Chapter 1: Analysis of Data — Mark Scheme

Standard methods used throughout: standard deviation $\sigma = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n}}$ *(or equivalent sum-of-squares form); for raw/discrete lists, quartiles are found as the median of the lower half and median of the upper half of the ordered data; for frequency tables/grouped data, quartiles and percentiles are read from the cumulative frequency (using the* $n/4$*,* $n/2$*,* $3n/4$ *positions, with linear interpolation for grouped continuous data); an outlier is a value more than* $1.5\times IQR$ *below* Q_1 *or above* Q_3*. Any equivalent correct method should be credited.*

Exercise 1A

1 Data types. **a)** Qualitative: ii (ice cream flavour), iv (test grade), x (gender), xii (nationality).

b) Discrete: i (number of emails), v (collar size), ix (membership fees).

c) Continuous: iii (mass of kittens), vi (length of fingers), vii (time reading), viii (petal area), xi (age).

2 **a)** Primary data: collect it yourself, e.g. hand out a questionnaire to a sample of students asking how they travel and how much they spend on transport.

b) Secondary data: use data already collected by someone else, e.g. published transport-survey statistics (government or college records).

3 **a)** Example questions: "Would you support building a new leisure centre?"; "How often would you use it?"; "What facilities would you like it to include?"; "How much would you be willing to pay for membership?" (any relevant, unbiased, answerable question credited).

b) i. Biased towards commuters travelling by train in the morning; excludes non-commuters, non-train-users, and afternoon/evening city users — not representative of the whole city. ii. Excludes residents without email or not registered with the council; likely low response rate (non-response bias). iii. Voluntary-response bias: only readers with strong opinions (and who read that paper) are likely to write in — unrepresentative of the wider population.

4 Unreliable: participants are self-selected and motivated by winning a prize rather than answering honestly, there is no check on identity or repeat entries, and the sample is not representative of the wider population.

5 **a)** Use the sixth form's student register as the sampling frame; divide into strata (e.g. year group); use random numbers to select 80 students proportionally from each stratum (stratified sampling).

b) Use a sampling frame for the city (e.g. electoral roll/postcode list); use stratified or cluster sampling with random numbers to select 800 people, balancing representativeness against practicality.

6 Decide quotas for age/gender/area in proportion to the town's known population; interviewers stop people in different locations/times until each quota of the 400 is filled (non-random selection within each quota).

7 **a)** Divide Scotland into geographical clusters (e.g. towns/postcode areas); randomly select some clusters; survey everyone (or a random sample) within each selected cluster.

b) No — only people in the selected clusters can be chosen (people in non-selected areas have zero chance), so it is not a simple random sample of the whole population, although it is practical for covering a large area.

8 Priya's method is a convenience sample: biased (only people in the canteen at that moment), unrepresentative (may miss students eating elsewhere/at other times), and people queuing together may not answer independently. Improvements: use a random or stratified sample based on the year-group register, and spread interviews across several days/times/locations.

9 Obtain the practice's patient list (sampling frame), number the patients, and use a random number generator to select a sample. Problems: some patients may be uncontactable or decline to respond (non-response bias); balancing sample size against cost/practicality of following up.

10 Set up coordinates over the field (e.g. x : 0–60 m, y : 0–35 m); use a random number generator to generate 15 pairs of coordinates and sample a 1 m^2 quadrat at each point.

11 **a)** Simple random sampling: number everyone in the frame, use random numbers to select. Adv: unbiased, every person equally likely. Disadv: needs a complete sampling frame; sample may be impractically spread out.

b) Stratified sampling: divide by age/gender in known population proportions, randomly sample within each stratum. Adv: guarantees representation of subgroups, more precise. Disadv: needs accurate data on subgroup sizes.

c) Cluster sampling: divide the city into areas, randomly select a few areas to survey fully/partially. Adv: cheaper and practical, no full frame needed. Disadv: less precise, selected areas may not reflect the whole city.

d) Quota sampling: interviewers fill quotas matching known population characteristics. Adv: quick and cheap, no sampling frame needed. Disadv: not random — subject to interviewer bias.

12 **a)** Sample of 180 from 795 employees (factor 180/795): Warehouse M-FT 71, M-PT 5, F-FT 22, F-PT 18; Depot M-FT 14, M-PT 4, F-FT 7, F-PT 3; Office M-FT 5, M-PT 2, F-FT 9, F-PT 11; Distribution M-FT 5, M-PT 1, F-FT 3, F-PT 0 (rounded values total 180).

b) Other relevant characteristics: age, length of service, job role/seniority, shift pattern.

c) Problems: some employees on leave/unavailable; low response rates; ensuring genuinely random selection within each sub-group; confidentiality.

