Data, Sampling and Reliability

IB Mathematics: Analysis and Approaches SL/HL – SL 4.1

Populations, data types, sampling methods, reliability, data quality and outliers

Learning goal	Choose and evaluate sampling methods, identify possible bias and data-quality problems, classify data, and apply the $1.5 \times$ IQR outlier rule.
Syllabus link	IB Mathematics AA SL/HL: **SL 4.1**. Population, sample and random sample; discrete and continuous data; reliability and bias; missing/incorrect data; outliers; simple random, convenience, systematic, quota and stratified sampling.
Structure	60 questions: 15 short response; 15 extended response; 20 original IB-style / Paper 1 questions; 10 AA HL-exclusive challenge questions; full answer key and solution outlines.
Calculator / technology	Calculator not required. The worksheet is designed to be fully accessible as AA Paper 1-style practice; quartiles and IQR values are chosen for exact hand calculation.

Instructions

Questions 1–50 are shared AA SL/HL practice. Questions 51–60 form an **AA HL-exclusive challenge section** using only SL 4.1 content at greater depth. Answer in context whenever a question asks you to explain or evaluate. When identifying bias, say *who may be over-represented or under-represented and why.* For outliers, use the strict rule: values exactly on an outlier fence are **not** outliers.

Section 1: Short response — core knowledge and fluency

1. A school wants to estimate the average number of hours slept by all 840 students. It surveys 70 students chosen from the school register.

(a) State the **population**.

(b) State the **sample**.

(c) State what would be meant by a **census** in this context.

2. Classify each variable as **discrete** or **continuous**.

(a) Number of siblings.

(b) Time taken to travel to school.

(c) Number of goals scored in a match.

(d) Mass of a school bag.

3. Explain the difference between a **random sample** and a sample formed by asking the first available people nearby.

4. Name the sampling method used in each situation.

(a) A random-number generator selects 60 identification numbers from a complete list of employees.

(b) A researcher interviews people who happen to be leaving a shopping centre.

(c) After a random starting position, every 15th name on a register is selected.

(d) An interviewer is told to obtain 30 adults and 20 teenagers, and approaches available people until both targets are full.

(e) A school randomly selects students separately from Grades 11 and 12 in proportion to the size of each grade.

5. A news article reports that "82% of teenagers prefer studying online" but gives no further information. State **four** questions you should ask before deciding whether to trust this claim.

6. A student surveys only people sitting in the school cafeteria at 12:30 to estimate how often all students buy lunch at school. Identify one likely source of bias and explain why it matters.

7. A database contains 840 customer records. A systematic sample of 60 records is required.

(a) Find the sampling interval k.

(b) A random start of 9 is chosen. Write down the next four selected record numbers.

8. A school has 480 students: 240 in Grade 10, 150 in Grade 11 and 90 in Grade 12. A proportional stratified sample of 64 students is required. Determine how many students should be selected from each grade.

9. In quota sampling, explain why matching the population proportions does **not** make the individual selection random.

10. For a data set, $Q_1 = 18$ and $Q_3 = 30$.

(a) Calculate the IQR and the two outlier fences.

(b) Decide whether each of the values -1, 0, 48 and 49 is an outlier.

11. State one sensible action for each data-quality problem.

(a) A participant ID appears twice with identical responses.

(b) A person's age is recorded as -6.

(c) A response is missing.

(d) A travel time is much larger than all the others but is still possible.

12. Explain the difference between an **outlier that is a genuine observation** and an **outlier caused by a recording error**.

13. Describe how to take a simple random sample of 40 students from a complete numbered list of 600 students.

14. Match each description to the most appropriate term: **undercoverage**, **non-response bias**, or **voluntary-response bias**.

(a) Many people selected for a survey do not reply, and non-responders may differ from responders.

(b) Anyone who sees an online poll may choose whether to participate.

(c) A sampling frame excludes students who study remotely.

15. A researcher wants every year group in a school to be represented in the same proportions as in the school population. State the most appropriate sampling method and explain why.

Section 2: Extended response — applying and evaluating methods

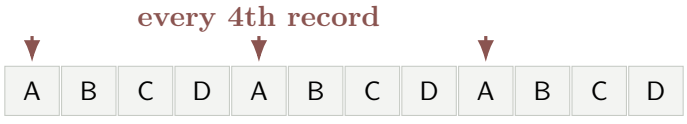
16. A city wants to estimate the mean cycling time of registered bicycle commuters. The city has a complete list of commuters and knows the district in which each person lives. Choose a suitable sampling method and justify your choice.

