IB Math AI HL — Topic 4
Statistics Toolkit

Sampling · Central Tendency · Dispersion · Frequency Tables · Linear Transformations · Outliers · Box & Whisker · Cumulative Frequency · Histograms · Interpreting Data

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 sf unless stated. GDC permitted throughout.

Key Formulae

Mean (grouped): $\bar{x} = \dfrac{\sum fx}{\sum f}$ (use midpoints x)

Population SD: $\sigma_n = \sqrt{\dfrac{\sum fx^2}{\sum f} - \bar{x}^2}$

Unbiased (sample) SD: $s_{n-1} = \sqrt{\dfrac{n}{n-1}}\,\sigma_n$ or $s^2_{n-1} = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

Outlier fences: Upper: $Q_3 + 1.5 \times \text{IQR}$; Lower: $Q_1 - 1.5 \times \text{IQR}$

IQR: $\text{IQR} = Q_3 - Q_1$

Linear transform: If $y = ax + b$ then $\bar{y} = a\bar{x} + b$ and $\sigma_y = |a|\,\sigma_x$

Freq density (histogram): $\text{fd} = \dfrac{\text{frequency}}{\text{class width}}$

95% CI for μ: $\bar{x} \pm 1.96\,\dfrac{s_{n-1}}{\sqrt{n}}$

GDC Tips

1. **Mean & SD from a list:** STAT → Edit → enter data in L1 → STAT → CALC → 1-Var Stats. Read $\bar{x}$, σx (population SD), Sx (sample SD), n.
2. **Mean & SD from a frequency table:** Enter midpoints in L1, frequencies in L2; 1-Var Stats L1, L2.
3. **Quartiles:** Scroll down in 1-Var Stats output to see MinX, Q1, Med, Q3, MaxX.
4. **Cumulative frequency:** Enter upper class boundaries in L1, cumulative frequencies in L2; use Stat Plot

→ Scatter to check interpolation.

5. **Histogram from frequency density:** Remember: frequency = fd × class width before entering data.

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

Scan me for help

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

The table gives the lengths (in cm) of 40 fish caught in a lake.

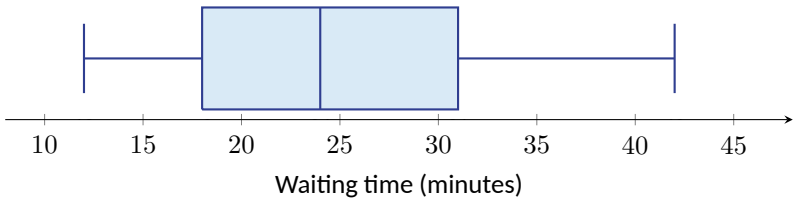
Length, l (cm)	Frequency
$10 \leq l < 20$	6
$20 \leq l < 30$	14
$30 \leq l < 40$	12
$40 \leq l < 50$	8

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

(b) Calculate an estimate of the mean length of the fish. [2]

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

The box and whisker diagram below shows the waiting times (in minutes) of customers at a bank.



(a) Write down the median waiting time. [1]

(b) Find the interquartile range. [2]

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

A school has 480 students across three year groups: Year 10 has 160 students, Year 11 has 180 students, and Year 12 has 140 students. A researcher selects a stratified random sample of 60 students.

(a) Write down the number of Year 10 students that should be selected. [1]

(b) Find the number of students from Year 11 and Year 12 that should be selected. [2]

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

The table gives the scores of students in a short quiz.

Score	Frequency
2	3
3	7
4	k
5	5

The total number of students is 20.

(a) Find the value of k. [1]

(b) Calculate the mean score. [2]

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

The table gives the ages of 30 volunteers at a charity event.

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

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

(b) Calculate an estimate of the standard deviation of the ages. [2]

6. EASY **[3 marks]**

The daily high temperatures recorded at a weather station over one month have a mean of $22.4\,^{\circ}$C and a standard deviation of $3.8\,^{\circ}$C. The temperatures are converted to degrees Fahrenheit using the formula $F = 1.8C + 32$.

(a) Find the mean temperature in degrees Fahrenheit. [2]

(b) Write down the standard deviation of the temperatures in degrees Fahrenheit. [1]

7. EASY **[3 marks]**

The masses (in grams) of ten stones collected on a beach are listed below.

$$45,\quad 52,\quad 58,\quad 63,\quad 67,\quad 71,\quad 78,\quad 81,\quad 89,\quad 124$$

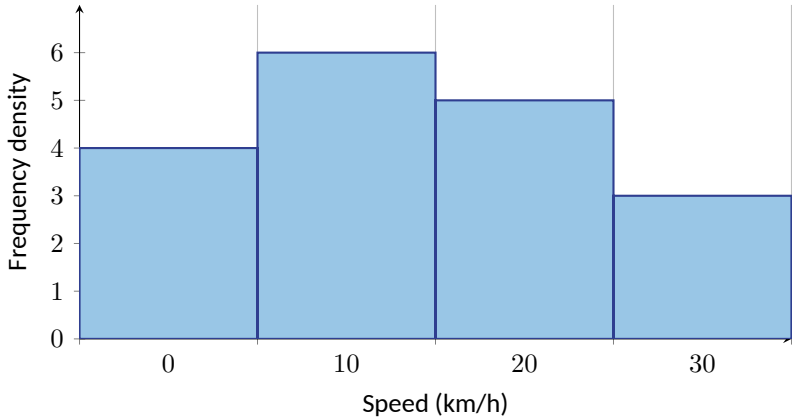
For this data set: $Q_1 = 58$, $Q_3 = 78$.