13 **a)** i. Sample of 6000 from 201 620 (factor 6000/201620): Female 18: 1137, Female 19: 644, Female 20: 277, Female Other: 518; Male 18: 1531, Male 19: 843, Male 20: 359, Male Other: 690 (rounded values total ≈6000). ii. Other relevant characteristics: sector/type of apprenticeship, region, level of apprenticeship. iii. Problems: obtaining up-to-date contact details; non-response; the wide "Other" age band. Could overcome by using multiple contact methods and following up non-responders.

b) Open research task: credit a short report referencing a genuine named annual apprentice survey, summarising its purpose and method.

14 **a)** Use the college register as the sampling frame; split into strata (e.g. year group/subject); randomly select 20% from each stratum using random numbers so the sample reflects the college's composition.

b) Example questions: "Do you support a four-day week?"; "How would it affect your studies/travel/part-time job?" (closed/Likert-scale questions are easiest to analyse).

c) Use a bar chart or pie chart for the (categorical) opinions, since these make it easy to compare proportions; justify choice of diagram.

Exercise 1B

1 **a)** Theme park prices ($n = 20$): mode = £165; median = £165; mean = £172.35; range = £260−£119 = £141.

b) Sponsorship earnings ($n = 19$): mode = £1.1m; median = £1.6m; mean ≈ £1.81m; range = £6.8m−£0.3m = £6.5m.

2 **a)** Deshi is wrong because the groups have different sizes (38 pickers, 11 operators), so simply averaging the two given means ignores the different weightings — a weighted mean is needed.

b) Overall mean $= \dfrac{38 \times 10.20 + 11 \times 14.80}{49} = \dfrac{387.60 + 162.80}{49} = \dfrac{550.40}{49} = £11.23$ per hour (2dp).

3 Total needed over 7 matches $= 21 \times 7 = 147$ goals; current total $= 18 \times 6 = 108$ goals; goals needed next match $= 147 - 108 =$ **39**.

4 **a)** They can all be correct because they are calculating different statistics/using different subsets of the data — e.g. different averages (mean vs. median vs. modal class size) or the mean taken over different groupings (e.g. one gender only vs. both). Each is a valid figure answering a slightly different question.

b) The overall mean class size (total students ÷ number of classes) generally best represents "typical" class size as it uses all the data, unless one class is unusually large/small, in which case the median is preferable. (Exact figures depend on reading the bar heights from the chart; credit sound reasoning consistent with the chart.)

5 **a)** i. Minibus lateness ($n = 20$): median = 1 min; $Q_1 = 0$, $Q_3 = 1.5$, IQR = 1.5 min. ii. Tomatoes ($n = 10$): median = 32.5; $Q_1 = 30$, $Q_3 = 35$, IQR = 5. iii. Coffee spend ($n = 9$): median = £2.40; $Q_1 =$ £2.15, $Q_3 =$ £3.45, IQR = £1.30.

b) Each dataset has an extreme outlier (23 minutes; 8 tomatoes; £8.60) which would distort the mean and range, whereas the median and IQR are resistant to extreme values and give a more useful "typical" picture.

6 **a)** Test marks: mean = 34; $\sigma \approx 7.23$.

b) Temperatures: mean ≈ 21.71°C; $\sigma \approx 1.98$°C.

c) Journey times: mean = 30 min; $\sigma \approx 5.90$ min.

d) Netball points: mean ≈ 34.67; $\sigma \approx 25.90$.

7 **a)** Discrete: (a) test marks, (d) netball points.

b) Continuous: (b) temperatures, (c) journey times.

8 **a)** Heights ($n = 19$, all values unique so no mode): median = 165 cm; mean ≈ 164.16 cm; range = 183 − 132 = 51 cm; $Q_1 = 156$, $Q_3 = 175$, IQR = 19 cm; $\sigma \approx 12.44$ cm.

b) Median and IQR are most suitable: the lowest value (132 cm) is noticeably separated from the rest, which would distort the mean/range, and with all values unique the mode is not useful.

9 **a)** Stem-and-leaf (stem = tens digit): 0|6 7; 1|2 2 4 5 7 8; 2|0 3 6 8 9 9; 3|0 5 6 7 8 8 9; 4|0 0 1 4 7 8 9; 5|1 1 1 1 2 2 2 2 3 3 3. Key: 1 | 2 means 12 minutes.

b) i. Mode = 51 and 52 min (bimodal, both occur 4 times). ii. Range = 53 − 6 = 47 min. iii. Median = 38 min. iv. IQR = 51 − 23 = 28 min.

c) Model paragraph: stays vary widely (6 to 53 min, range 47). The distribution is skewed towards longer stays, with a cluster around 51–53 minutes (mode), suggesting many drivers park for just over a common time limit. The median (38 min) and IQR (28 min) show a wide spread of parking durations.

