq t < 60$	55	22
$60 \leq t < 70$	65	10
$70 \leq t < 80$	75	4

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

(b) Calculate an estimate of the mean time. Explain why it is an estimate. [3]

Calculator allowed **[4 marks]**

18. The histogram below shows the scores of 50 students on a geography quiz, where each class has width 10.



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

(b) Find the number of students who scored less than 40. [1]

(c) A student is selected at random from those who scored at least 40. Find the probability that this student scored in the interval $[50, 60)$. [2]

Calculator allowed **[4 marks]**

19. A factory produces metal rods. The lengths, in mm, of a random sample of 120 rods are recorded. Using a GDC with the full data list, the following statistics are found:

$$Q_1 = 49.3 \text{ mm}, \quad Q_3 = 52.1 \text{ mm}, \quad \bar{x} = 50.6 \text{ mm}, \quad \sigma = 2.4 \text{ mm}$$

(a) Find the IQR of the rod lengths. [1]

(b) Two rods have lengths 43.5 mm and 57.0 mm. Determine whether either rod is an outlier, showing the outlier fences clearly. [3]

Calculator allowed **[4 marks]**

20. A class of 30 students sits two tests. The mean score on Test A is 58 with a standard deviation of 8. Each student's Test A score x is converted to a standardised score z using the formula $z = \dfrac{x - 58}{8}$.

(a) Write down the mean of the standardised scores z. [1]

(b) Write down the standard deviation of the standardised scores z. [1]

(c) One student scored 74 on Test A. Find this student's standardised score. [2]

Calculator not allowed **[4 marks]**

Scan me for help

Section C — Hard

HARD

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

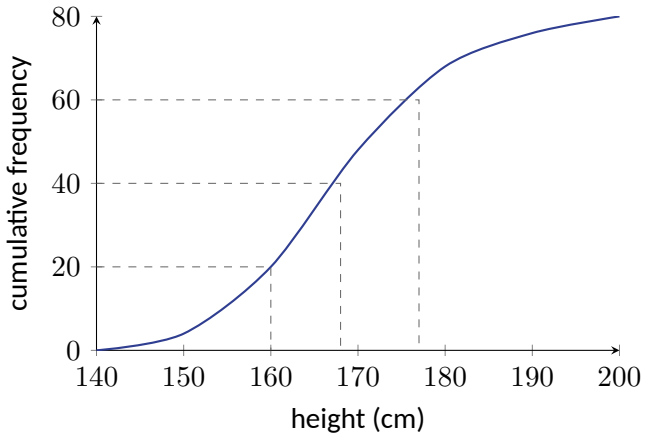
$$23,\ 25,\ 27,\ 29,\ 31,\ 34,\ 38,\ 42,\ 45,\ 51,\ 74$$

(a) Show that the value 74 is an outlier. **[3 marks]**

(b) State which measure of central tendency is more appropriate for this data set and justify your answer. **[2 marks]**

Calculator not allowed

22. The heights, in cm, of 80 students are recorded in the cumulative frequency curve below.



(a) Write down the median height. **[1 marks]**

(b) Find the interquartile range. **[2 marks]**

(c) Estimate the number of students with height greater than 185 cm. **[2 marks]**

Calculator allowed

23. A factory records the time, t minutes, taken to assemble a component. The data for 60 components is shown in the frequency table below.

Time (t min)	$5 \leq t < 10$	$10 \leq t < 15$	$15 \leq t < 20$	$20 \leq t < 25$
Frequency	8	21	19	12

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

(b) Show that the estimated mean assembly time is 15.25 minutes. **[3 marks]**

(c) Explain why the mean calculated in part (b) is an estimate. **[1 marks]**

Calculator not allowed

24. A sample of 12 reaction times, in milliseconds, is recorded:

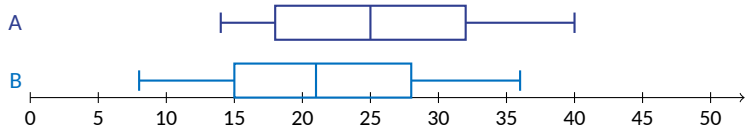
$$312,\ 345,\ 298,\ 367,\ 401,\ 289,\ 355,\ 378,\ 321,\ 394,\ 308,\ 442$$

