



Cleaning data and outliers

About 50 minutes Calculator allowed 45 marks

Date _____

A Do Now

The first question is what today's lesson needs. The others bring back earlier topics.

A1 Needed today [2 marks] **A2** Last lesson [2 marks]

The mean of x , $x + 7$, $2x$ and $3x - 6$ is 16. Find x .

Answer

The probability that a bus is full is $\frac{5}{9}$. Owen uses two-digit random numbers: 00 to 49 for full and 50 to 89 for not full, ignoring 90 to 99. Use these numbers to estimate the probability that a bus is full: 71, 38, 42, 21, 70, 40, 24, 97.

Answer

A3 From chapter 3 [2 marks] **A4** Maths skill [2 marks]

In a trial of a new sleep spray, 14 of the 40 people who used it slept for more than 7 hours. In the control group, 78 of the 120 people slept for more than 7 hours. Work out both percentages. Does the sleep spray appear to work?

Answer

Write $\frac{5}{6} : \frac{2}{3}$ as a ratio of whole numbers in its simplest form.

Answer

B Example, then your turn

Your teacher works through each example with you. Then try the one beside it.

EXAMPLE 1

Six students' heights were recorded in centimetres:

162, 1.58, 171, 0, 1650, 164

Identify the values that need cleaning and say what to do with each.

Hint Look for a different unit, an impossible value and a typing error.

B1 Seven students' times, in minutes, to walk to school: [2 marks]

14, 22, 9, 0.25, 17, 140, 19

Identify two values that need cleaning and say what to do with each.

B2 The times, in minutes, of 15 deliveries: [3 marks]

8, 22, 25, 27, 28, 30, 31, 32, 33,

35, 36, 38, 40, 44, 61

Using the $1.5 \times \text{IQR}$ rule, find any outliers.

Answer

EXAMPLE 2

The masses of a farm's apples have mean 152 g and standard deviation 11 g. Use the rule: an outlier lies outside mean $- 3 \times \text{sd}$ to mean $+ 3 \times \text{sd}$. Which of 118 g, 125 g and 190 g are outliers?

Hint Work out the mean minus 3 standard deviations, and the mean plus 3 standard deviations.

B3 Runners' times for 400 m have [3 marks]

mean 68.4 s and standard deviation 3.2 s. Using the mean ± 3 standard deviations rule, which of 57.9 s, 77.6 s and 79.5 s are outliers?

Answer

EXAMPLE 3

Reaction times have a mean of 0.25 s. Two outliers are 0.02 s and 0.71 s. Say whether each should be removed.

Hint Could each time really happen?

B4 In a survey of Year 11 students, [2 marks]

one age is recorded as 4 years and one height as 198 cm. Both are outliers. Say whether each should be removed, with a reason.

C Practice

C1 Six hand spans, in cm: 18, 21, 19, 210, 20, 22. The 210 is a recording error. Work out the mean of the other values. [2 marks]

Answer

C2 The masses of five parcels were recorded as 2.4 kg, 1800 g, 3.1 kg, 950 g and 2.0 kg. Clean the data and work out the mean mass in kilograms. [2 marks]

Answer

C3 For a set of test marks, the lower quartile is 34 and the upper quartile is 46. Using the $1.5 \times \text{IQR}$ rule, show that a mark of 66 is an outlier. [2 marks]

C4 Give two reasons why data should be cleaned before it is analysed. [2 marks]

C5 Exam marks have mean 54 and standard deviation 12. Using the mean ± 3 standard deviations rule, is a mark of 15 an outlier? Show your working. [2 marks]

C6 In a survey of shoe sizes, one entry is left blank and one is written as "seven". Say what should be done with each before the data are analysed. [2 marks]

D Exam-style questions

D1 Jo says, "An outlier is always a mistake, so it should always be removed." [2 marks]
Is Jo correct? Give a reason for your answer.

- D2** Rafi measures the lengths of 40 leaves in centimetres. One length is recorded as 61 cm. For the other 39 lengths the mean is 6.2 cm and the standard deviation is 0.9 cm; they include 2.1 cm, 3.9 cm, 8.4 cm and 9.1 cm. [6 marks]

Rafi needs to clean the data before he analyses it. Using the mean $\pm$ 3 standard deviations rule, explain how he should clean the data. You must show your working.

- D3** The number of visitors to a museum on 11 days:

212, 230, 236, 241, 245, 250, 254, 260, 263, 270, 412

- (a) Using the $1.5 \times \text{IQR}$ rule, show that 412 is an outlier. [3 marks]

- (b) The 412 visitors came on a bank holiday. The museum is planning staff for ordinary days. Should the 412 be removed? Give a reason. [1 mark]

Total for D3: 4 marks

E Extension

No route is given. The hints are at the end of the sheet. Use one at a time.

- E1** Here are 11 values in order: 23, 27, 28, 31, 33, 34, 36, 38, 40, 44, x , where x is a whole number greater than 44. A value is an outlier if it is more than $1.5 \times \text{IQR}$ above the upper quartile. Find the greatest value of x for which the data have no outliers. [3 marks]

Answer

Extension hints

Cover these with paper. Uncover one at a time.

- E1

Hint 1

Does x change the quartiles? It is the largest value, the 11th.
- Hint 2

Q_1 is the 3rd value, 28, and Q_3 is the 9th value, 40, so the IQR is 12.
- Hint 3

The upper limit is $40 + 18 = 58$. A value of 58 is not more than 58, so the greatest x is 58.

How did it go?

Colour one circle on each line: not yet, nearly, yes.

- I can spot errors in raw data and say what to do with each.
- I can find outliers with $1.5 \times \text{IQR}$ or with mean ± 3 standard deviations.
- I can decide whether a value should be removed, and say why.