17. A factory list repeats the job roles **A, B, C, D** in that order throughout 400 employee records. A researcher needs a systematic sample of 100 employees.

(a) Calculate the systematic sampling interval.

(b) Explain why this interval can produce a seriously biased sample.

(c) Suggest one way to avoid this problem while still using the employee list.



18. A community event has attendees in three age groups: 50% adults, 30% teenagers and 20% children. An interviewer wants a quota sample of 40 attendees.

(a) Calculate the three quotas.

(b) Describe how the interviewer might fill these quotas in practice.

(c) Explain one reason the resulting sample may still be biased.

19. The following extract is taken from a survey database.

ID	Age	Study hours/week	Travel time (min)
A101	17	9	24
A102	16	11	31
A102	16	11	31
A103	−4	8	19
A104	17	–	27
A105	16	10	420

Identify **four** data-quality issues or values that should be investigated. For each one, state an appropriate action before analysis.

20. The following sorted waiting times, in minutes, were recorded at a clinic:

$$12,\ 14,\ 15,\ 16,\ 17,\ 18,\ 19,\ 20,\ 21,\ 22,\ 23,\ 40.$$

(a) Find Q_1 and Q_3.

(b) Calculate the IQR and the outlier fences.

(c) Identify any outlier(s).

21. In Question 20, suppose the value 40 minutes is confirmed by the clinic: the patient had to wait because an emergency case arrived. Explain why 40 should not automatically be deleted from the data set.

22. Two sources report the proportion of city residents who support a new tram line.

	Source A	Source B
Organisation	City transport agency	Anonymous social-media account
Date	October 2026	Date not stated
Sample	1200 residents selected randomly from the electoral register	25 000 voluntary online responses
Response	91% of selected residents responded	Anyone seeing the post could vote

Which source would you regard as more reliable for estimating the opinion of all city residents? Give **three** reasons. Explain why the much larger number of responses in Source B does not by itself make Source B more reliable.

23. A school wants to estimate the weekly study time of its Diploma Programme students. Design a trustworthy investigation. Your answer should state:

- the population and variable;
- a suitable sampling method;
- how the sample would actually be selected;
- two checks that should be made before analysing the data;
- one possible limitation that could remain.

24. A company has a complete list of 3000 customers. Compare **simple random sampling** and **systematic sampling** for selecting 150 customers. Give one advantage and one possible disadvantage of each method.

25. A college has 600 students distributed as shown.

Programme	Students	Sample of 80
A	240	
B	180	
C	120	
D	60	
Total	600	80

(a) Calculate the proportional stratified allocation for each programme.

(b) Explain why random selection is still required within each programme.

26. A music platform asks users visiting its homepage, "Do you agree that our excellent new recommendation system is better than the old one?" Users may choose whether to answer.

(a) Identify two different sources of bias.

(b) Rewrite the question in a more neutral way.

(c) Suggest a better way to select participants.

27. A school survey receives 420 responses from 600 selected students. Of the 180 students who did not respond, most are final-year students during examination week.

(a) Explain why this is a non-response problem rather than a missing-data problem within a completed response.

(b) Explain how the pattern of non-response could bias the results.

(c) Suggest one practical response by the researcher.

28. A health website publishes average daily step counts from data collected automatically by its own smartphone app. Evaluate the reliability of using these data to estimate the average daily step count of **all adults in the country**. Give at least three considerations.

29. A data set has $Q_1 = 42$ and $Q_3 = 50$.

(a) Calculate the lower and upper outlier fences.

(b) Decide whether 30, 31, 62 and 63 are outliers.

(c) Explain why identifying a value as an outlier does not prove that the value is incorrect.

30. A supermarket wants to estimate customer waiting times at its checkouts across an entire week. Explain why recording only the first 30 customers on Monday morning would be weak. Propose a sampling plan that better represents the week, and identify one remaining limitation.

Section 3: IB-style / challenge — original extended-response questions

31. School travel survey [8 marks]

A school has 720 students. The school wants a systematic sample of 60 students to estimate the mean time taken to travel to school.

(a) Calculate the sampling interval. [1 mark]

(b) A random starting position of 8 is chosen. Write down the next three selected positions. [1 mark]

The school later discovers that the register is arranged in repeating blocks of 12 students: the first six positions in every block are Grade 11 students and the next six are Grade 12 students.

(c) State which grade will be selected when the random start is 8, and explain your answer. [2 marks]

(d) Explain why the systematic sample is biased in this situation. [2 marks]

(e) Suggest a more reliable sampling plan for the two grades. [2 marks]

32. Sports-centre membership [8 marks]

A sports centre has 1800 members. Their age groups are shown below.