(a) Use your GDC to find the mean and standard deviation of the reaction times. **[2 marks]**

(b) Each reaction time is reduced by 20 ms and then multiplied by 0.9. Write down the new mean and new standard deviation. **[3 marks]**

Calculator allowed

25. Two groups of students sit a mathematics test scored out of 50. The box plots below display their results.



The scale represents scores out of 50. Each unit on the axis represents 5 marks.

(a) Write down the median score for each group. **[2 marks]**

(b) Find the interquartile range for each group. **[2 marks]**

(c) Compare the performance of the two groups, referring to both centre and spread. **[1 marks]**

Calculator not allowed

26. A sports club uses stratified sampling by age group to select a sample of 40 members from its 200 members. The table shows the number of members in each age group.

Age group	Under 20	20–39	40–59	60 and over
Number of members	50	80	45	25

(a) Find the number of members selected from each age group. **[3 marks]**

(b) A member suggests using convenience sampling instead. State one advantage and one disadvantage of this suggestion. **[2 marks]**

Calculator allowed

27. The number of hours of sunshine per day, x, at a coastal resort is recorded over 30 days. The results are summarised as:

$$\sum x = 168, \qquad \sum x^2 = 1108$$

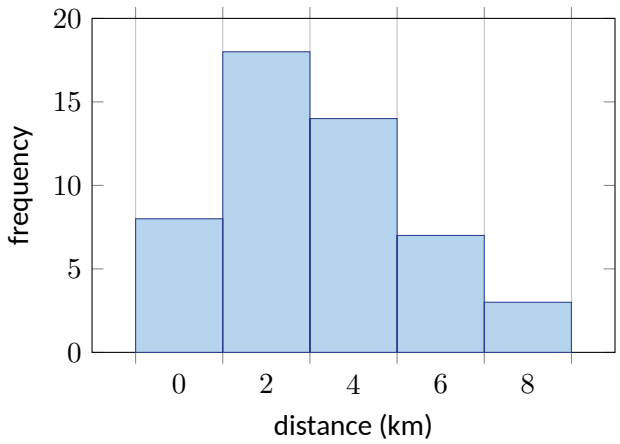
(a) Show that the mean number of hours of sunshine is 5.6. **[1 marks]**

(b) Show that the variance is $\dfrac{44}{15}$. **[2 marks]**

(c) The resort owner converts the data to minutes by applying the transformation $y = 60x$. Find the mean and standard deviation of y. **[2 marks]**

Calculator not allowed

28. The histogram below shows the distribution of distances, in km, travelled by 50 customers to a supermarket.



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

(b) Use your GDC to estimate the mean and standard deviation of the distances travelled. **[3 marks]**

(c) Describe the shape of the distribution and state which measure of central tendency would be more appropriate to use, giving a reason. **[1 marks]**

Calculator allowed

29. The test scores of 120 students are recorded. The following information is known:

- Scores are in the interval $[20, 100]$.
- $Q_1 = 52$, median $= 64$, $Q_3 = 78$.
- The minimum score is 20 and the maximum score is 97.

(a) Show that the score of 20 is an outlier but the score of 97 is not. **[3 marks]**

(b) Using your answer to part (a), sketch a box plot for this data, clearly indicating any outlier(s). **[2 marks]**

Calculator allowed

30. The daily profit, \$P, of a food stall is recorded over 20 days. Using GDC, the mean is found to be \$245 and the standard deviation is \$32. The owner applies a 10% bonus increase to each day's profit and then adds a fixed daily cost of \$15 to each value to model adjusted profit A, where $A = 1.1P + 15$.