(a) Calculate the upper outlier fence. [2]

(b) Determine whether the value 124 is an outlier. Justify your answer. [1]

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

The histogram shows the speeds (in km/h) of 180 vehicles recorded at a checkpoint. All class widths are 10 km/h.



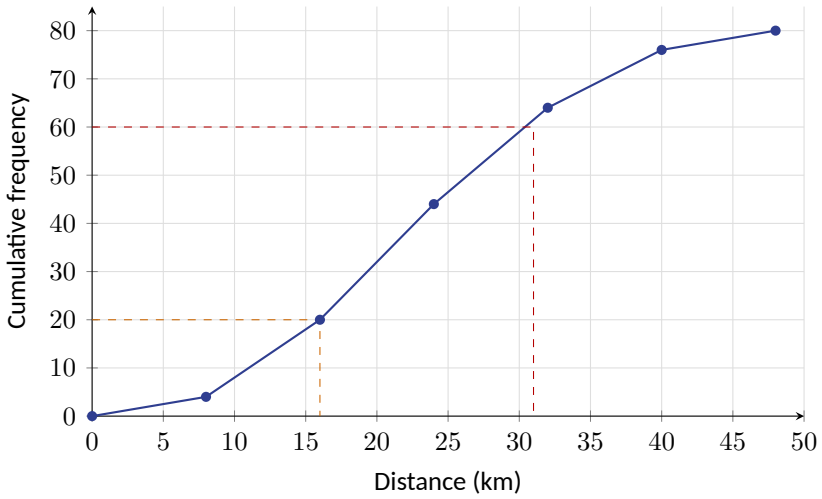
(a) Write down the modal class. [1]

(b) Find the number of vehicles with speed $30 \leq v < 40$ km/h. [1]

(c) Find the total number of vehicles. [1]

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

The cumulative frequency graph shows the distances (in km) travelled to work by 80 employees.



(a) Write down the lower quartile, Q_1. [1]

(b) Use the graph to find Q_3. [1]

(c) Hence find the interquartile range. [1]

10. EASY **[3 marks]**

A researcher surveys people leaving a supermarket to find out about their weekly spending habits. She stops and interviews every fifth person who exits the store over a two-hour period on a Tuesday morning.

(a) Name the sampling method being used. [1]

(b) State one reason why this method may not produce a representative sample of all supermarket customers. [1]

(c) Suggest one improvement to make the sample more representative. [1]

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

Scan me for help

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

The table gives the distances (in km) cycled by 36 participants in a charity ride.

Distance, d (km)	Frequency
$0 \leq d < 10$	3
$10 \leq d < 20$	8
$20 \leq d < 30$	12
$30 \leq d < 40$	9
$40 \leq d < 50$	4

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

(b) Calculate an estimate of the mean distance cycled. [2]

(c) Find the percentage of participants who cycled at least 30 km. [1]

12. MEDIUM **[4 marks]**

A sample of 25 parcels delivered by a courier service has the following five-number summary for their masses (in kg):

$$\text{Minimum} = 0.8, \quad Q_1 = 2.4, \quad \text{Median} = 4.1, \quad Q_3 = 6.5, \quad \text{Maximum} = 9.7$$

(a) Find the interquartile range. [1]

(b) Determine whether 9.7 kg could be an outlier. Show all working and justify your answer. [2]

(c) Describe the skewness of the data. Justify your answer using values from the five-number summary. [1]

13. MEDIUM **[4 marks]**

A hotel manager wants to survey 30 employees about workplace satisfaction. The hotel employs 45 reception staff, 75 kitchen staff, and 30 housekeeping staff.

(a) Name the type of sampling that ensures each staff group is represented in proportion to its size. [1]

(b) Find the number of employees from each group that should be included in the sample. [2]

(c) Suggest one reason why this survey may not produce fully reliable data. [1]

14. MEDIUM **[4 marks]**

A scientist records the pH of a water sample on six occasions:

$$6.8, \quad 7.1, \quad 7.0, \quad 6.9, \quad 7.2, \quad 7.0$$

(a) Calculate the mean pH. [1]

(b) Calculate the unbiased estimate of the population standard deviation, s_{n-1}. Give your answer correct to 3 significant figures. [3]

15. MEDIUM **[4 marks]**

In an end-of-term test, students' scores out of 60 are recorded. The scores have a mean of $\bar{x} = 38.4$ and a standard deviation of $\sigma = 7.2$. The teacher rescales all scores to a percentage using the formula $P = \dfrac{100}{60}\,s$, where s is the original score.

(a) Find the mean percentage score. [2]

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

(c) A student scored 45 out of 60. State their percentage score. [1]

16. MEDIUM **[4 marks]**

A researcher codes a dataset using $y = \dfrac{x - 50}{5}$. The coded data has a mean of $\bar{y} = 2.4$ and a standard deviation of $\sigma_y = 1.6$.