Age group	Number of members	Sample size
Under 30	900	
30–59	600	
60 or over	300	
Total	1800	90

The manager decides to take a proportional stratified sample of 90 members.

(a) Calculate the number of members that should be selected from each age group. [3 marks]

(b) Describe how the manager should select the required people within each age group. [2 marks]

Twelve of the selected members do not respond. Ten of these twelve are aged 60 or over.

(c) Explain why this non-response could bias the results. [3 marks]

33. Delivery times and outliers [9 marks]

A company records the following delivery times, in minutes, for 15 deliveries:

$$21,\ 22,\ 23,\ 23,\ 24,\ 24,\ 25,\ 25,\ 26,\ 26,\ 27,\ 28,\ 29,\ 30,\ 51.$$

(a) Find Q_1 and Q_3. [2 marks]

(b) Hence calculate the IQR and the lower and upper outlier fences. [3 marks]

(c) Identify any outlier. [1 mark]

The driver confirms that the 51-minute delivery was delayed by a road closure.

(d) Explain whether the company should automatically remove the value 51 before analysing delivery performance. [3 marks]

34. School-start survey [7 marks]

A student newspaper wants to estimate the proportion of all students who support a later school start time. Reporters stand outside the library during lunch and ask the first 100 students who are willing to respond.

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

(b) Give two reasons why the sample may be biased. [4 marks]

(c) Suggest one practical improvement that would make the sample more representative. [2 marks]

35. Quota versus stratified sampling [10 marks]

A town has 2000 residents in the following three age categories.

Category	Residents	Quota for a sample of 100
Adults	1000	
Teenagers	600	
Older residents	400	

An interviewer uses proportional quota sampling and interviews people at a shopping centre until every quota is full.

(a) Calculate the quota for each category. [3 marks]

(b) Explain why this is not a random sample even though the population proportions are matched. [2 marks]

(c) State one source of bias that could remain. [2 marks]

(d) Describe how the same three categories could instead be used to obtain a proportional **stratified random sample**. [3 marks]

36. Hospital appointment survey [6 marks]

A hospital wants to estimate the mean waiting time of all patients attending outpatient appointments during one month. A researcher samples every 20th patient on the appointment list after choosing a random starting position.

(a) Name the sampling method. [1 mark]

(b) State one condition under which this method is likely to give a representative sample. [2 marks]

(c) Describe one feature of the appointment list that could make the method biased. [2 marks]

(d) State one alternative sampling method. [1 mark]

37. Streaming-service survey [7 marks]

A streaming company reports that 93% of users "love" a new feature. The figure comes from a pop-up poll shown to users immediately after they use the feature; 7% of users who saw the pop-up responded.

(a) Identify one reason the response may be affected by voluntary-response bias. [2 marks]

(b) Explain why the low response rate is important. [2 marks]

(c) State three pieces of additional information that would help you judge the reliability of the claim. [3 marks]

38. Air-quality records [8 marks]

A sensor records hourly particulate concentration. An extract is shown.

Time	Concentration	Status
09:00	18.4	valid
10:00	19.1	valid
11:00	–	missing
12:00	18.9	valid
13:00	1890	valid
14:00	18.7	valid
14:00	18.7	valid

(a) Identify three entries or features that should be investigated before analysis. [3 marks]

(b) For each, state an appropriate action. [3 marks]

(c) Explain why the value 1890 should be checked rather than automatically deleted. [2 marks]

39. Stratifying by year group [7 marks]

A college has 500 students: 200 in Year 1, 150 in Year 2, 100 in Year 3 and 50 in Year 4. A proportional stratified sample of 75 students is required.

(a) Calculate the expected proportional allocation before rounding. [4 marks]

(b) Give one reasonable whole-number allocation that totals 75. [1 mark]

(c) State how the students should be selected within each year group. [2 marks]

40. Library satisfaction [6 marks]

A library randomly selects 300 members and emails them a satisfaction questionnaire. Only 102 reply.

(a) Calculate the response rate as a percentage. [1 mark]

(b) Explain why the original random selection does not guarantee that the final 102 responses are representative. [3 marks]

(c) Suggest one action that may reduce the problem. [2 marks]

41. Repeating weekly pattern [8 marks]

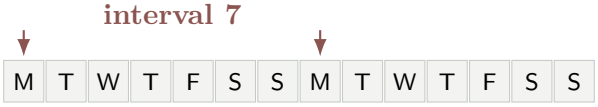
A restaurant stores one sales record for each day of the year in chronological order. A researcher selects every 7th record after a random start.