10 **a)** Route C ($n = 14$, no mode): range $= 44 - 14 = 30$ min; median $= 28.5$ min; IQR $= 36 - 23 = 13$ min. Route D ($n = 13$, no mode): range $= 41 - 16 = 25$ min; median $= 29$ min; IQR $= 33 - 23 = 10$ min.

b) i. Both routes have a similar median time (28.5 vs 29 min), but Route C has a larger range (30 vs 25) and larger IQR (13 vs 10), so Route C's times are more variable/less consistent than Route D's. ii. Femi has 35 minutes available. Route D's upper quartile (33 min) is comfortably under 35, whereas Route C's (36 min) is slightly over — Route D is the more reliable choice even though its recorded maximum (41 min) is higher than Route C's typical spread suggests care is still needed.

11 Report question — draw a back-to-back stem-and-leaf (stems = tens digit) for the ages. Illustrative summary: women ($n = 18$): mean ≈ 28.4, median $= 28$, $\sigma \approx 7.1$ years; men ($n = 20$): mean ≈ 32.2, median $= 31.5$, $\sigma \approx 7.2$ years. Model paragraph: the men who married this week were on average about 4 years older than the women (both mean and median agree), while the spread of ages was similar for both groups (SD ≈ 7 years each) — the men are consistently a little older rather than a few much older men skewing the mean.

Exercise 1C

1 Trainee ages: female ($n = 34$): mode $= 18$; range $= 20 - 17 = 3$ (no 21-year-olds); median $= 18$; mean ≈ 18.32. Male ($n = 43$): mode $= 21$; range $= 21 - 17 = 4$; median $= 19$; mean ≈ 18.98. Comparison: males include an older group (mode 21) with no female equivalent; males' ages are slightly more spread (range 4 vs 3) and on average about half a year to a year older (mean/median) than the females.

2 **a)** Draw two box plots from: Marcus ($n = 11$): min 0, $Q_1 = 4$, median $= 24$, $Q_3 = 41$, max 82. Devon ($n = 11$): min 11, $Q_1 = 21$, median $= 24$, $Q_3 = 41$, max 53.

b) Marcus: mean ≈ 28.55, $\sigma \approx 25.67$. Devon: mean ≈ 29.91, $\sigma \approx 13.17$. Devon has a slightly higher mean but a much lower SD — his scores are far more consistent than Marcus's, which range from 0 to 82.

c) Pick Devon: similar mean/median score, but Devon is much more consistent (lower SD/IQR), making his performance more reliable/predictable, whereas Marcus is prone to very low as well as very high scores.

3 **a)** Wages ($n = 22$): median $= £19\,500$; $Q_1 = £19\,500$, $Q_3 = £22\,000$, IQR $= £2\,500$.

b) Upper fence $= Q_3 + 1.5\times\text{IQR} = 22\,000 + 1.5(2\,500) = £25\,750$. Since $£64\,000 > £25\,750$, it is an outlier.

c) The true mean is $\approx £22\,091$, inflated by the single £64 000 outlier; in fact 16 of the 22 workers (73%) earn only £19 500, well below any quoted "average" near £21 000–£22 000 — the mean (or the manager's figure) is not representative of a typical worker's wage; the median (£19 500) gives a fairer picture.

4 **a)** Elmwood ($n = 12$): mean ≈ 18.33, $\sigma \approx 9.17$, median $= 15$, $Q_1 = 11.5$, $Q_3 = 27$, IQR $= 15.5$. Fairview ($n = 12$): mean ≈ 17.67, $\sigma \approx 8.52$, median $= 17.5$, $Q_1 = 9.5$, $Q_3 = 23$, IQR $= 13.5$.

b) Box plots from the five-figure summaries above.

c) i. Naomi (Elmwood safer): supported by Fairview's higher median (17.5 vs 15), suggesting Fairview typically has more break-ins. ii. Dean (Fairview safer): supported by Elmwood's higher mean (18.33 vs 17.67) and larger spread (SD 9.17 vs 8.52, IQR 15.5 vs 13.5) — Elmwood has more extreme high months (25, 29, 32, 35) pulling its mean and spread up.

5 **a)** i. Mode = 13. ii. Range $= 17 - 12 = 5$.

b) i. Median = 14. ii. IQR $= 15 - 13 = 2$. iii. Box plot: min 12, $Q_1 = 13$, median 14, $Q_3 = 15$, max 17.

c) i. Mean ≈ 14.07. ii. $\sigma \approx 1.08$.

6 **a)** i. Mode = 50 seeds; median = 50; mean = 49.5. ii. "Even though a few packets contain fewer than 50 seeds, the mode and median are exactly 50 and the mean (49.5) is very close, so on average the claim of '50 seeds' is fair and accurate."

b) $Q_1 = 49$, $Q_3 = 50$, IQR = 1; upper fence $= 50 + 1.5(1) = 51.5$; since $54 > 51.5$, 54 is an outlier.

c) Two improvements: (1) sample more than 100 packets / a larger proportion of stock, as the first 100 may not represent the whole production run; (2) select packets randomly from across the stock/different batches rather than just "the first 100", to avoid bias from a single batch.