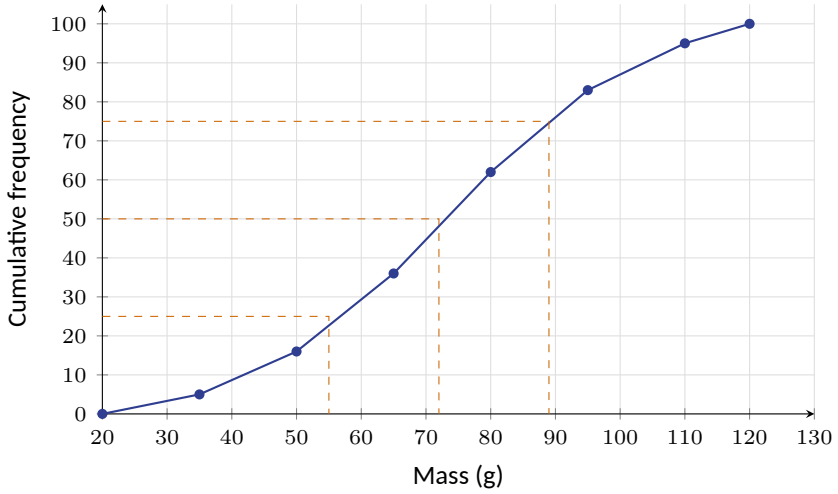
(a) Find the mean of the original data, $\bar{x}$. [2]

(b) Find the standard deviation of the original data, σ_x. [1]

(c) Find the variance of the original data. [1]

17. MEDIUM GDC **[4 marks]**

The cumulative frequency graph shows the masses (in grams) of 100 pebbles collected from a riverbed.



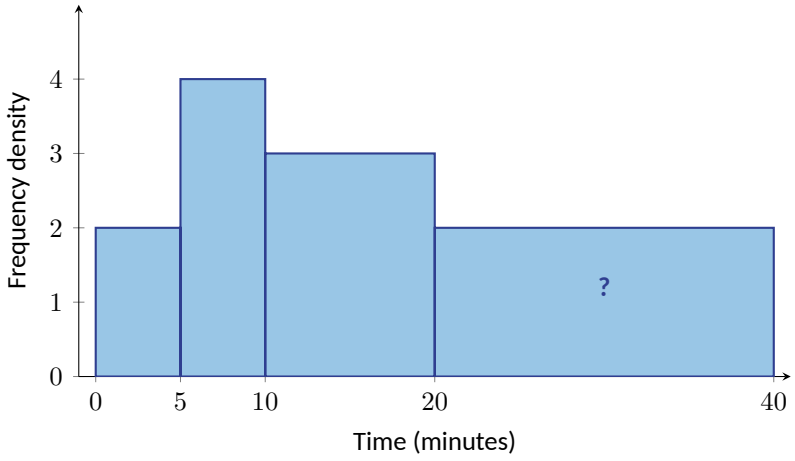
(a) Use the graph to find the median mass. [1]

(b) Find the interquartile range. [2]

(c) Estimate the number of pebbles with mass greater than 95 g. [1]

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

The histogram shows the times (in minutes) taken by 100 customers to complete an online purchase. The class widths are not all equal.



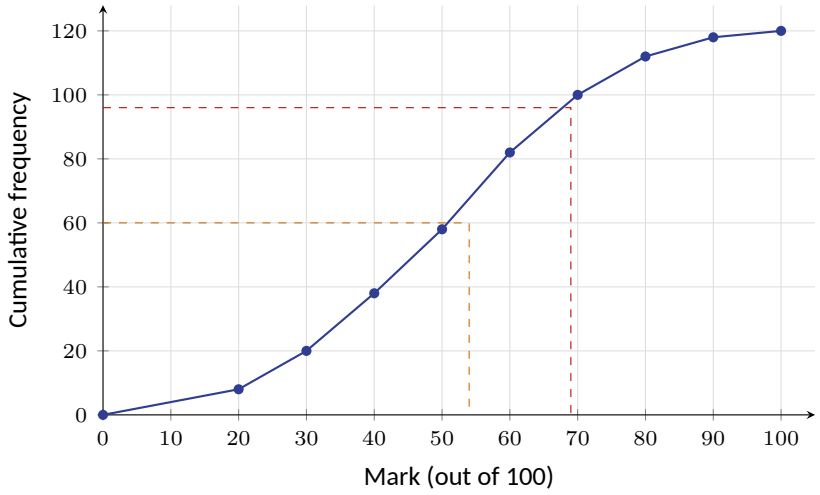
(a) Write down the frequency for the class $5 \leq t < 10$. [1]

(b) Given that the total number of customers is 100, find the frequency and frequency density for the class $20 \leq t < 40$. [2]

(c) State the modal class. [1]

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

The cumulative frequency graph shows the marks (out of 100) obtained by 120 students in a mathematics assessment.



(a) Use the graph to find the median mark. [1]

(b) Find the 80th percentile. [1]

(c) The pass mark is 45. Estimate the number of students who did not pass. [2]

20. MEDIUM **[4 marks]**

A market research company surveys 200 people about their coffee-drinking habits. They decide to interview 100 males and 100 females regardless of the actual proportion of males and females in the population.

(a) Name this sampling method. [1]

(b) Explain one disadvantage of this method compared with stratified random sampling. [2]

(c) Suggest one way to improve the validity of the data collected. [1]

Section C — Hard HARD (10 questions × 5 marks = 50 marks)

Scan me for help

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

The table gives the scores of a group of students in a problem-solving task.

Score	Frequency
2	a
4	5
6	b
8	3

The total number of students is 20 and the mean score is 4.8.

(a) Write down an equation connecting a and b. Hence find $a + b$. [2]

(b) Form a second equation using the mean. Solve to find the values of a and b. [3]

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

A botanist measures the heights (cm) of 36 plants. The sample gives $\bar{x} = 24.7$ cm and the population standard deviation $\sigma_n = 4.2$ cm.