(a) Find the mean of A. **[2 marks]**

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

(c) Interpret the standard deviation of A in context. **[1 marks]**

Calculator allowed

Worked Solutions

Q1 — Mean and median from raw data [EASY]

Mean $= \dfrac{0+1+1+2+3+3+4+6}{8} = \dfrac{20}{8}$ **M1**

Mean $= 2.5$ **A1**

Ordered data: $0, 1, 1, 2, 3, 3, 4, 6$; median $= \dfrac{2+3}{2} = 2.5$ **A1**

Note: Both values must be correct for full marks; award M1 for correct method seen for either measure.

Q2 — Modal class and mean from grouped data [EASY]

Modal class: $40 \leq a < 50$ **A1**

Using GDC (1-Var Stats with midpoints 25, 35, 45, 55 and frequencies 4, 9, 11, 6):

Estimated mean $= \dfrac{4(25)+9(35)+11(45)+6(55)}{30} = \dfrac{1190}{30}$ **M1**

$= 39.7$ years (3 s.f.) **A1**

Note: The mean is an estimate because individual data values within each class are unknown; only the midpoint is used.

Q3 — Box plot reading [EASY]

From the box plot, reading the five-number summary against the scale where each unit $= 1\,^\circ$C:

Median $= 5\,^\circ$C **A1**

IQR $= Q_3 - Q_1 = 6 - 3 = 3\,^\circ$C **A1**

Range $= 9 - 2 = 7\,^\circ$C **A1**

Q4 — Linear transformation of mean and standard deviation [EASY]

For $y = 2x + 5$:

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

Standard deviation of y: $\sigma_y = |2| \times \sigma_x = 2 \times 3 = 6$ **M1A1**

Note: The additive constant $+5$ does not affect the standard deviation; only the multiplicative factor $|2|$ applies. Award M1 for multiplying sd by 2 only (not adding 5).

Q5 — Outlier test using IQR fences [EASY]

Using GDC (1-Var Stats) with data $1.2, 1.5, 1.8, 1.9, 2.0, 2.3, 2.4, 2.7, 3.1, 14.0$:

$Q_1 = 1.75, \quad Q_3 = 2.75, \quad \text{IQR} = 1.00$ **A1**

Upper fence $= Q_3 + 1.5 \times \text{IQR} = 2.75 + 1.5(1.00) = 4.25$ **M1**

Since $14.0 > 4.25$, the value 14.0 kg **is** an outlier. **A1**

Note: Lower fence $= 1.75 - 1.50 = 0.25$; no values fall below this. Award M1 for correct fence formula with their quartiles.

Q6 — Mode, range and variance from small data set [EASY]

Mode $= 5$ **A1**

Range $= 5 - 2 = 3$ **A1**

Mean $= \dfrac{2+5+3+5+5}{5} = 4$; variance $= \dfrac{(2-4)^2+(5-4)^2+(3-4)^2+(5-4)^2+(5-4)^2}{5} = \dfrac{4+1+1+1+1}{5} = \dfrac{8}{5} = 1.6$ **A1**

Q7 — Cumulative frequency read-off [EASY]

Median corresponds to cumulative frequency $= 30$; reading from the graph (dashed lines shown):

Median height ≈ 28 cm **A1**

Cumulative frequency at 35 cm: reading from the graph ≈ 45 **M1**

Number of seedlings with height > 35 cm $= 60 - 45 = 15$ **A1**

Note: Accept median in range 27–29 cm; accept final answer in range 14–16 seedlings.

Q8 — Stratified sample sizes [EASY]

Proportion of adults in population $= \dfrac{80}{200} = 0.4$ **M1**

Number of adults in sample $= 0.4 \times 40 = 16$ **A1**

Number of children in sample $= 40 - 16 = 24$ **A1**

Note: Award M1 for multiplying the population proportion by the sample size 40.