7 Compare the boxplots' medians, spread (IQR/range/whisker length) and any skew for revenue vs. running costs. Model points: state which distribution has the higher/lower median, which is more spread out (bigger gap between best- and worst-funded clubs), and whether either shows skew (e.g. a long upper whisker suggesting a few very high-revenue clubs). (Exact reading depends on the printed chart.)

8 **a)** i. Theo has used the frequency column's mode instead of the age with the highest frequency, and calculated the range using an age (18) that has zero students. ii. Correct mode = 14; correct range $= 17 - 12 = 5$.

b) i. Olu has likely entered the frequencies as if they were data values (or omitted the frequency list) in his calculator. ii. Correct mean ≈ 14.26 years; correct $\sigma \approx 1.50$ years.

c) i. Checking with the cumulative frequencies $(8, 19, 35, 41, 53, 57, 57)$ actually confirms median = 14 and IQR $= 16 - 13 = 3$ — so, unlike Theo and Olu, Freya's final values are correct; the teaching point is to verify using cumulative frequencies rather than assume an error must exist. ii. Median = 14, IQR = 3. iii. Box plot: min 12, $Q_1 = 13$, median 14, $Q_3 = 16$, max 17. The box's upper half (14 to 16) is wider than its lower half (13 to 14), so the ages are slightly positively (right-)skewed, with most students aged 13–16.

9 **a)** Before ($n = 40$): min 5, $Q_1 = 6$, median 7, $Q_3 = 8$, max 10. After ($n = 38$): min 3, $Q_1 = 4$, median 5, $Q_3 = 6$, max 7.

b) After training, both the mean (5.0 vs 7.1) and median (5 vs 7) number of errors fell, while the spread stayed similar (IQR = 2 in both cases, $\sigma \approx 1.15$ in both) — the training reduced the typical error count without much change in consistency; the whole distribution shifted to fewer errors.

10 **a)** Day of test ($n = 30$): min 5, $Q_1 = 8$, median 10, $Q_3 = 11$, max 12. After 1 week ($n = 24$): min 1, $Q_1 = 3$, median 4.5, $Q_3 = 6$, max 8.

b) Mean/SD: day of test ≈ 9.5, $\sigma \approx 2.01$; after 1 week ≈ 4.38, $\sigma \approx 1.84$.

c) Volunteers remembered far more items immediately (median 10 of 12) than after a week (median 4.5) — recall roughly halved. Spread was similar (IQR 3 in both), so the decline was fairly consistent across volunteers, with no overlap between the two typical (IQR) ranges, confirming a genuine decline in recall.

11 **a)** i. Method: read the frequency for each number of bicycles from the bar chart; mode = tallest bar; median from cumulative frequencies; mean = Σ(bicycles $\times$ freq)/Σfreq. (Exact values depend on the bar heights shown.)

b) Bukky may prefer the mean because it uses every value and is useful for further calculation (e.g. total bicycles for planning).

c) Tariq may prefer the mode/median as they are not affected by a few households with unusually many bicycles, giving a more typical, whole-number value.

) i. The bar chart shows the exact frequency for every value and the shape of the distribution. ii. A box plot would show median, quartiles and outliers directly, aiding comparison of spread, but loses the individual frequency detail.

12 **a)** i. Marni is correct — a box plot must show a vertical line inside the box at the median (here, 2 items); Folake has omitted it. ii. Further error: the whisker runs to 6 even though only 1 customer bought 6 items; the outlier test (upper fence = $Q_3 + 1.5 \times \text{IQR} = 2 + 1.5(1) = 3.5$) shows 6 is an outlier, so the whisker should stop at the highest non-outlier value with 6 shown as an isolated point.

b) i. Check bars have equal width with gaps (discrete data), correct/labelled axes, and a linear scale starting at 0; an improvement is to add frequency labels above each bar. ii. The bar chart is likely more useful day-to-day (shows exact numbers for stocking decisions); the box plot is more useful for a quick statistical summary (e.g. comparing different Saturdays).

13 **a)** Hana ($n = 32$): min 68, $Q_1 = 69$, median 70, $Q_3 = 71$, max 72. Jess ($n = 32$): min 69, $Q_1 = 70$, median 70, $Q_3 = 71$, max 72.

b) A box plot is better for comparing typical time and consistency at a glance; a bar chart shows exact frequencies at each time — recommend box plots since the aim is comparison.

) Since the data is fairly symmetric with no extreme outliers, mean and SD (or equivalently median and IQR) are both suitable; Jess: mean ≈ 70.44 s, $\sigma \approx 0.93$ s (or median 70 s, IQR 1 s).

