End of Topic Test – Chapter 1 (Test 2: Topics 4–6) — Mark Scheme

Total: 43 marks

1. Comparing distributions using box plots

Reading from the box plots (approximate values, tolerance ±2 allowed): Cohort Y: min 40, LQ ≈ 55, median ≈ 68, UQ ≈ 85, max 90. Cohort X: min 45, LQ ≈ 50, median ≈ 70, UQ ≈ 90, max 95.

(a) B1 Cohort X. B1 correct comparison of median positions (X's median line, ≈ 70, lies to the right of/above Y's, ≈ 68). *[2]*

(b) B1 ranges are equal (Y: $90 - 40 = 50$; X: $95 - 45 = 50$). B1 interquartile ranges differ (Y: $85 - 55 = 30$; X: $90 - 50 = 40$) – Cohort X has the larger IQR. B1 conclusion: Cohort X shows greater variability, since IQR (a measure unaffected by the single most extreme value) is larger for X. *[3]*

(c) B1 the maximum is a single, possibly unrepresentative value/potential outlier, not a measure of the cohort's overall performance. B1 the median (the "typical" student) is only marginally higher for X (≈ 70 vs ≈ 68), and X's lower quartile (≈ 50) is actually *below* Y's (≈ 55), meaning the weakest quarter of Cohort X did worse than the weakest quarter of Cohort Y. B1 balanced conclusion: the claim is not well supported – overall performance (median, spread) is similar or mixed, so "Cohort X performed better overall" is not justified by one extreme value alone. *[3]*

2. Grouped data calculations

Speed (km/h)	30–39	40–49	50–59	(continued) 60–69: freq 50, midpoint 64.5; 70–79: freq 20, midpoint 74.5. Total $n = 200$.
Frequency	12	38	70	
Midpoint	34.5	44.5	54.5	

(a) M1 $\sum fx = 12(34.5) + 38(44.5) + 70(54.5) + 50(64.5) + 20(74.5)$. M1 $\sum fx = 414 + 1691 + 3815 + 3225 + 1490 = 10635$. M1 mean $= 10635 \div 200$. A1 mean $= 53.2$ km/h (1 d.p.). *[4]*

(b) M1 median class located using $n/2 = 100$th value; cumulative frequencies: 12, 50, 120, 170, 200 $\Rightarrow$ median lies in the 50–59 class. M1 linear interpolation set up: $\text{median} = 49.5 + \dfrac{100 - 50}{70} \times 10$. A1 median $= 49.5 + 7.14 = 56.6$ (1 d.p.). A1 correct interpolation method shown (class boundaries used, not just class limits). *[4]*

(c) B2 (or B1+B1) valid limitation, e.g. the exact values within each class are unknown, so using the midpoint assumes data is evenly spread within the class, which may not be true; this makes the mean/median only an *estimate*, not an exact value. *[2]*

3. Histograms and frequency density

Length (cm)	Frequency	Width	Freq. density (freq/width)
10–14	8	4	2.0
14–18	22	4	5.5
18–22	40	4	10.0
22–30	30	8	3.75

(a) M1 class widths identified correctly (4, 4, 4, 8). M1 frequency density = frequency ÷ width calculated for at least 2 classes. A1 all four frequency densities correct: 2.0, 5.5, 10.0, 3.75. *[3]*

(b) B1 correct horizontal scale (length, 10–30 cm) and vertical scale (frequency density) with axes labelled. B1 all four bars drawn with the correct heights (matching part (a)) and correct (unequal) widths, with no gaps between bars. B1 bars correctly aligned to class boundaries (10, 14, 18, 22, 30). *[3]*

(c) M1 proportion of fish over 19 cm needed: within the 18–22 class (freq. density 10.0), the portion above 19 cm is 22 to 19, i.e. 3 cm wide ⇒ estimated frequency = $3 \times 10 = 30$ (out of 40 in that class), plus all of the 22–30 class (30 fish). M1 total suitable in sample = $30 + 30 = 60$ out of $n = 100$ fish sampled, i.e. proportion = $60/100 = 0.6$. M1 scale up: $0.6 \times 1{,}000{,}000$. A1 estimated 600,000 fish suitable to catch and sell (accept answers using a sensibly justified equivalent method, e.g. assuming uniform spread within the 18–22 class). *[4]*

4. Cumulative frequency and percentiles

(clinic A, $n = 200$; reading tolerance ±2 minutes allowed throughout this question)

(a) B1 median (value at cumulative frequency 100) ≈ 21 minutes. *[1]*

(b) B1 75th percentile (value at cumulative frequency 150) ≈ 28 minutes. *[1]*

(c) Clinic B (given): min 5, LQ 15, median 25, UQ 35, max 50. Clinic A (read from the c.f. graph): min ≈ 0, LQ (c.f. 50) ≈ 13, median ≈ 21, UQ (c.f. 150) ≈ 28, max ≈ 60.

B1 Clinic A box plot drawn with a correctly scaled axis (0–60) and a box from LQ to UQ with median marked, and whiskers to min/max, using the candidate's own reasonable readings from the graph. B1 comparison of medians: Clinic A's median (≈ 21) is lower than Clinic B's (25), so typical waits are shorter at Clinic A. B1 comparison of spread: Clinic A's IQR ($\approx 28 - 13 = 15$) is smaller than Clinic B's ($35 - 15 = 20$), so the middle 50% of waiting times are less variable at Clinic A. B1 Clinic A's maximum (≈ 60) is higher than Clinic B's (50), showing Clinic A has a longer extreme wait / positive skew despite shorter typical waits. B1 overall conclusion: Clinic A is generally faster and more consistent for most patients, but has a small number of much longer waits than Clinic B. *[5]*

5. Interpreting survey results

(a) B1+B1+B1 (any three) valid reasons, e.g. different sampling methods reach different populations (online panel vs. in-person at the park); different response/non-response patterns; timing/weather/day of the survey differed; question wording or interviewer effects differed between methods; sample sizes or demographic makeup differed. *[3]*

(b) B1 face-to-face interviews at park entrances only sample people who are already visiting/using the park, who are likely to be more satisfied than non-users. B1 this is a form of sampling bias (a biased/non-random sampling frame – only park users are reached, and only at the times/locations interviewers were present). B1 interviewer presence may also cause responders to give more positive/socially desirable answers face-to-face than they would anonymously online. *[3]*

(c) B2 (or B1+B1) valid suggestion, e.g. use the same method (e.g. both online, or both face-to-face) for both surveys; or use a broader, randomly selected sample for the face-to-face survey (not just park visitors) so both surveys draw from a comparable population. *[2]*