



H01-test Data Cleaning and Outliers (Higher): Test Pack

EDEXCEL 1ST0 · Calculator allowed · about 25 minutes

Total Marks

Name: _____ Date: ____ / ____ / ____

Answer ALL questions. Show all your working.

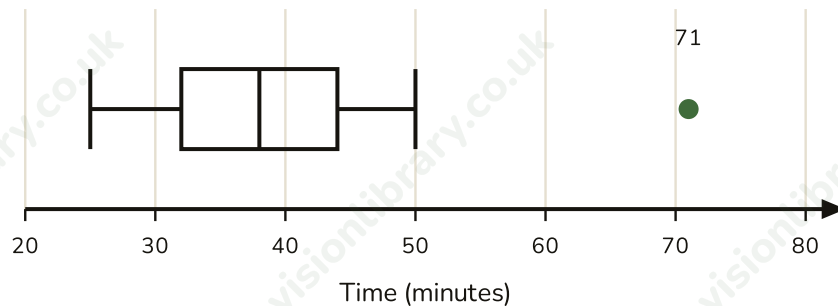
- 1** A cycling club logs the distance, in km, ridden by each of its 8 members on a Saturday club run. The office assistant's raw log is shown below.

Row	Rider	Distance recorded
1	Priya	42 km
2	Tom	38 km
3	Freya	45 km
4	Oscar	51 km
5	Freya	45 km
6	Nadia	29 miles
7	Callum	47 km
8	Isla	33 km

- (a) State which two row numbers are a duplicate entry for the same ride. (1)
- (b) Explain why recording row 6's distance in miles, when every other row is in km, could cause a problem when the club works out the total distance ridden. (1)
- (c) Given that 1 mile = 1.6 km, convert Nadia's distance to km, correct to 1 decimal place. (2)

.....
(Total for Question 1 is 4 marks)

- 2 The box plot shows the time, in minutes, taken by 40 runners to finish a fun run. One runner's time is marked separately with a dot, since it was not included when the box and whiskers were drawn.



- (a) Write down the median time and the interquartile range (IQR) shown on the box plot. (2)
- (b) One runner's time is marked separately at 71 minutes. Using the rule that a value is an outlier if it is above $Q3 + 1.5 \times \text{IQR}$, show that 71 minutes is an outlier. (2)
- (c) Give a possible genuine reason why this runner's time was so much higher than everyone else's, rather than assuming it must be a data-entry error. (1)

(Total for Question 2 is 5 marks)

3 Aaron is investigating how many hours Year 11 students spend revising per week. He designs a questionnaire, distributes it to a random sample of 30 students, and enters their responses into a spreadsheet. (The statistical enquiry cycle has four stages: plan, collect, process, interpret.)

(a) Aaron notices that two students have left the revising-hours question blank. At which stage of the statistical enquiry cycle should Aaron decide how to deal with these missing responses? (1)

(b) One response reads '20 hours', much higher than every other response (all between 2 and 12 hours). Aaron calculates that $Q1 = 4$ hours and $Q3 = 9$ hours for the sample. Use the rule $Q3 + 1.5 \times IQR$ to decide whether 20 hours should be treated as an outlier. (3)

(c) Aaron decides to leave the two blank responses out of his analysis rather than guessing a value for them, but to investigate the '20 hours' response further rather than deleting it immediately. Explain why both of these are reasonable decisions. (2)

(Total for Question 3 is 6 marks)

- 4 A car park attendant's spreadsheet automatically logs the length of stay, in minutes, of vehicles using the car park. While checking the log for one day, she discovers that 6 vehicles which left during the $60 < t \leq 90$ minute class were also mistakenly logged a second time in the $90 < t \leq 120$ minute class, because of a fault in the automatic timer. The table below shows the frequencies as they were first recorded (before this fault was found).