(a) Find the unbiased estimate of the population variance. [2]

(b) Calculate a 95% confidence interval for the true mean height. [2]

(c) State what is meant by "95% confidence interval" in this context. [1]

23. HARD **[5 marks]**

The reaction times (ms) of 60 participants in a cognitive study have the following five-number summary:

$$\text{Minimum} = 140, \quad Q_1 = 200, \quad \text{Median} = 248, \quad Q_3 = 280, \quad \text{Maximum} = 420$$

(a) Show that the value 420 ms is an outlier. [3]

(b) Using values from the five-number summary, describe the skewness of the distribution. [1]

(c) State the effect on the mean of removing the outlier from the data set. [1]

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

A filling machine is set to dispense 500 mL of juice into bottles. A quality inspector selects a random sample of 25 bottles and measures the volume in each. The sample gives $\bar{x} = 498.4$ mL with an unbiased sample standard deviation of $s_{n-1} = 6.8$ mL.

(a) Calculate a 95% confidence interval for the true mean volume dispensed. [3]

(b) The manufacturer claims the machine is correctly calibrated. Using your confidence interval, comment on this claim. [2]

25. HARD **[5 marks]**

A social scientist wants to investigate the study hours per week of students across three faculties at a university: Science (240 students), Arts (180 students), and Engineering (120 students). She plans to use a stratified random sample of 60 students.

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

(b) Explain why stratified random sampling is more appropriate than systematic sampling for this investigation. [1]

(c) State one limitation of using a self-reported questionnaire to collect data on study hours. [1]

(d) Suggest one way to improve the reliability of the questionnaire. [1]

26. HARD **[5 marks]**

A data set x has mean μ and lower and upper quartiles Q_1 and Q_3 respectively. A new variable is defined by $y = 3x - 8$.

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

(b) The standard deviation of x is σ. Write down the standard deviation of y. [1]

(c) Find an expression for the upper outlier fence of y in terms of Q_1 and Q_3. Simplify your answer. [3]

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

The table gives the cumulative frequency distribution for the masses (in kg) of 80 parcels at a sorting depot.

Mass, m (kg)	Cumulative frequency
$m \leq 10$	0
$m \leq 20$	12
$m \leq 30$	36
$m \leq 40$	60
$m \leq 50$	72
$m \leq 60$	80

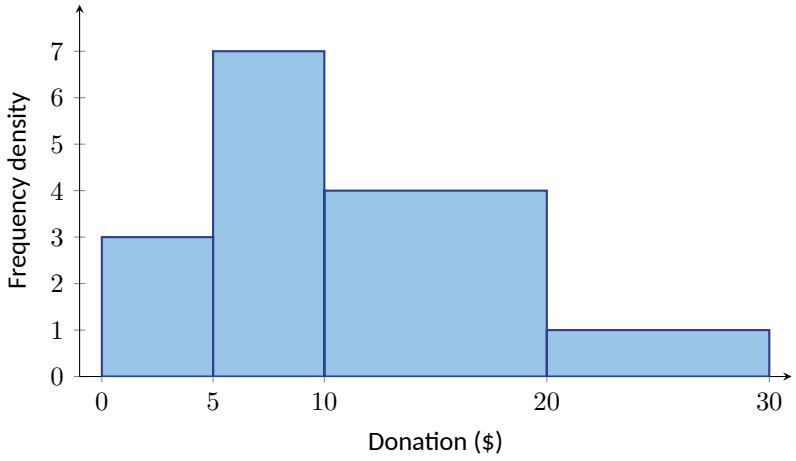
(a) Using linear interpolation, find the median mass. [2]

(b) Find Q_1 using linear interpolation. [2]

(c) Comment on the skewness of the data using your values of Q_1, the median, and Q_3. [1]

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

The histogram gives the amounts (in $) donated by visitors to a museum in one day. The class widths are not all equal.



(a) Use the histogram to complete the frequency table below. [2]

Donation, d ($)	Freq. density	Frequency
$0 \leq d < 5$	3	
$5 \leq d < 10$	7	
$10 \leq d < 20$	4	
$20 \leq d < 30$	1	

(b) Calculate an estimate of the mean donation. [2]

(c) Explain why this is an estimate rather than an exact value. [1]

29. HARD **[5 marks]**

The scores of 15 athletes in a strength test are listed in ascending order:

$$55,\ 62,\ 66,\ 70,\ 72,\ 75,\ 78,\ 80,\ 82,\ 84,\ 86,\ 89,\ 92,\ 96,\ 112$$

(a) Find the values of Q_1, Q_3 and the IQR. [2]

(b) Determine whether 112 is an outlier. Show all working. [2]

(c) A second team has mean score 81 and standard deviation 4. Compare the performance of the two teams, referring to both the mean and the consistency of scores. [1]

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

A transport company records the daily travel times (minutes) of 50 drivers. The results are summarised in the table below.

Time, t (min)	Frequency	Cumulative frequency
$0 \leq t < 20$	6	6
$20 \leq t < 40$	18	24
$40 \leq t < 60$	16	40
$60 \leq t < 80$	7	47
$80 \leq t < 100$	3	50

(a) Write down the cumulative frequency at $t = 40$. [1]

(b) Using linear interpolation, find the median travel time. [2]

(c) Calculate an estimate of the mean travel time. The company offers a bonus to all drivers if the mean is less than 45 minutes. Determine whether the bonus applies. [2]

Worked Solutions — Official Mark Scheme Style

Section A — Easy

Q1 — Frequency Table: Modal Class and Mean EASY

(a) Modal class $= 20 \leq l < 30$ **A1**

(b) Midpoints: 15, 25, 35, 45 **(M1)**

$$\bar{x} = \frac{6(15) + 14(25) + 12(35) + 8(45)}{40} = \frac{90 + 350 + 420 + 360}{40} = \frac{1220}{40}$$

$\bar{x} = 30.5$ cm **A1**

Note: Award (M1) for correct use of midpoints with given frequencies. FT their midpoints.