(a) Explain why the sample can contain records from only one day of the week. [2 marks]

(b) Explain why this may bias an estimate of mean daily sales. [2 marks]

(c) Suggest a sampling method that would ensure all seven weekdays are represented. [2 marks]

(d) Describe how to carry out your proposed method. [2 marks]



42. Salaries and an outlier [7 marks]

The annual salaries, in thousands of euros, of 12 employees are

$$24,\ 25,\ 25,\ 26,\ 26,\ 27,\ 27,\ 28,\ 29,\ 30,\ 31,\ 90.$$

(a) Find Q_1, Q_3 and the IQR. [3 marks]

(b) Use the $1.5 \times$ IQR rule to identify any outlier. [2 marks]

(c) Give one contextual reason why an analyst might keep the outlier in the data. [2 marks]

43. Fitness-app data [6 marks]

A fitness app records the number of exercise sessions per week and the duration of each session for its users.

(a) State which variable is discrete and which is continuous. [2 marks]

(b) Explain one reason why app users may not form a representative sample of the country's adult population. [2 marks]

(c) State two questions you would ask about the data source before using it in a report. [2 marks]

44. Census or sample? [6 marks]

A small club has 42 members and wants to know which of three meeting times members prefer.

(a) Give one advantage of conducting a census. [2 marks]

(b) Give one disadvantage of conducting a census. [2 marks]

(c) If only a sample is used, state one suitable method and briefly describe it. [2 marks]

45. Online election poll [7 marks]

A news website asks readers to click one of two buttons to indicate which candidate they support. It receives 80 000 responses.

(a) Explain why the poll is not a random sample of all voters. [2 marks]

(b) Give two reasons the result may be biased. [3 marks]

(c) Explain why the large sample size does not remove these problems. [2 marks]

46. Supermarket queue study [8 marks]

A supermarket wants to compare checkout waiting times on weekdays and weekends.

(a) Define a suitable population for the investigation. [1 mark]

(b) State a suitable variable and classify it as discrete or continuous. [2 marks]

(c) Propose a sampling method that represents both weekdays and weekends. [2 marks]

(d) Describe how the sample should be selected in practice. [2 marks]

(e) State one possible remaining source of bias or limitation. [1 mark]

47. School-bus routes [8 marks]

A school has 800 bus users on four routes: North 320, South 240, East 160 and West 80. A sample of 100 bus users is required.

(a) Calculate the proportional stratified sample size from each route. [4 marks]

(b) Explain how the school should select students from each route. [2 marks]

(c) State one reason why sampling only the North route would be inappropriate. [2 marks]

48. Data-quality audit [8 marks]

A data table contains the following records.

ID	Age	Height (cm)	Visits/month
P01	34	171	3
P02	41	168	4
P03	29	–	2
P04	37	1760	5
P04	37	1760	5
P05	310	182	3

(a) Identify four issues that should be investigated. [4 marks]

(b) For any two of these issues, state an appropriate action. [2 marks]

(c) Explain why data cleaning should be documented rather than performed silently. [2 marks]

49. Comparing two sampling plans [7 marks]

A cinema wants to survey customers about ticket prices. Plan A chooses 120 customers using a random-number generator from all online and counter purchases during one month. Plan B asks an interviewer to obtain quotas of 60 adults and 60 teenagers outside the cinema on Saturday evening.

(a) Name the sampling method in Plan A. [1 mark]

(b) Name the sampling method in Plan B. [1 mark]

(c) Give two reasons why Plan A is likely to be more representative of all monthly customers. [4 marks]

(d) State one situation in which Plan B might nevertheless be useful. [1 mark]

50. Planning a statistical investigation [10 marks]

A city wants to estimate the mean weekly number of hours that secondary-school students spend on homework. The city has a complete list of students including school and year group.

(a) State the population and the variable, and classify the variable as discrete or continuous. [3 marks]

(b) Propose a suitable sampling method and explain why it is appropriate. [2 marks]

(c) Describe how the sample should be selected. [2 marks]

(d) State two data-quality checks that should be made before analysis. [2 marks]

(e) State one limitation that may remain even after a careful sample is selected. [1 mark]

Section 4: AA HL exclusive challenge – deeper SL 4.1 reasoning

AA HL only. These questions use only the shared SL 4.1 syllabus, but require less scaffolding, stronger justification and more careful analysis of sampling structure and data quality.