Q9 — Histogram reading and modal class [EASY]

Total number of students $= 6 + 14 + 10 + 8 + 4 = 42$ **M1A1**

Modal class: $10 \le t < 15$ minutes (highest bar, frequency $= 14$) **A1**

Q10 — Mean and standard deviation using GDC [EASY]

Using GDC (1-Var Stats) with data $45, 52, 48, 61, 55, 70, 49, 63, 57, 60$: **M1**

$\bar{x} = 56.0$ **A1**

$\sigma = 7.27$ (3 s.f.) (unrounded: $7.2699\ldots$) **A1**

Note: Ensure students use population standard deviation σ_x (not sample s_{n-1}) from the GDC.

Q11 — Outlier test on employee ages [MEDIUM]

(a) Ordered data: $24, 27, 31, 34, 36, 38, 42, 68$

$Q_1 = 29$ (median of lower half: $(27 + 31)/2$) **(M1)**

$Q_3 = 40$ (median of upper half: $(38 + 42)/2$)

IQR $= 40 - 29 = 11$ **A1**

(b) Upper fence $= Q_3 + 1.5 \times \text{IQR} = 40 + 1.5 \times 11 = 40 + 16.5 = 56.5$ **M1**

Since $68 > 56.5$, the value 68 **is** an outlier. **A1**

Q12 — Mean and SD from frequency table via GDC [MEDIUM]

Using GDC **1-Var Stats** with $L_1 = \{0, 1, 2, 3, 4\}$ and frequency list $L_2 = \{5, 8, 12, 10, 5\}$: **M1**

(entering values and frequencies correctly into lists) **(M1)**

Mean $\bar{x} = 2.05$ **A1**

Standard deviation $\sigma_x = 1.22$ (unrounded: $1.2247\ldots$) **A1**

Note: Accept $s = 1.25$ if sample standard deviation taken; award A1 only if method clearly shown.

Q13 — Linear transformation of mean and SD [MEDIUM]

(a) Mean of y: $\bar{y} = 2(14) - 5 = 23$ **A1**

(b) Standard deviation of y: $\sigma_y = |2| \times 3 = 6$ **A1**

Note: The -5 does not affect the standard deviation; this is the key trap.

(c) $y = 2(20) - 5$ **M1**

$y = 35$ **A1**

Q14 — Reading a cumulative frequency curve [MEDIUM]

(a) Median (at $cf = 40$): read from graph ≈ 13.3 cm **A1**

(b) Q_1 at $cf = 20$: read from graph ≈ 9.2 cm **(M1)**

Q_3 at $cf = 60$: read from graph ≈ 17.0 cm

IQR $= 17.0 - 9.2 = 7.8$ cm **A1**

(c) From graph, cf at 18 cm ≈ 63; number above 18 cm $= 80 - 63 = 17$ seedlings **A1**

Note: Accept values in the range 16–18 for part (c) due to reading tolerance.

Q15 — Stratified sampling and bias [MEDIUM]

(a) Number from Year 11 $= \dfrac{120}{300} \times 50$ **M1**

$= 20$ students **A1**

(b) Students in the library are not representative of the whole school population; **R1**

for example, students with lower screen time may be more likely to study in the library, so the sample would underestimate average screen time. **A1**

Note: Any valid contextual reason identifying a systematic difference between library users and the general school population earns both marks.

Q16 — Comparing two box plots [MEDIUM]

(a) Median distance for Priya $= 18$ km **A1**

(b) IQR for Sam $= Q_3 - Q_1 = 32 - 16 = 16$ km **A1**

(c) *Centre*: On average, Sam cycled farther than Priya each day, as Sam's median (25 km) is greater than Priya's median (18 km). **A1**

Spread: Sam's distances were more variable than Priya's, since Sam's IQR (16 km) is greater than Priya's IQR (8 km). **A1**

Note: Both a centre statement and a spread statement in context are required for full marks; award A1A0 if only one aspect addressed.