Q2 — Box Plot: Median and IQR EASY

(a) Median $= 24$ minutes **A1**

(b) IQR $= Q_3 - Q_1 = 31 - 18 = 13$ minutes **M1 A1**

Q3 — Stratified Sampling EASY

(a) Year 10: $\frac{160}{480} \times 60 = 20$ students **A1**

(b) Year 11: $\frac{180}{480} \times 60 = 22.5 \approx 23$ students **M1**

Year 12: $\frac{140}{480} \times 60 = 17.5 \approx 17$ students; total check $20 + 23 + 17 = 60$ **A1**

Note: Accept Year 11 = 22, Year 12 = 18 (total = 60). Award M1 for correct proportion method; A1 for both values consistent with total = 60.

Q4 — Frequency Table with Unknown EASY

(a) $3 + 7 + k + 5 = 20 \implies k = 5$ **A1**

(b) $\bar{x} = \frac{3(2) + 7(3) + 5(4) + 5(5)}{20} = \frac{6 + 21 + 20 + 25}{20} = \frac{72}{20} = 3.6$ **M1 A1**

Note: Award M1 for correct Σfx attempt with their k. FT from (a).

Q5 — Frequency Table: Modal Class and SD EASY

(a) Modal class $= 40 \leq a < 50$ **A1**

(b) Midpoints: 25, 35, 45, 55

$$\bar{x} = \frac{4(25)+9(35)+11(45)+6(55)}{30} = \frac{100+315+495+330}{30} = \frac{1240}{30} = 41.\overline{3}$$ **(M1)**

Using GDC (or formula $\sigma_n = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2}$): $\sum fx^2 = 100(625) + \ldots = 53\,950$

$$\sigma_n = \sqrt{\frac{53950}{30} - \left(\frac{1240}{30}\right)^2} = \sqrt{1798.3\overline{3} - 1708.4\overline{4}} = \sqrt{89.8\overline{8}} = 9.480\ldots \approx 9.48$$ **A1**

Note: Accept 9.48 (3 sf). Award (M1) implied if correct answer seen from GDC.

Q6 — Linear Transformation of Data EASY

(a) New mean $= 1.8(22.4) + 32 = 40.32 + 32 = 72.32\ldots \approx 72.3\,^\circ$F **M1 A1**

(b) New SD $= 1.8 \times 3.8 = 6.84\,^\circ$F **A1**

Note: In part (b), adding a constant does not affect the SD. Award A1 for 6.84 only.

Q7 — Outlier Test EASY

(a) IQR $= 78 - 58 = 20$ **(M1)**

Upper outlier fence $= Q_3 + 1.5 \times$ IQR $= 78 + 1.5(20) = 78 + 30 = 108$ g **A1**

(b) Since $124 > 108$, the value 124 g **is** an outlier. **R1**

Note: R1 requires explicit comparison $124 > 108$ and conclusion; statement alone without comparison scores R0.

Q8 — Histogram Reading EASY

(a) Modal class $= 10 \le v < 20$ km/h (highest frequency density = 6) **A1**

(b) Frequency $=$ fd $\times$ class width $= 3 \times 10 = 30$ vehicles **A1**

(c) Total $= 40 + 60 + 50 + 30 = 180$ vehicles **A1**

Note: For (c), award A1 for correct sum of all four class frequencies.

Q9 — Cumulative Frequency: Quartiles EASY

(a) Q_1: read at cumulative frequency $= 20 \implies Q_1 = 16$ km **A1**

(b) Q_3: read at cumulative frequency $= 60 \implies Q_3 = 31$ km **A1**

(c) IQR $= Q_3 - Q_1 = 31 - 16 = 15$ km **A1**

Note: Accept values within ± 1 km of the stated answers. FT from (a) and (b) for (c).

Q10 — Sampling Method and Validity EASY

(a) Systematic sampling **A1**
(b) Any one of: only Tuesday morning shoppers are sampled / the sample may not represent weekend or evening shoppers / certain customer groups (e.g. working people) are under-represented. **R1**
(c) Any valid improvement, e.g.: survey at different times and days of the week / use random sampling from a customer database. **A1**
Note: Award R1 only if the reason is linked to the specific context (time/day restriction). Award A1 for a concrete, actionable suggestion.

Section B — Medium

Q11 — Grouped Frequency Table MEDIUM

(a) Modal class $= 20 \leq d < 30$ km **A1**
(b) Midpoints: 5, 15, 25, 35, 45

$$\bar{d} = \frac{3(5) + 8(15) + 12(25) + 9(35) + 4(45)}{36} = \frac{15 + 120 + 300 + 315 + 180}{36} = \frac{930}{36} = 25.8\overline{3}\ldots \approx 25.8 \text{ km}$$ **M1 A1**

(c) Cyclists with $d \geq 30$: $9 + 4 = 13$

$$\text{Percentage} = \frac{13}{36} \times 100 = 36.1\ldots\% \approx 36.1\%$$ **A1**

Note: Award M1 for correct Σfd with midpoints. FT in (c) from their $n = 36$ total.