51. Systematic sampling and periodic structure [7 marks]

A personnel list repeats the following pattern of departments throughout the file:

$$A, A, A,\ B, B, B,\ C, C, C,\ D, D, D.$$

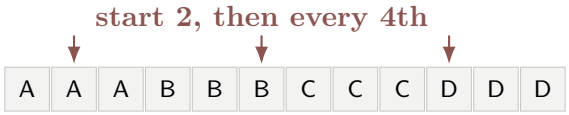
A systematic sample selects every 4th record. The random starting position is 2.

(a) Write down the departments of the first six selected employees. [2 marks]

(b) State which department is never selected. [1 mark]

(c) Explain why a random starting position does not remove the bias in this case. [2 marks]

(d) Suggest a sampling method that would guarantee representation from all four departments. [2 marks]



52. Response distortion after stratified sampling [8 marks]

A college has four year groups containing 400, 300, 200 and 100 students. A proportional stratified sample of 100 students is selected. The response rates in the four groups are respectively 75%, 80%, 90% and 100%.

(a) Determine the number initially selected from each year group. [2 marks]

(b) Determine the number of respondents from each year group and the total number of respondents. [3 marks]

(c) Explain why the final responding sample is no longer proportional to the college population. [2 marks]

(d) State one action the researcher could take before analysis. [1 mark]

53. Marginal quotas can hide imbalance [8 marks]

A population is 50% female and 50% male. It is also 30% teenagers and 70% adults. An interviewer is told to obtain a quota sample of 100 people with 50 females, 50 males, 30 teenagers and 70 adults.

(a) Construct one possible sample that satisfies all four quotas but has **all 30 teenagers female**. [3 marks]

(b) Explain why satisfying the separate gender and age quotas does not guarantee that the sample represents the joint age–gender structure of the population. [3 marks]

(c) State one improvement to the quota design. [2 marks]

54. Outlier fences in algebraic form [6 marks]

For a data set, $Q_1 = a$ and $Q_3 = a + 8$.

(a) Show that the upper outlier fence is $a + 20$. [2 marks]

(b) An observation has value $a + 20$. State whether it is an outlier and justify your answer. [2 marks]

(c) Another observation has value $a + 20.01$. State whether it is an outlier and explain what this illustrates about the $1.5 \times$ IQR rule. [2 marks]

55. Identical rows: duplicate or repeated measurement? [7 marks]

A researcher finds two completely identical rows in a data file.

(a) Explain why the researcher should not automatically delete one row. [2 marks]

(b) Give one situation in which the rows are likely to be a true duplicate and one situation in which they could be two valid observations. [3 marks]

(c) State two pieces of information that could help decide which situation applies. [2 marks]

56. Missing data and systematic bias [7 marks]

A survey of household spending is sent to a random sample of 1000 households. The income question is left blank much more often by high-income households than by other households.

(a) Explain why simply deleting every record with missing income could bias an analysis involving income. [3 marks]

(b) Explain why this is different from income values being missing completely at random. [2 marks]

(c) Suggest one practical step the researchers could take before deciding how to analyse the data. [2 marks]

57. Comparing sample size with sample quality [8 marks]

Study X selects 350 residents randomly from a complete register and obtains responses from 182 of them. Study Y obtains 5000 responses from a voluntary poll promoted on a campaign website.

(a) Calculate the response rate for Study X. [2 marks]

(b) Give two reasons why Study X may still be more useful for estimating the opinion of all residents, despite having far fewer responses. [4 marks]

(c) State one important weakness that Study X still has. [2 marks]

58. Sampling frame versus target population [7 marks]

A school wants to estimate the average journey time of all students. It selects a simple random sample from the school's transport-pass database. Students who walk or cycle do not appear in this database.

(a) Explain why the sample can be random but still biased. [3 marks]

(b) Name the problem with the sampling frame. [1 mark]

(c) Suggest a better sampling frame or sampling plan. [3 marks]

59. Choosing between two systematic intervals [8 marks]

A factory has 720 records arranged in repeating blocks of 12, where position within each block is related to work shift. A researcher needs a systematic sample of 60 records.

(a) Show that the natural sampling interval is 12. [1 mark]

(b) Explain why using this interval is risky. [3 marks]

(c) Another researcher suggests randomly choosing one record from each block of 12 instead. Explain why this avoids the particular periodicity problem. [2 marks]

(d) State one reason why proportional stratified sampling by work shift might be preferable to both plans. [2 marks]

60. Designing and defending an investigation [12 marks]

A university wants to estimate the mean number of hours per week that its 2400 students spend in paid employment. The university knows each student's faculty and year of study. The student register is sorted first by faculty, then by year.