14 Every shop's "size 10" and "size 12" waist/hip measurements exceed the British Standard equivalent — e.g. mean size-10 waist across the 8 shops ≈ 28.4 in vs. BS 26.5 in, a difference of ≈ 1.9 in (≈ 4.9 cm). Several shops' "size 10" hip measurement already exceeds the BS "size 12" hip (37.0 cm/14.6 in). This evidence supports the claim of "vanity sizing" — clothes are cut more generously than the official size standard, consistently across shops.

15 Open class-survey task — credit sensible use of averages/spread and/or diagrams (chosen appropriately for skew/outliers) to summarise and compare estimates with the class.

Exercise 1D

1 Pumpkins ($n = 66$, cum. freq. 2, 7, 21, 45, 62, 66 at boundaries 5, 10, 15, 20, 25, 35). **a)** Modal class: $15 < m \leqslant 20$.

b) Plot cumulative frequency against upper class boundary.

c) i. median ≈ 17.5 kg. ii. $Q_1 \approx 13.4$ kg. iii. $Q_3 \approx 21.3$ kg. iv. IQR ≈ 7.9 kg. v. 30th percentile ≈ 14.6 kg. vi. 70th percentile ≈ 20.4 kg.

d) i. mean ≈ 17.27 kg. ii. $\sigma \approx 5.96$ kg.

2 Attendance ($n = 44$). **a)** Modal class: 600–800.

b) i. CF graph from cumulative freq. 1, 4, 15, 36, 42, 44, 44 at 200, 400, 600, 800, 1200, 1600, 2000. ii. median ≈ 667; IQR ≈ 244 ($Q_1 \approx 527$, $Q_3 \approx 771$). iii. Interpolating within 800–1200: estimated meetings $> 1000 \approx 5$ (11.4% of 44), which is $> 8\%$, so the graph confirms the organiser's claim.

c) i. mean ≈ 682. ii. $\sigma \approx 252$.

d) Compared with last season (mean 720, SD 240): this season's mean attendance is lower (down ≈ 38) and slightly more variable (SD up ≈ 12) — attendances have fallen a little and become marginally less consistent.

3 **a)** Read the median (at cumulative frequency $= n/2$) and quartiles (at $n/4$ and $3n/4$) from each curve; IQR $= Q_3 - Q_1$.

b) After coaching, both the median time and IQR should be smaller than before coaching (the "after" curve lies to the left of "before" throughout), showing the trainees became both faster on average and more consistent as a result of the coaching. (Exact figures depend on the printed graph.)

4 Nursery weights ($n = 49$, boundaries $12.95, 13.95, \ldots, 19.95$). **a)** mean ≈ 16.25 kg; $\sigma \approx 1.47$ kg.

b) i. CF diagram from cumulative freq. 3, 9, 21, 35, 42, 47, 49. ii. Our median (16.2 kg) matches the book's 50th percentile (16.2 kg) almost exactly; our Q_1 (≈ 15.2 kg) matches its 25th percentile (15.2 kg); our Q_3 (≈ 17.2 kg) is close to its 75th percentile (17.4 kg) — this sample is closely in line with the national figures.

5 Gaming ($n = 60$): modal class 30–59; mean ≈ 62.5 min, $\sigma \approx 28.1$ min, median ≈ 61.4, $Q_1 \approx 40.9$, $Q_3 \approx 82.9$, IQR ≈ 41.9. Reading ($n = 55$): modal class 60–89; mean ≈ 73.9 min, $\sigma \approx 28.6$ min, median ≈ 70.2, $Q_1 \approx 51.7$, $Q_3 \approx 86.7$, IQR ≈ 35.0. **a)** Modal classes as above.

b) Students spend on average ≈ 11 minutes longer reading than gaming (mean 73.9 vs 62.5; medians agree, 70.2 vs 61.4); spreads are similar, so this is a genuine general pattern, not driven by a few extreme values.

c) Reading the CF graphs gives similar median/IQR values, supporting the same conclusion across the whole distribution, not just the mean.

6 (Cumulative-% chart, video app ages.) **a)** The curve rises steadily; based on the labelled points (28.6%, 57.2%, 85.8%) the distribution is weighted towards younger users, with the curve flattening at older ages — a positively skewed age distribution.

b) i. Priya: a CF graph lets you read exact median/percentile ages and the precise shape of the distribution. ii. Connor: a box plot is quicker to interpret and clearly shows median, quartiles and outliers without reading values off a curve.

c) Neither diagram shows the sample size or how the sample was selected, both needed to judge reliability/representativeness.

7 Cycling event: women mean ≈ 3.75 hrs, $\sigma \approx 0.92$; men mean ≈ 3.18 hrs, $\sigma \approx 0.89$; women median ≈ 3.66, $Q_1 \approx 3.07$, $Q_3 \approx 4.31$, IQR ≈ 1.24; men median ≈ 3.07, $Q_1 \approx 2.54$, $Q_3 \approx 3.73$, IQR ≈ 1.19. **a)** i. CF graph for each group. ii. Men typically finished faster than women (median 3.07 vs 3.66 hrs), with a similar spread for both (IQR ≈ 1.2 hrs) — the whole distribution is shifted rather than differing in consistency.

b) i. Compare the median line and box/whisker lengths on the previous year's plots: identify which group has the lower (faster) median and which has the wider spread. ii. Compare this year's (CF graph) results with last year's (box plots): comment on whether the gap between men's and women's typical times, and the variability, has changed.