Q12 — Box Plot and Outlier Test MEDIUM

(a) IQR $= Q_3 - Q_1 = 6.5 - 2.4 = 4.1$ kg **A1**
(b) Upper fence $= Q_3 + 1.5 \times \text{IQR} = 6.5 + 1.5(4.1) = 6.5 + 6.15 = 12.65$ kg **M1**
Since $9.7 < 12.65$, the value 9.7 kg is **not** an outlier. **A1**
(c) $Q_3 - Q_2 = 6.5 - 4.1 = 2.4$; $Q_2 - Q_1 = 4.1 - 2.4 = 1.7$
Since $Q_3 - Q_2 > Q_2 - Q_1$, the data is **positively skewed** (skewed right). **R1**
Note: R1 requires both the numerical comparison and the correct skewness conclusion.

Q13 — Stratified Sampling and Bias MEDIUM

(a) Stratified random sampling **A1**
(b) Total staff $= 45 + 75 + 30 = 150$ **(M1)**
Reception: $\frac{45}{150} \times 30 = 9$; Kitchen: $\frac{75}{150} \times 30 = 15$; Housekeeping: $\frac{30}{150} \times 30 = 6$ **A1A1**
(c) Any valid reason, e.g.: employees may not answer honestly due to fear of repercussions / response bias in self-reported satisfaction. **R1**

Q14 — Sample Standard Deviation MEDIUM

(a) $\bar{x} = \dfrac{6.8 + 7.1 + 7.0 + 6.9 + 7.2 + 7.0}{6} = \dfrac{42.0}{6} = 7.00$ **A1**

(b) $\sum(x_i - \bar{x})^2 = (0.04) + (0.01) + (0) + (0.01) + (0.04) + (0) = 0.10$ **M1**

$s_{n-1} = \sqrt{\dfrac{0.10}{5}} = \sqrt{0.02} = 0.14142\ldots \approx 0.141$ **A1 A1**

Note: Award M1 for correct $\Sigma(x_i - \bar{x})^2$ attempt using their $\bar{x}$. Award A1 for 0.10 (or correct SS); A1 for 0.141 (3 sf). Also accept GDC value of s_{n-1} directly.

Q15 — Linear Transformation: Exam Scores MEDIUM

(a) New mean $= \dfrac{100}{60} \times 38.4 = \dfrac{3840}{60} = 64.0\%$ **M1 A1**

(b) New SD $= \dfrac{100}{60} \times 7.2 = 12.0\%$ **A1**

(c) Student's percentage $= \dfrac{100}{60} \times 45 = 75.0\%$ **A1**

Note: In (b), adding a constant does not affect SD; only the multiplicative factor $100/60$ applies.

Q16 — Coded Data MEDIUM

(a) From $y = \dfrac{x - 50}{5}$: $x = 5y + 50$

$\bar{x} = 5\bar{y} + 50 = 5(2.4) + 50 = 12 + 50 = 62.0$ **M1 A1**

(b) $\sigma_x = 5\sigma_y = 5(1.6) = 8.0$ **A1**

(c) Variance $= \sigma_x^2 = 8.0^2 = 64$ **A1**

Note: Award M1 for correctly inverting the coding formula. Adding constants does not change SD or variance.

Q17 — Cumulative Frequency Graph MEDIUM

(a) Median: read at cf $= 50 \implies$ median $= 72$ g **A1**

(b) Q_1: cf $= 25 \implies Q_1 = 55$ g **M1**

Q_3: cf $= 75 \implies Q_3 = 89$ g

IQR $= 89 - 55 = 34$ g **A1**

(c) At $m = 95$ g: cf ≈ 83 (reading from graph)

Number with mass > 95 g $= 100 - 83 = 17$ pebbles **A1**

Note: Accept cf at 95 in range 81–85; FT accordingly. Accept ± 2 on median and quartile readings.

Q18 — Histogram with Unequal Class Widths MEDIUM

(a) Frequency for $5 \leq t < 10$: fd $\times$ width $= 4 \times 5 = 20$ customers **A1**

(b) Known frequencies: $[0, 5) : 2 \times 5 = 10$; $[5, 10) : 20$; $[10, 20) : 3 \times 10 = 30$

Frequency for $[20, 40) = 100 - 10 - 20 - 30 = 40$ **M1**

fd for $[20, 40) = \dfrac{40}{20} = 2.0$ **A1**

(c) Modal class $= 5 \leq t < 10$ (highest frequency density $= 4$) **A1**

Note: In (c), modal class uses highest frequency density for grouped data. Award A1 for $5 \leq t < 10$ only.

Q19 — Cumulative Frequency: Percentiles and Pass Rate MEDIUM

(a) Median: read at cf $= 60 \implies$ median $= 54$ marks **A1**

(b) 80th percentile: 80% of $120 = 96$th student; read at cf $= 96 \implies 69$ marks **A1**

(c) At mark $= 45$: read cf ≈ 44 from graph **M1**

Estimated number who did not pass ≈ 44 students **A1**

Note: Accept cf at 45 in range 42–46. In (b), award A1 for correct reading of 96th student. FT from graph for (c).

Q20 — Quota Sampling and Validity MEDIUM

(a) Quota sampling **A1**

(b) Quota sampling does not use random selection within each group, so the sample may be biased by the interviewer's choices / does not reflect the actual proportion of males to females in the population, so results may not generalise. **R1 R1**

Award R1 for each distinct valid point up to 2.