(a) Propose a sampling method and explain why it is appropriate. [3 marks]

(b) Describe clearly how the sample should be selected in practice. [3 marks]

(c) Explain why simply taking every 20th name from the current register could be risky. [2 marks]

(d) State two data-quality checks that should be completed before calculating the mean. [2 marks]

(e) A very large reported value is confirmed as genuine. Explain how it should be treated and why. [2 marks]

Answer key and solution outlines

Section 1: Short response

1. (a) All 840 students. (b) The 70 surveyed students. (c) Surveying all 840 students.

2. (a) discrete; (b) continuous; (c) discrete; (d) continuous.

3. A random sample uses a chance process; convenience sampling depends on who is nearby/available, so inclusion probabilities are not controlled and some groups may be under-represented.

4. (a) simple random; (b) convenience; (c) systematic; (d) quota; (e) stratified.

5. Any four sensible checks: population, sample size, selection method, response rate, date, organisation/funder, question wording, handling of missing data.

6. Cafeteria users may buy lunch more often than other students, so frequent buyers can be over-represented and the estimate biased upward.

7. (a) $k = 840/60 = 14$. (b) $23, 37, 51, 65$.

8. Grade 10: 32; Grade 11: 20; Grade 12: 12.

9. Quotas fix only how many are required from each group; the interviewer can still choose convenient/available/willing people within groups.

10. IQR $= 12$; fences 0 and 48. Outliers: -1 and 49. Values exactly on fences are not outliers.

11. Duplicate: verify/remove duplicate if same response; impossible age: check source/correct if known or code missing; missing response: keep as missing/investigate; unusually large but possible value: verify before deciding whether genuine.

12. A genuine outlier is unusual but correctly recorded and should normally remain unless exclusion is justified. An error is a value produced by incorrect measurement/entry and should be corrected if possible or treated appropriately as missing.

13. Number students 1–600 and use a random-number generator (without replacement) to select 40 distinct numbers.

14. (a) non-response bias; (b) voluntary-response bias; (c) undercoverage.

15. Proportional stratified random sampling by year group: it preserves the population proportions while using random selection within each group.

Section 2: Extended response

16. Proportional stratified random sampling by district is suitable: allocate the sample in proportion to district size and randomly select within each district. A simple random sample is also defensible if district representation is not a concern.

17. (a) $k = 400/100 = 4$. (b) The list repeats every four records, so every fourth record selects the same role/position in the cycle. (c) Shuffle/randomise the list first, use simple random sampling, or stratify by role.

18. Adults 20, teenagers 12, children 8. Fill each quota by approaching attendees until the target is reached. Bias can remain because selection within each group depends on availability, willingness and interviewer choice.

19. Duplicate A102; impossible age -4; missing study-hours entry; travel time 420 should be checked as an unusual value. Appropriate actions are to verify against the source, remove a true duplicate, correct known errors, code unresolved entries as missing, and retain a genuine unusual observation.

20. $Q_1 = 15.5$, $Q_3 = 21.5$, IQR $= 6$; fences 6.5 and 30.5; 40 is an outlier.

21. The value is a confirmed genuine observation with a contextual explanation. It should not be deleted merely because it is unusual; summaries with and without it may be compared if influence is relevant.

22. Source A is more reliable: identifiable organisation, recent date, random selection and high response rate. Source B is self-selected/voluntary and may not represent all residents. A very large biased sample can still estimate the wrong population.

23. Example: population = all DP students; variable = weekly study time (continuous). Use proportional stratified random sampling by year group from complete lists; randomly select within each stratum. Check missing values, duplicates, impossible values and unusual outliers. Remaining limitations include self-reporting or non-response.

24. Simple random: every listed customer can be selected by chance and periodic order cannot distort selection; disadvantage: may require generating/sorting many random IDs and may not guarantee subgroup balance. Systematic: simple and quick after a random start; disadvantage: periodic structure/order in the list can create bias.

25. A: $80(240/600) = 32$; B: 24; C: 16; D: 8. Random selection within strata is needed so the researcher does not choose convenient or preferred individuals.

26. Biases: voluntary-response/self-selection and leading/loaded wording. Neutral wording: e.g. "Which recommendation system do you prefer: the new system, the old system, or no preference?" Better selection: a random sample of users who have experienced both systems.

27. Non-response means selected people fail to provide a completed response; missing data refers to absent values within records/responses. Final-year students are under-represented, so results can be biased if their behaviour/opinions differ. Follow up non-responders, extend the response window, or weight cautiously if justified.

28. App users may be younger/more health-conscious/wealthier than all adults; people without compatible smartphones are excluded; phone-carrying and permissions affect measurement; definitions and data-cleaning methods matter. The app data can be precise for its users but not representative of all adults.