8 Percentile table (hours worked / rate of pay, male vs. female). **a)** i. Sami compares means (£13.25 vs £12.70, men ahead by 55p); Lola compares medians (£9.65 vs £9.12, women ahead by 53p) — both are correct for the statistic used. ii. Pay is positively skewed: a small number of very highly paid male jobs (90th percentile £25.85 vs £22.10) pull the male mean up much more than the male median, reversing the comparison.

b) IQR of pay: male = £14.25 − £7.50 = £6.75; female = £14.00 − £7.85 = £6.15 — male rates are very slightly more spread in the middle 50%.

c) Women's hours percentiles (e.g. median 18.6 hrs) are consistently a little higher than men's (median 17.1 hrs) almost throughout, and mean hours are higher for women (17.7 vs 16.8) — women in part-time work tend to work slightly more hours on average, though both groups show wide spread.

d) We are not given the true minimum/maximum (only 10th/90th percentiles), so the whiskers of a standard box plot cannot be drawn precisely from this table.

9 Marriage ages (assume "70 and over" has the same width, 10 years, as the previous class). Male 1990: mean ≈ 25.17, $\sigma \approx 6.02$, median ≈ 23.67, IQR ≈ 6.26. Male 2020: mean ≈ 32.21, $\sigma \approx 7.35$, median ≈ 30.94, IQR ≈ 10.00. Female 1990: mean ≈ 22.80, $\sigma \approx 5.52$, median ≈ 21.91, IQR ≈ 4.64. Female 2020: mean ≈ 30.15, $\sigma \approx 6.96$, median ≈ 28.56, IQR ≈ 9.35. **a)** Assumption: the open "70 and over" class is treated as if 10 years wide, matching the previous class.

b) Median/IQR values as above.

c) Between 1990 and 2020 the mean age at first marriage rose by ≈ 7 years for both men and women, and the spread (SD, IQR) also increased for both — people are marrying later on average and there is now much more variation in age at first marriage.

d) Use stratified sampling by age group/gender from marriage records; problems: contact details may be outdated, the topic is sensitive so response rates may be low, and timing the survey appropriately after the wedding is important.

10 Open research task — credit a genuine cited data source, appropriate measures/diagrams, and clear comparison of age distributions.

11 Sunshine: Brightside mean ≈ 230 hrs, $\sigma \approx 39.3$; Mistport mean ≈ 182 hrs, $\sigma \approx 31.4$. Rainfall: Brightside mean ≈ 53.0 mm, $\sigma \approx 32.6$; Mistport mean ≈ 62.6 mm, $\sigma \approx 29.0$. **a)** Brightside is sunnier on average (230 vs 182 hrs) with a similar spread to Mistport. Mistport is on average wetter (62.6 vs 53.0 mm), though Brightside's rainfall is more variable and more positively skewed (median 43.3 mm is well below its mean). Overall Brightside suits a sunshine-focused holiday; Mistport is more reliably wetter/overcast.

b) Open research task — credit appropriate comparison using real found data.

Exercise 1E

1 **a)** Frequency densities: $P < 20$: 3.2; $20 \le P < 40$: 4.4; $40 \le P < 80$: 3.55; $80 \le P < 120$: 1.775; $120 \le P < 200$: 0.2375; $P \ge 200$ (assumed 200–280): 0.075. Assumption: the last class is assumed to have the same width (80) as the previous class.

b) Draw a histogram with these frequency densities as bar heights over the stated (unequal) class boundaries, no gaps between bars.

2 **a)** Frequency densities: 0–4: 730; 5–15: 930; 16–44: ≈ 685.2; 45–64: 755.5 (boundaries 45–65); 65–74: 564 (65–75); 75–89: ≈ 274.7 (75–90); 90+: 32 (assumed 90–105, same width as previous class). Assumptions: ages treated as continuous, classes run on exactly from each other, and the open "90+" class assumed 15 years wide.

b) Histogram using these frequency densities over the given (unequal) boundaries.

3 Reading ages ($n = 500$). **a)** i. mean ≈ 17.17; $\sigma \approx 3.38$. ii. This year's mean (17.17) is very close to last year's (17.2), but this year's SD (3.38) is lower than last year's (3.6) — similarly centred but less spread out/more consistent this year.

b) i. Frequency densities: 10–12: 9; 12–14: 32; 14–16: 60.5; 16–18: 57.5; 18–20: 48.5; 20–24: 17.75; 24–30: ≈ 2.33. ii. Interpolating within 16–18: ≈ 240 students ($\approx 48\%$) have a reading age above 17.