(c) Any one of: use a stratified random sample in proportion to the actual gender split / select participants randomly from a customer database. **A1**

Section C — Hard

Q21 — Reconstruction from Mean and Total HARD

(a) Total frequency equation: $a + 5 + b + 3 = 20 \implies a + b = 12$ **M1 A1**

(b) Mean equation: $\dfrac{2a + 4(5) + 6b + 8(3)}{20} = 4.8$ **M1**

$2a + 20 + 6b + 24 = 96 \implies 2a + 6b = 52 \implies a + 3b = 26$

Substituting $a = 12 - b$: $(12 - b) + 3b = 26 \implies 2b = 14 \implies b = 7$ **A1**

$a = 12 - 7 = 5$ **A1**

Q22 — Unbiased Variance and Confidence Interval HARD

(a) $s_{n-1}^2 = \dfrac{n}{n-1} \cdot \sigma_n^2 = \dfrac{36}{35} \times 4.2^2 = \dfrac{36}{35} \times 17.64 = 18.144$ **M1 A1**

(b) $s_{n-1} = \sqrt{18.144} = 4.2595\ldots$

95% CI: $24.7 \pm 1.96 \times \dfrac{4.2595\ldots}{\sqrt{36}} = 24.7 \pm 1.96 \times 0.7099\ldots = 24.7 \pm 1.391\ldots$ **M1**

CI: $(23.3, 26.1)$ cm (to 3 sf) **A1**

(c) If this process were repeated many times, 95% of the confidence intervals constructed would contain the true mean height of the plants. **R1**

Note: In (c) do not award R1 for "the probability that the true mean is in this interval is 95%".

Q23 — Box Plot Outlier and Skewness HARD

(a) IQR $= Q_3 - Q_1 = 280 - 200 = 80$ ms **M1**

Upper outlier fence $= Q_3 + 1.5 \times \text{IQR} = 280 + 1.5(80) = 280 + 120 = 400$ ms **A1**

Since $420 > 400$, the value 420 ms is an outlier. **R1**

(b) $Q_2 - Q_1 = 248 - 200 = 48$; $Q_3 - Q_2 = 280 - 248 = 32$

Since $Q_2 - Q_1 > Q_3 - Q_2$, the data is **negatively skewed** (skewed left). **R1**

(c) Removing the outlier will **increase** the mean, since 420 ms pulls the mean upward. **R1**

Note: In (a), full marks require calculation of fence, comparison with 420, and explicit conclusion.

Q24 — Confidence Interval and Decision HARD

(a) 95% CI: $\bar{x} \pm 1.96 \cdot \dfrac{s_{n-1}}{\sqrt{n}} = 498.4 \pm 1.96 \times \dfrac{6.8}{\sqrt{25}}$ **M1**

$= 498.4 \pm 1.96 \times 1.36 = 498.4 \pm 2.6656$

CI: $(495.7, 501.1)$ mL (to 4 sf) **A1 A1**

(b) Since 500 mL lies within the confidence interval $(495.7, 501.1)$, there is insufficient evidence to conclude that the machine is not correctly calibrated at the 95% level. **R1 R1**

Note: Award R1 for noting 500 is inside CI; R1 for a non-assertive conclusion in context.

Q25 — Stratified Sampling Design and Critique HARD

(a) Total students $= 240 + 180 + 120 = 540$

Science: $\dfrac{240}{540} \times 60 = \dfrac{4}{9} \times 60 = 26.\overline{6} \approx 27$

Arts: $\dfrac{180}{540} \times 60 = 20$; Engineering: $\dfrac{120}{540} \times 60 = 13.\overline{3} \approx 13$ **M1 A1**

(Accept Science = 27, Arts = 20, Engineering = 13; total = 60. Or Science = 26 with Engineering = 14.)

(b) Stratified sampling ensures each faculty is represented in proportion to its size, reducing sampling bias; systematic sampling from a combined list might over- or under-represent smaller faculties. **R1**
(c) Students may over-report study hours / responses are subject to social desirability bias. **R1**
(d) Use a test–retest method: administer the questionnaire twice and compare responses / pilot the questionnaire and refine ambiguous questions. **A1**

Q26 — Linear Transform and Outlier Expression HARD

(a) $\bar{y} = 3\mu - 8$ **A1**
(b) SD of $y = 3\sigma$ **A1**
(c) $Q_{3,y} = 3Q_3 - 8; \quad Q_{1,y} = 3Q_1 - 8$ **M1**
$\text{IQR}_y = Q_{3,y} - Q_{1,y} = (3Q_3 - 8) - (3Q_1 - 8) = 3(Q_3 - Q_1) = 3\,\text{IQR}_x$ **A1**
Upper fence $= Q_{3,y} + 1.5\,\text{IQR}_y = (3Q_3 - 8) + 1.5 \times 3(Q_3 - Q_1)$
$= 3Q_3 - 8 + 4.5(Q_3 - Q_1) = 7.5Q_3 - 4.5Q_1 - 8$ **A1**
Note: Award M1 for attempting to express $Q_{3,y}$ or IQR_y in terms of Q_1, Q_3. Award A1A1 for correct IQR and correct final fence expression.