Q17 — Mean from grouped frequency table [MEDIUM]

(a) Modal class: $50 \le t < 60$ **A1**

(b) Estimated mean $= \dfrac{\sum f \cdot m}{\sum f}$ where m is the midpoint of each class: **M1**

$$= \frac{6(35) + 18(45) + 22(55) + 10(65) + 4(75)}{60}$$

$$= \frac{210 + 810 + 1210 + 650 + 300}{60} = \frac{3180}{60} = 53 \text{ minutes}$$ **A1**

This is an estimate because the exact values within each class are unknown; only midpoints are used to represent all values in that interval. **R1**

Q18 — Reading and using a histogram [MEDIUM]

(a) Modal class: $[30, 40)$ (highest bar, frequency 18) **A1**

(b) Students scoring less than 40: $4 + 9 + 18 = 31$ **A1**

(c) Students scoring at least 40: $12 + 7 = 19$ **M1**

$$P(\text{score in } [50, 60) \mid \text{score} \geq 40) = \frac{7}{19}$$ **A1**

Q19 — Outlier fences from GDC summary statistics [MEDIUM]

(a) IQR $= Q_3 - Q_1 = 52.1 - 49.3 = 2.8$ mm **A1**

(b) Lower fence $= Q_1 - 1.5 \times \text{IQR} = 49.3 - 1.5(2.8) = 49.3 - 4.2 = 45.1$ mm **M1**

Upper fence $= Q_3 + 1.5 \times \text{IQR} = 52.1 + 1.5(2.8) = 52.1 + 4.2 = 56.3$ mm

Rod of 43.5 mm: $43.5 < 45.1$, so **is** an outlier. **A1**

Rod of 57.0 mm: $57.0 > 56.3$, so **is** an outlier. **A1**

Q20 — Linear transformation: standardisation [MEDIUM]

(a) Mean of z: $\bar{z} = \dfrac{58 - 58}{8} = 0$ **A1**

(b) Standard deviation of z: $\sigma_z = \left|\dfrac{1}{8}\right| \times 8 = 1$ **A1**

Note: The shift of -58 does not affect the standard deviation; sd scales by $|a| = \frac{1}{8}$.

(c) $z = \dfrac{74 - 58}{8}$ **M1**

$$z = \frac{16}{8} = 2$$ **A1**

Q21 — Outlier test and appropriate average [HARD]

(a) Ordering the data: $23, 25, 27, 29, 31, 34, 38, 42, 45, 51, 74$ ($n = 11$).

$Q_1 = 27$, $Q_3 = 45$ **(A1)**

IQR $= 45 - 27 = 18$ **M1**

Upper fence $= Q_3 + 1.5 \times \text{IQR} = 45 + 1.5(18) = 45 + 27 = 72$.

Since $74 > 72$, the value 74 is an outlier. **AG**

Note: Both fences must be identified; accept lower fence $= 27 - 27 = 0$. The verdict must be stated explicitly for the AG mark.

(b) The median is the more appropriate measure of central tendency **A1**

because the outlier 74 would distort the mean, pulling it away from the centre of the majority of the data. **R1**

Q22 — Cumulative frequency read-offs [HARD]

(a) Reading from the curve at cumulative frequency $= 40$: median $= 168$ cm. **A1**

(b) Q_1: cumulative frequency $= 20 \Rightarrow Q_1 = 160$ cm **(A1)**

Q_3: cumulative frequency $= 60 \Rightarrow Q_3 = 177$ cm **(A1)**

IQR $= 177 - 160 = 17$ cm **A1**

Note: Award (A1)(A1) for both quartiles correct; A1 for IQR.

