



H01-10min Data Cleaning and Outliers (Higher): 10-minute Pack

EDEXCEL 1ST0 · Calculator allowed · about 10 minutes

Total Marks

Name: _____ Date: ____ / ____ / ____

Answer ALL questions. Show all your working.

- 1** A small museum records the number of visitors on each of 11 consecutive days.

Day	1	2	3	4	5	6	7	8	9	10	11
Visitors	84	79	91	88	76	95	810	82	89	77	85

- (a) Write down the value in the list that is most likely to be a data-entry error. (1)
-
- (b) Suggest what the correct number of visitors was likely to be on day 7, and explain your reasoning. (1)
- (c) Work out the range of the visitor numbers exactly as recorded, including the value 810. (2)
-
- (d) The value is corrected to 81. Work out the range of the corrected data set. (2)
-
- (e) Explain the effect that this data-entry error had on the range. (1)

(Total for Question 1 is 7 marks)

2 For each statement about data cleaning and outliers, write down whether it is True or False.

(i) The rule $Q1 - 1.5 \times IQR$ is used to find an upper boundary for identifying outliers. **(1)**

.....

(ii) A single very large outlier typically increases the standard deviation of a data set. **(1)**

.....

(iii) If a missing value is imputed using the mean of the remaining data, the overall mean of the completed data set stays the same. **(1)**

.....

(iv) Deleting a data point simply because it looks unusual, without first checking whether it is a genuine value, is good statistical practice. **(1)**

.....

(Total for Question 2 is 4 marks)

Mark scheme · H01-10min Data Cleaning and Outliers (Higher): 10-minute Pack

Question 1

- (a) **B1** 810 cao
- (a) **Answer: 810**
- (b) **B1** 81 (an extra 0 was probably typed after 81), oe reasoning that links to the range of the other values
- (b) **Answer: 81 visitors (an extra 0 was probably typed by mistake).**
- (c) **M1** 810 - 76, oe
- (c) **A1** 734 cao
- (c) **Answer: 734**
- (d) **M1** 95 - 76, oe
- (d) **A1** 19 cao
- (d) **Answer: 19**
- (e) **B1** the error made the range far larger than it should be (734 instead of 19), because the range only uses the two most extreme values, so one incorrect extreme value distorts it heavily oe
- (e) **Answer: The error inflated the range hugely (734 instead of the true 19), because the range depends only on the biggest and smallest values, so a single wrong extreme value badly distorts it.**

Question 2

- (i) **B1** False cao
- (i) **Answer: False**
- (ii) **B1** True cao
- (ii) **Answer: True**
- (iii) **B1** True cao
- (iii) **Answer: True**
- (iv) **B1** False cao
- (iv) **Answer: False**