Q27 — Linear Interpolation: Median and Q1 HARD

(a) Median: 40th value. From table: cf at $m = 30$ is 36, cf at $m = 40$ is 60.
$\text{Median} = 30 + \dfrac{40 - 36}{60 - 36} \times 10 = 30 + \dfrac{4}{24} \times 10 = 30 + 1.6\overline{6} = 31.7$ kg **M1 A1**
(b) Q_1: 20th value. cf at $m = 20$ is 12, cf at $m = 30$ is 36.
$Q_1 = 20 + \dfrac{20 - 12}{36 - 12} \times 10 = 20 + \dfrac{8}{24} \times 10 = 20 + 3.\overline{3} = 23.3$ kg **M1 A1**
(c) Q_3: 60th value $\implies Q_3 = 40$ kg (exactly at boundary).
$Q_3 - \text{Med} = 40 - 31.7 = 8.3$; $\text{Med} - Q_1 = 31.7 - 23.3 = 8.4$
Since the two halves are approximately equal, the distribution is approximately **symmetric**. **R1**

Q28 — Histogram: Frequency Table and Mean HARD

(a) Frequencies from histogram (freq = fd × width): **M1**
$[0,5)$: $3 \times 5 = 15$; $[5,10)$: $7 \times 5 = 35$; $[10,20)$: $4 \times 10 = 40$; $[20,30)$: $1 \times 10 = 10$ **A1**
(b) Midpoints: 2.5, 7.5, 15, 25; total $n = 100$
$\bar{d} = \dfrac{15(2.5) + 35(7.5) + 40(15) + 10(25)}{100} = \dfrac{37.5 + 262.5 + 600 + 250}{100} = \dfrac{1150}{100} = \$11.50$ **M1 A1**
(c) The midpoint of each class is used to represent all values in that class; the actual values within each class are unknown, so the mean is an estimate. **R1**

Q29 — Quartiles, Outlier Test, and Comparison HARD

(a) $n = 15$; Q_1 is the 4th value $= 70$; Q_3 is the 12th value $= 89$

IQR $= 89 - 70 = 19$ **M1 A1**

(b) Upper outlier fence $= Q_3 + 1.5 \times$ IQR $= 89 + 1.5(19) = 89 + 28.5 = 117.5$ **M1**

Since $112 < 117.5$, the value 112 is **not** an outlier. **A1**

(c) Team A: $\bar{x} = \dfrac{55 + 62 + 66 + 70 + 72 + 75 + 78 + 80 + 82 + 84 + 86 + 89 + 92 + 96 + 112}{15} = \dfrac{1199}{15} = 79.9\ldots \approx 79.9$

Team B has a higher mean ($81 > 79.9$), so on average Team B performed better. **A1**

Team B (SD $= 4$) is far more consistent than Team A, which has much greater variability in scores. **R1**

Note: Award A1 for mean comparison with values; R1 for consistency/spread comparison referencing SD. Award R1 only with a contextual statement.

Q30 — Cumulative Frequency Table: Median and Mean with Decision HARD

(a) Cumulative frequency at $t = 40$: **24** **A1**

(b) Median: 25th value. From table: cf at $t = 40$ is 24; cf at $t = 60$ is 40.

25th value falls in $40 \leq t < 60$:

Median $= 40 + \dfrac{25 - 24}{16} \times 20 = 40 + \dfrac{1}{16} \times 20 = 40 + 1.25 = 41.25 \approx 41.3$ min **M1 A1**

(c) Midpoints: 10, 30, 50, 70, 90

$\bar{t} = \dfrac{6(10) + 18(30) + 16(50) + 7(70) + 3(90)}{50} = \dfrac{60 + 540 + 800 + 490 + 270}{50} = \dfrac{2160}{50} = 43.2$ min **M1**

Since $43.2 < 45$, the mean travel time is less than 45 minutes, so the **bonus applies**. **A1**

Note: Award M1 for correct Σft with midpoints; A1 for 43.2 and a clear decision referencing the 45-minute threshold.

Common Mistakes to Avoid

1. **Modal class vs mode.** In grouped data the *modal class* is the class with the highest frequency (or highest frequency density for unequal widths); never give a single number as the "mode" for grouped data.
2. **Outlier fence direction.** To test an unusually *large* value, use the **upper** fence $Q_3 + 1.5$ IQR; for an unusually *small* value, use $Q_1 - 1.5$ IQR. Mixing these up reverses your conclusion.
3. **s_n vs s_{n-1}.** Your GDC reports both: $\sigma x = s_n$ (population SD) and $Sx = s_{n-1}$ (sample/unbiased SD). For a confidence interval or unbiased estimate, **always use** Sx.
4. **Linear transform and SD.** Adding a constant b shifts the mean but does *not* change the SD or IQR. Only

the multiplicative factor a in $y = ax + b$ scales the spread: $\sigma_y = |a|\,\sigma_x$.

5. **Frequency density in histograms.** For unequal class widths, **do not read frequency directly from the y-axis**. Always multiply: frequency $=$ fd $\times$ class width.
6. **Cumulative frequency interpolation.** The *median* is the $\frac{n}{2}$th value, **not** the value at height $\frac{n}{2} + 1$. Read Q_1 at height $\frac{n}{4}$ and Q_3 at $\frac{3n}{4}$.
7. **Confidence interval wording.** Do **not** say "the probability the true mean is in this interval is 95%". The correct statement is: "95% of such intervals, if constructed repeatedly, would contain the true mean."
8. **Skewness from a box plot.** If $Q_3 - Q_2 > Q_2 - Q_1$, the upper half is more spread out $\Rightarrow$ *positive* (right) skew. If $Q_2 - Q_1 > Q_3 - Q_2 \Rightarrow$ *negative* (left) skew.