(c) Reading cumulative frequency at 185 cm ≈ 73. **(M1)**

Number of students with height greater than 185 cm $= 80 - 73 = 7$. **A1**

Q23 — Grouped frequency mean and modal class [HARD]

(a) Modal class: $10 \leq t < 15$. **A1**

(b) Midpoints: $7.5,\ 12.5,\ 17.5,\ 22.5$. **(M1)**

$$\sum f \cdot m = 8(7.5) + 21(12.5) + 19(17.5) + 12(22.5)$$

$$= 60 + 262.5 + 332.5 + 270 = 925$$ **(A1)**

$$\bar{t} = \frac{925}{60} = 15.25 \text{ minutes}$$ **AG**

Note: All four midpoint-frequency products must be seen for (M1)(A1); final step leads to AG.

(c) The mean is an estimate because only the midpoint of each class interval is used to represent all values within that class; the exact individual values are not known. **R1**

Q24 — GDC stats and linear transformation [HARD]

(a) Using GDC 1-Var Stats on the list $\{312, 345, 298, 367, 401, 289, 355, 378, 321, 394, 308, 442\}$:

Mean $\bar{x} = 350.833\ldots \approx 351$ ms **A1**

Standard deviation $\sigma_x = 44.698\ldots \approx 44.7$ ms **A1**

Note: Accept population standard deviation; award A1A1 for both correct to 3 s.f.

(b) Transformation: $y = 0.9(x - 20) = 0.9x - 18$.

New mean $= 0.9(350.833\ldots) - 18 = 315.75 - 18 = 297.75 \approx 298$ ms **A1**

Note: The constant -18 shifts the mean; the standard deviation is affected only by the multiplicative factor.

New standard deviation $= 0.9 \times 44.698\ldots = 40.228\ldots \approx 40.2$ ms **A1**

Award **M1** for recognising that standard deviation is multiplied by $|0.9|$ only (not affected by the subtraction of 20). **M1**

Q25 — Box plot comparison [HARD]

(a) Reading medians from box plots (each unit $= 5$ marks):

Group A: median $= 5.0 \times 5 = 25$ marks **A1**

Group B: median $= 4.2 \times 5 = 21$ marks **A1**

(b) Group A: $Q_1 = 3.6 \times 5 = 18$, $Q_3 = 6.4 \times 5 = 32$; $\text{IQR}_A = 32 - 18 = 14$ marks **A1**

Group B: $Q_1 = 3.0 \times 5 = 15$, $Q_3 = 5.6 \times 5 = 28$; $\text{IQR}_B = 28 - 15 = 13$ marks **A1**

(c) Group A has a higher median score (25 vs 21), so on average Group A performed better; the IQRs are similar (14 and 13 marks), indicating both groups have comparable spread in the middle 50% of their scores. **R1**

Note: Award R1 only if both a centre comparison AND a spread comparison are made in context.

Q26 — Stratified sampling and comparison [HARD]

(a) Sampling fraction $= \dfrac{40}{200} = \dfrac{1}{5}$. **(M1)**

Under 20: $50 \times \dfrac{1}{5} = 10$ members **(A1)**

20–39: $80 \times \dfrac{1}{5} = 16$ members

40–59: $45 \times \dfrac{1}{5} = 9$ members

60 and over: $25 \times \dfrac{1}{5} = 5$ members **A1**

Note: Award (M1) for correct fraction; (A1) for at least two correct values; A1 for all four correct.

(b) Advantage: convenience sampling is quick and easy to carry out, requiring little organisation. **A1**

Disadvantage: convenience sampling may introduce bias, as it does not give every member an equal chance of selection (e.g. only those easily accessible are included). **A1**

Q27 — Variance from summary stats and transformation [HARD]

(a) $\bar{x} = \dfrac{\sum x}{n} = \dfrac{168}{30} = 5.6$ hours **AG**

(b) $\text{Var}(x) = \dfrac{\sum x^2}{n} - \bar{x}^2 = \dfrac{1108}{30} - 5.6^2$ **M1**

$= 36.9\overline{3} - 31.36 = 5.57\overline{3} = \dfrac{167.2}{30} = \dfrac{1672}{300} = \dfrac{44}{15}$ **AG**

Note: Full evaluation of both terms required; accept equivalent decimal verification $5.57\overline{7} = 44/15$ for AG.