Length of stay, t (min)	$0 < t \leq 30$	$30 < t \leq 60$	$60 < t \leq 90$	$90 < t \leq 120$
Frequency (as recorded)	8	14	23	21

- (a) Explain why the total of the frequencies in the table (66) does not match the 60 vehicles that actually used the car park that day. (1)
- (b) Write down the corrected frequency for the $90 < t \leq 120$ class. (1)
-
- (c) Using mid-interval values, work out an estimate for the mean length of stay using the corrected frequencies (8, 14, 23, 15). (3)
-
- (d) Given that $\sum fx^2 = 324900$ for the corrected data, work out an estimate for the standard deviation of the corrected data. Give your answer correct to 1 decimal place. (2)
-
- (e) Using the UNCORRECTED (originally recorded) frequencies, the estimated mean is 70.9 minutes and the estimated standard deviation is 29.9 minutes (both 1 dp). Comment on the effect that the duplicate-logging error had on these estimates. (1)

(Total for Question 4 is 8 marks)

Mark scheme · H01-test Data Cleaning and Outliers (Higher): Test Pack

Question 1

- (a) **B1** 3 and 5 cao
- (a) **Answer: Rows 3 and 5**
- (b) **B1** explains that mixing units would make the total/sum of distances meaningless unless all values are converted to the same unit first oe
- (b) **Answer: If the 29 is added to the other distances without converting it, the total distance ridden would be wrong, because 29 miles is not the same as 29 km.**
- (c) **M1** 29×1.6 , oe
- (c) **A1** 46.4 km cao
- (c) **Answer: 46.4 km**

Question 2

- (a) **B1** median = 38 cao
- (a) **B1** $IQR = 44 - 32 = 12$ cao
- (a) **Answer: Median = 38 minutes, IQR = 12 minutes**
- (b) **M1** $44 + 1.5 \times 12$, oe
- (b) **A1** 62, and $71 > 62$ so it is an outlier, cao
- (b) **Answer: 62 minutes; since $71 > 62$, it is an outlier.**
- (c) **B1** e.g. the runner may have stopped to help an injured runner, walked most of the course, or been injured themselves, oe
- (c) **Answer: The runner may have stopped to help someone else, been injured, or walked most of the course, so 71 minutes could be a genuine time.**

Question 3

- (a) **B1** process cao
- (a) **Answer: The process stage.**
- (b) **M1** $IQR = 9 - 4 = 5$, oe
- (b) **M1** $9 + 1.5 \times 5$, oe
- (b) **A1** 16.5, and $20 > 16.5$ so yes it is an outlier, cao
- (b) **Answer: Yes; the boundary is 16.5 hours, and 20 hours is greater than this.**
- (c) **B1** blank responses contain no genuine information, so guessing a value could introduce bias/inaccuracy that a real value would not have oe
- (c) **B1** an unusually high but not impossible value like 20 hours could be genuine, e.g. a student revising intensively before exams, so deleting it straight away could lose real, useful information oe
- (c) **Answer: Blank responses have no genuine data to guess from, so estimating them risks bias. 20 hours could be a genuine value (e.g. a student cramming before exams), so it should be checked rather than deleted straight away.**

Question 4

- (a) **B1** the 6 vehicles that left during the 60-90 min class were logged twice - once correctly and once in error in the 90-120 min class - so these 6 extra rows do not represent extra vehicles oe
- (a) **Answer: The 6 vehicles that left in the 60-90 minute class were logged a second time by mistake in the 90-120 minute class, so those 6 extra entries do not represent extra vehicles.**
- (b) **B1** 15 cao (21 - 6)
- (b) **Answer: 15.**
- (c) **M1** mid-values 15, 45, 75, 105 used, oe
- (c) **M1** sum of $f \times \text{mid-value} = 4050$, oe
- (c) **A1** mean = 67.5 minutes cao
- (c) **Answer: 67.5 minutes.**
- (d) **M1** $324900/60 - 67.5^2$, oe
- (d) **A1** 29.3 (awrt 29.3) cao
- (d) **Answer: 29.3 minutes (awrt).**
- (e) **B1** the error increased both the estimated mean (by about 3.4 minutes) and the estimated standard deviation (by about 0.6 minutes), because it wrongly added extra vehicles into a higher time class oe
- (e) **Answer: The error increased the estimated mean (by about 3.4 minutes) and the estimated standard deviation (by about 0.6 minutes), since it wrongly shifted 6 extra 'vehicles' into a higher time class than they really belonged to.**