4 Train delays. **a)** i. The histogram plots raw frequency (not frequency density) despite very unequal class widths, so bar areas do not represent frequency correctly and short-delay classes are hugely exaggerated; the x-axis also appears evenly spaced (0, 30, 60, ...) even though the true class widths (10.5, 10, 25, 75, 120, 120) are very different. ii. Using continuity-corrected boundaries $0, 10.5, 20.5, 45.5, 120.5, 240.5, 360.5$ (last class assumed width 120, matching the previous class), frequency densities are: Early–10: ≈ 937.1; 11–20: 198.0; 21–45: 49.6; 46–120: ≈ 10.1; 121–240: ≈ 1.21; > 240: 0.15.

b) i. mean ≈ 15.6 minutes; $\sigma \approx 27.1$ minutes. ii. These are estimates because the exact individual delay times within each class are unknown — the calculation assumes the data is spread evenly (uniformly) across each class (e.g. using midpoints), which is unlikely to be exactly true, especially for the very wide final classes.

5 Fire doors ($n = 389$, boundaries $695, 745, 845, 945, 1045, 1145$). **a)** Interpolating within 700–740: ≈ 11 doors ($\frac{725-695}{50} \times 19 \approx 11.4$) are narrower than the 725 mm guideline — these do not comply.

b) Interpolating within 750–840: doors < 800 mm $\approx 19 + \frac{800-745}{100} \times 168 \approx 111$, i.e. $\approx 28.6\%$ of the 389 doors would need widening.

6 Method: read the frequency density (patients per hour) for each interval from the histogram, multiply by class width to find the frequency in each class, and build a grouped frequency table. Use this to estimate the percentage of patients seen within 4 hours (compare to the 95% target) and the mean waiting time in minutes (compare to the national average of 110 minutes). State whether the clinic is meeting or falling short of each benchmark based on these estimates. (Exact percentages/mean depend on reading the printed histogram.)

Exercise 1F

1 Newborn weights ($n = 88$). **a)** Bea's CF graph: the class widths are unequal (first class is twice the width of the others) but her plotted values do not reflect a genuine running total against a true continuous weight scale plotted at upper class boundaries starting from $(1.5, 0)$ — check the values are non-decreasing and correctly positioned. Theo's histogram: he has plotted frequency rather than frequency density; since the first class has a different width, using frequency directly (rather than frequency ÷ width) makes the bar areas fail to represent the true frequencies; bars must be drawn with no gaps, aligned to the true class boundaries.

b) Correct cumulative frequencies (plotted at upper boundary): $(2.5, 0)$, $(3, 14)$, $(3.5, 47)$, $(4, 74)$, $(4.5, 84)$, $(5, 88)$. Correct frequency densities for the histogram: 1.5–2.5: 0; 2.5–3: 28; 3–3.5: 66; 3.5–4: 54; 4–4.5: 20; 4.5–5: 8.

2 Household size: mean household size TQ4 ≈ 2.02, TQ5 ≈ 2.24, TQ6 ≈ 2.22; all three areas have median household size 2 and are positively skewed (long tail to "8+"). Age: mean age TQ4 ≈ 43.4, TQ5 ≈ 42.6, TQ6 ≈ 48.0; TQ6's modal age band (60–75) is older than TQ4/TQ5's (45–60). Model conclusion: TQ6 has smaller households on average combined with an older population than TQ4/TQ5, suggesting TQ6 may be a retirement/older-resident area, while TQ4/TQ5 have

younger, larger households, consistent with more families with children. Support with comparative histograms of age/household size and the calculated means/medians above.

Consolidation Exercise 1

1 **a)** i. Random sample: every member of the population has an equal, known, non-zero chance of selection. ii. Representative sample: a sample whose characteristics reflect those of the whole population being studied.

b) i. Cluster sampling example: surveying all pupils in a small number of randomly chosen schools rather than pupils from every school. ii. Quota sampling example: a researcher stopping shoppers until they have interviewed set numbers of men/women in each age band.

c) Table totals: Year 1, 30; Year 2, 32; Year 3, 24; overall 86. i. Stratified sample of 24 (factor 24/86): Year 1 M ≈ 6, F ≈ 3; Year 2 M ≈ 7, F ≈ 2; Year 3 M ≈ 5, F ≈ 2 (values rounded to total 24; allow small rounding adjustment). ii. Use a random number generator applied to a numbered list within each stratum; or systematic sampling (e.g. every nth student on an alphabetical list) within each stratum.