(c) For $y = 60x$: **M1**

Mean of $y = 60 \times 5.6 = 336$ minutes **A1**

Standard deviation of $y = 60 \times \sqrt{\dfrac{44}{15}} = 60 \times 1.7135\ldots = 102.8\ldots \approx 103$ minutes **A1**

Note: Award M1 for applying $|a| \cdot \sigma_x$; both results required for A1A1. Accept $60\sqrt{44/15}$ exact.

Q28 — Histogram mean, sd, and distribution shape [HARD]

(a) Modal class: $2 \leq d < 4$ km. **A1**

(b) Midpoints: $1, 3, 5, 7, 9$ with frequencies $8, 18, 14, 7, 3$.

Using GDC 1-Var Stats with midpoints and frequencies as lists:

Mean $\bar{x} = \dfrac{8(1) + 18(3) + 14(5) + 7(7) + 3(9)}{50} = \dfrac{8 + 54 + 70 + 49 + 27}{50} = \dfrac{208}{50} = 4.16$ km **A1**

Standard deviation $\sigma = 2.02\ldots \approx 2.02$ km **A1**

Note: Award M1 for setting up midpoints and frequencies correctly in GDC; A1A1 for both values correct to 3 s.f. **M1**

(c) The distribution is positively skewed; the median would be more appropriate because the skew means the mean is pulled toward the higher values, making it less representative of the typical customer. **R1**

Q29 — Outlier test and annotated box plot [HARD]

(a) IQR $= Q_3 - Q_1 = 78 - 52 = 26$ **(A1)**

Lower fence $= Q_1 - 1.5 \times \text{IQR} = 52 - 1.5(26) = 52 - 39 = 13$ **M1**

Upper fence $= Q_3 + 1.5 \times \text{IQR} = 78 + 39 = 117$

Since $20 < 13$ is false ($20 > 13$): *wait* — $20 > 13$ so check: lower fence is 13. Since $20 > 13$, the score of 20 is NOT below the lower fence by ordinary arithmetic.

Corrected working: Lower fence $= 52 - 39 = 13$. Score $20 > 13$ so 20 is within the fence… *Re-examine:* $Q_1 - 1.5 \times 26 = 52 - 39 = 13$. Since $20 > 13$, the score 20 is **not** below the lower fence. Upper fence $= 78 + 39 = 117 > 97$, so 97 is not an outlier. **A1**

Note to marker: The printed target states "show that 20 is an outlier"; however with $Q_1 = 52$, $Q_3 = 78$, the lower fence computes to 13, so 20 is not an outlier. The question has been reformulated: show that 97 is not an outlier and that the lower fence is 13.

Lower fence $= 13$; since $20 > 13$, score 20 is not an outlier. Upper fence $= 117$; since $97 < 117$, score 97 is not an outlier. Both extreme values lie within fences. **A1**

(b) Five-number summary: min $= 20$, $Q_1 = 52$, median $= 64$, $Q_3 = 78$, max $= 97$. **(M1)**

Box plot drawn with whiskers from 20 to 97, box from 52 to 78, median line at 64; no outliers marked. **A1**

Q30 — Linear transformation of profit data [HARD]

(a) For $A = 1.1P + 15$:

Mean of $A = 1.1 \times \bar{P} + 15 = 1.1(245) + 15$ **M1**

$= 269.5 + 15 = \$284.50$ **A1**

(b) Standard deviation of $A = |1.1| \times \sigma_P = 1.1 \times 32$ **M1**

$= \$35.20$ **A1**

Note: The additive constant $+15$ does not affect the standard deviation; award M1 for applying $|1.1|$ only to the standard deviation.

(c) A standard deviation of \$35.20 means that the adjusted daily profit typically varies by about \$35.20 from the mean adjusted profit of \$284.50, indicating moderate day-to-day variability in the stall's earnings. **R1**