29. IQR $= 8$; fences $42 - 12 = 30$ and $50 + 12 = 62$. Outliers are values strictly below 30 or above 62: 63 is an outlier; 30, 31 and 62 are not. Outlier status indicates unusualness, not incorrectness.

30. Monday morning customers may differ from customers at other times/days. A better plan samples across weekday/weekend and multiple time blocks, for example stratified by day/time and randomly or systematically selecting customers within blocks. A limitation could be very busy periods causing missed observations or customers leaving before measurement.

Section 3: IB-style / challenge

31. (a) $k = 12$. [1] (b) $20, 32, 44$. [1] (c) Start 8 lies in the Grade 12 half of each 12-person block, so every selected position is Grade 12. [2] (d) The interval matches the repeating structure, systematically excluding Grade 11. [2] (e) Use proportional stratified random sampling by grade. [2]

32. (a) 45, 30, 15. [3] (b) Use complete lists and random selection within each age stratum. [2] (c) Older members become strongly under-represented among respondents, so any age-related differences can bias results. [3]

33. $Q_1 = 23$, $Q_3 = 28$, IQR $= 5$; fences 15.5 and 35.5; 51 is an outlier. Since 51 is a confirmed real road-closure delay, it should not automatically be removed; exclusion depends on the analysis purpose and must be justified. [9]

34. Convenience sampling. Bias may arise because library users differ from others, only lunchtime attendees are available, and willingness to reply can be related to opinion. Improve with a random or stratified random sample from the full register. [7]

35. Quotas: 50 adults, 30 teenagers, 20 older residents. The counts are controlled but individual identities are chosen non-randomly at the shopping centre. Bias may come from location/time or interviewer selection. For stratified random sampling, split a complete resident list into the three age groups and randomly select 50/30/20. [10]

36. (a) Systematic. [1] (b) The appointment list should have no repeating structure related to waiting time and should cover the whole month/population. [2] (c) A periodic list pattern, for example clinics/specialties repeating every 20 records, could over-select one type. [2] (d) Simple random or stratified random. [1]

37. Strong opinions may make users more likely to respond, and only 7% replied, so respondents may differ greatly from non-respondents. Useful information includes sample size, exact question wording, date/time, who saw the pop-up, whether repeat responses were possible, and respondent characteristics. [7]

38. Investigate the missing 11:00 value, the extreme 1890, and duplicated 14:00 record. Check the source/sensor logs; keep missing as missing unless recoverable; verify units/decimal placement for 1890; verify whether the duplicate is a repeated record and remove one if appropriate. 1890 could represent a real pollution event, so it must be verified rather than deleted automatically. [8]

39. Expected allocations: Year 1 = 30, Year 2 = 22.5, Year 3 = 15, Year 4 = 7.5. One valid total-75 allocation is $30, 23, 15, 7$ (or $30, 22, 15, 8$ with a justified rounding rule). Randomly select the allocated number within each year group. [7]

40. Response rate $= 102/300 \times 100 = 34\%$. Random selection applies to the 300 invited members, but response

is not random; responders may differ from non-responders. Follow up non-responders, offer alternative response modes, or extend the response period. [6]

41. Every 7th chronological record has the same weekday because the calendar repeats every seven days. If weekday sales differ, the estimate is biased. Stratify by weekday and randomly sample within each weekday (or use a simple random sample across all days). [8]

42. Lower six give $Q_1 = (25 + 26)/2 = 25.5$; upper six give $Q_3 = (29 + 30)/2 = 29.5$; IQR $= 4$; fences 19.5 and 35.5; 90 is an outlier. It may be a genuine senior-management salary and therefore part of the population being described. [7]

43. Sessions/week is discrete; duration is continuous. App users are self-selected and may differ from all adults in fitness, age or smartphone access. Ask about selection/coverage, date, measurement method, missing-data handling, sample size, or who operates the app. [6]

44. Census advantage: all 42 members are included, avoiding sampling error and giving every member a voice. Disadvantage: more time/effort and non-response can still occur. A suitable sample could be simple random: number all 42 and randomly select the required number. [6]

45. Readers choose themselves and website readers are not all voters. Bias may come from self-selection, readership demographics, repeat voting or strength of opinion. Large size reduces random sampling variability but does not correct systematic selection bias. [7]

46. Population: checkout customers/transactions over the defined week(s). Waiting time is continuous. Stratify by weekday/weekend (and ideally time block), then randomly/systematically sample within each stratum using a plan that avoids periodicity. Remaining limitations include missed observations during busy periods or unusual events. [8]