2 Plant heights: Shade ($n = 25$): mode = 19 cm, range = $28 - 8 = 20$ cm, median = 17, $Q_1 = 13$, $Q_3 = 21.5$, IQR = 8.5, mean ≈ 17.16, $\sigma \approx 5.66$. Sunlight ($n = 15$): mode = 23 cm, range = $30 - 8 = 22$ cm, median = 23, $Q_1 = 14$, $Q_3 = 26$, IQR = 12, mean ≈ 20.33, $\sigma \approx 6.75$. **a)** i. Back-to-back stem-and-leaf (stems = tens digit). ii. Modes as above. iii. Ranges as above.

b) i. Box plots from the five-figure summaries above. ii. Sunlight-grown plants are taller on average than shade-grown plants (median 23 vs 17 cm; mean 20.3 vs 17.2 cm) and more variable (IQR 12 vs 8.5 cm), though both groups have the same minimum height (8 cm) — growing in sunlight is associated with greater height.

3 German marks ($n = 72$, cumulative freq. 0, 6, 20, 46, 65, 72 at boundaries $10.5, 20.5, \ldots, 60.5$). **a)** i. CF graph from the cumulative frequencies above. ii. 35th percentile ≈ 32.5 marks — about 35% of students scored 32.5 marks or below. iii. Interpreting the 35% pass mark as 21 out of 60, interpolation gives ≈ 51 students ($\approx 70\%$) estimated to have passed.

b) Box plot: min ≈ 11 (lowest non-empty class), $Q_1 \approx 29.1$, median ≈ 36.7, $Q_3 \approx 44.7$, max ≈ 60.

4 **a)** Marcus is only partly correct: the amount spent is continuous data, but as he collected it himself directly from students, it is primary (not secondary) data.

b) Frequency densities: $P < 15$: 0.733; $15 \leq P < 25$: 2.3; $25 \leq P < 35$: 3.1; $35 \leq P < 50$: 1.067; $50 \leq P < 80$: 0.3; $P \geq 80$ (assumed 80–110): 0.1.

c) mean $\approx$ £32.50; $\sigma \approx$ £18.98.

5 **a)** Improvements: clear axis labels (units, "Year", "Charge-points (1000s)"), a chart title, a consistent non-misleading scale (y-axis starting at 0, evenly spaced gridlines), avoiding distortion, and using the same scale on both charts for fair comparison.

b) The number of charge-points grew steadily/rapidly from 2015 to 2024, e.g. reaching around 48 000 (top gridline) by 2024, showing strong sustained growth in EV infrastructure. (Exact figures depend on reading the chart.)

6 **a)** Two reasons: (1) choosing only one school/college from each group means most schools have zero chance of selection, so the sample cannot capture variation between schools; (2) interviewing only 3 teachers per chosen school gives a very small, non-random sub-sample (9 teachers total) that is too small to represent the whole area.

b) Better method: stratified sampling — treat primary, secondary and college as three strata and randomly select a proportionate number of schools/colleges (and/or teachers) from each stratum using random numbers, with a much larger overall sample.

7 Documentary box office ($n = 461$). **a)** Modal class: $0 < T < 0.05$ (by far the largest frequency, 286).

b) i. mean $\approx$ £0.73m. ii. $\sigma \approx$ £2.38m.

c) Estimated median $\approx$ £0.04m, $Q_1 \approx$ £0.02m, $Q_3 \approx$ £0.336m. Problem: both the median and Q_1 fall inside the huge first class (which alone contains 286 of 461 films, $\approx 62\%$), so linear interpolation (assuming an even spread within the class) is unreliable, since box-office takings are very unlikely to be spread uniformly between £0 and £50 000.

d) Around 62% of documentaries earn less than £50 000 at the box office. A small number earn very large amounts (up to £25 million), creating a strongly positively-skewed distribution. This means the mean (£0.73m) is far higher than the median ($\approx$£0.04m) and is not representative of a "typical" documentary's earnings.

8 Cumulative %: < 4wks 3%; 4–8wks 22%; 9–13wks 51%; 14–15wks 84%; 16wks 95%; 17wks 99%; 18wks 100%. Comment: 95% of patients are seen within 16 weeks, meeting the target at exactly that threshold; however 5% wait 17 or 18 weeks, so the clinic is not meeting the standard for every patient. The median waiting time falls within the 14–15 week band, showing most patients wait close to (not comfortably within) the target, leaving little margin.

9 Simple average of the 5 regional means $= \dfrac{612 + 588 + 597 + 724 + 812}{5} =$ £666.60. Weighted overall mean (weighted by employee numbers) $\approx$ £691.31. These differ because the simple average treats each region equally regardless of workforce size, while the weighted mean correctly reflects that regions with more employees (Capital, South East) — which also pay more — have greater influence on the true overall figure. Two further questions (any sensible, e.g.): "Which region's mean is furthest from the overall weighted mean?"; "How does the range of regional means (£812−£588=£224) compare with the variation in employee numbers across regions?"

10 Open research task — credit genuine researched data, appropriate statistical measures/diagrams, and clear discussion of regional pay inequality.