47. North 40, South 30, East 20, West 10. Select randomly within each route from complete route lists. Sampling North only under-covers the other routes and may not represent different travel experiences. [8]

48. Issues: missing height for P03, height 1760 for P04, duplicate P04 row, age 310 for P05. Actions: verify/correct from source; remove true duplicate; treat unresolved errors as missing. Document cleaning so decisions are transparent, reproducible and do not silently alter the evidence. [8]

49. Plan A: simple random; Plan B: quota. Plan A covers purchases across a full month and uses chance selection; Plan B covers only Saturday evening shoppers and chooses individuals non-randomly within quotas. Plan B can be useful when rapid approximate feedback from specified categories is needed and no sampling frame is available. [7]

50. Population: all secondary-school students in the city; variable: weekly homework hours, continuous. Use proportional stratified random sampling, for example by school and year group, because known strata can be represented proportionally. Allocate numbers, then randomly select from each stratum. Check missing/duplicate/impossible values and investigate outliers. A limitation may be inaccurate self-reporting or non-response. [10]

Section 4: AA HL exclusive challenge

51. Selected positions are $2, 6, 10, 14, 18, 22, \ldots$, giving departments $A, B, D, A, B, D, \ldots$; department C is never selected. A random start changes the cycle but, because the interval 4 and list period 12 share a factor, only three positions in the 12-record pattern are visited. Stratified random sampling by department would guarantee representation. [7]

52. Initial allocation: $40, 30, 20, 10$. Respondents: $30, 24, 18, 10$, total 82. Their respondent proportions are therefore approximately 36.6%, 29.3%, 22.0%, 12.2%, not the original 40%, 30%, 20%, 10%. Differential non-response has distorted the intended stratification. Follow up non-responders, especially in the lowest-response groups, before analysis. [8]

53. One valid construction is: 30 female teenagers, 20 female adults, 0 male teenagers and 50 male adults. This meets 50 female, 50 male, 30 teenagers and 70 adults, but the joint age–gender composition can be badly distorted. Use joint quotas such as female teenagers, male teenagers, female adults and male adults based on population proportions, or preferably use stratified random sampling from a complete frame. [8]

54. IQR $= (a + 8) - a = 8$, so the upper fence is $(a + 8) + 1.5(8) = a + 20$. A value exactly equal to $a + 20$ is not an outlier because the definition uses *more than* 1.5IQR. The value $a + 20.01$ is above the fence and is an outlier. This shows that the rule uses a sharp threshold and does not by itself measure how substantively unusual or erroneous a value is. [6]

55. Identical rows can arise from accidental duplication, but they can also describe two distinct people or two distinct events with the same recorded values. A true duplicate could be the same participant imported twice; valid repeated observations could be two customers of the same age making identical purchases. Useful checks include unique IDs, timestamps, source-file row numbers, transaction IDs or collection logs. [7]

56. Deleting incomplete records would remove high-income households disproportionately, so the remaining data would under-represent that group and could bias relationships involving income. If missingness were completely random, missing records would not be systematically associated with income level. Researchers could follow up non-responders, investigate the pattern of missingness and compare characteristics of complete and incomplete records before deciding on treatment. [7]

57. Response rate for Study X is $182/350 \times 100 = 52\%$. Study X begins from a complete register and uses random selection, whereas Study Y is self-selected and tied to a campaign website; the 5000 responses therefore need not represent all residents. Study X still has substantial non-response, and responders may differ from non-responders. [8]

58. Randomness only applies within the transport-pass database. Students who walk or cycle have zero chance of selection, so the sampling frame undercovers the target population. This is undercoverage. Use the full student register as the sampling frame, then select randomly (or stratify by year/other relevant groups). [7]

59. $720/60 = 12$. Because the list itself repeats every 12 records and position is related to shift, the systematic sample can repeatedly select the same position/shift. Choosing one random record independently from each block breaks the fixed same-position pattern. Stratifying by work shift directly ensures the final sample represents shifts in chosen proportions and is therefore easier to defend. [8]

60. A strong plan is proportional stratified random sampling using faculty and year (or a justified combination of these strata). Allocate the target sample in proportion to stratum sizes and use random selection within each stratum from the full student register. Every 20th name is risky because the register is ordered by faculty/year, so periodic selection can interact with that structure and distort representation. Before analysis, check missing values/non-response, impossible entries, duplicates and unusual values. A confirmed extreme value is a genuine observation and should not be removed merely because it is an outlier; retain it unless the investigation purpose provides a clearly documented reason for exclusion, and consider reporting its influence. [12]