OCR

A Level

A Level Mathematics

Outliers and Cleaning Data (Answers)

Name:

M M E

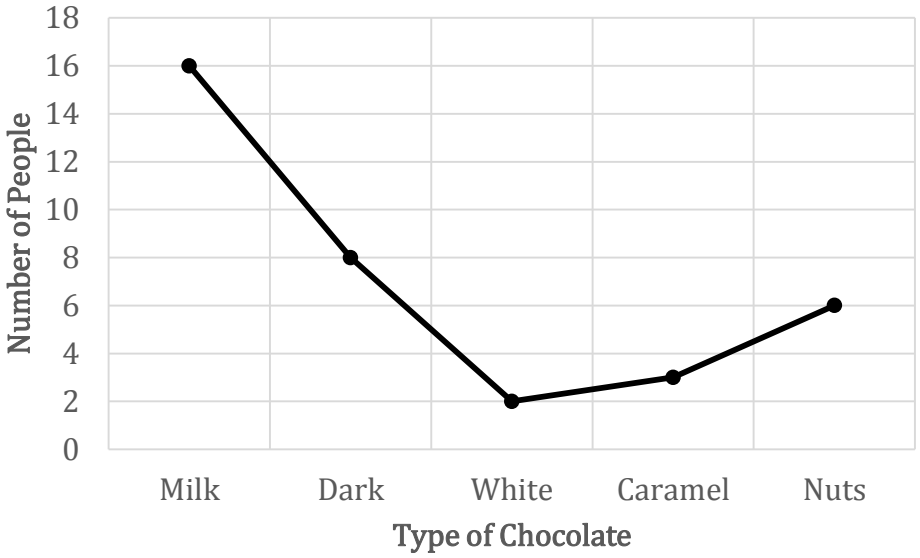
Mathsmadeeasy.co.uk

Total Marks:

L4- Outliers and Cleaning Data- Answers

OCR

1) **Kevin has the results of 35 people's favourite chocolate. He has represented it in a graph and calculated the mode as milk, the median as dark, the range as 14, and the mean as 7.**



Critique the following:

i) Kevin's choice of graph.

[1 mark]

Line graphs are suitable for indicating trends and representing continuous data. This means that for a continuous independent variable, the dependent variable can be predicted for unspecified points. Here, it would be incorrect to conclude that if we mixed dark and white chocolate together the number of people who would choose it as their favourite chocolate would be five.

ii) His summary statistics.

[1 mark]

Milk chocolate is the modal favourite chocolate.

[1 mark]

The median chocolate is not Dark. We do not know what the middle person likes as we have no method of ordering them. We cannot have a median of quantitative data.

[1 mark]

The mean cannot be calculated for favourite type. We cannot say the average favourite chocolate was x. But we can say, the useless statement, that the mean frequency is 7.

[1 mark]

Similarly to the mean, we cannot say the range is x – y. However, the range in the size of groups is 14.

iii) **His decision to remove *Fruit* because he considered it to be outlier.**

[1 mark]

Fruit type cannot be removed. The only way Fruit would be an outlier would be a) it is by far the most popular or b) it contains a negative number. If b) is correct, the survey/ recording of data is wrong. If a) is correct, this is the results of the survey and must be shown. This would be critical information for a chocolate manufacturer. We have no evidence that an error so great has occurred that it accounts for a high value.

2) **The number of spots on insects at a nature park was recorded for one day in July. The**

summary table of this experiment is shown in the grouped frequency table.

Number of Spots (s)	$0 < s \leq 2$	$2 < s \leq 4$	$4 < s \leq 6$	$6 < s \leq 8$	$8 < s \leq 10$
Frequency	**3**	**8**	**13**	**14**	**6**

i) Calculate the mean and range number of spots.

[1 mark]

The range in the number of spots is $10 - 0 = 10$.

[1 mark]

The modal number of spots is in the interval $4 < s \leq 6$.

[1 mark]

For the mean, we need to estimate the midpoint for each group.

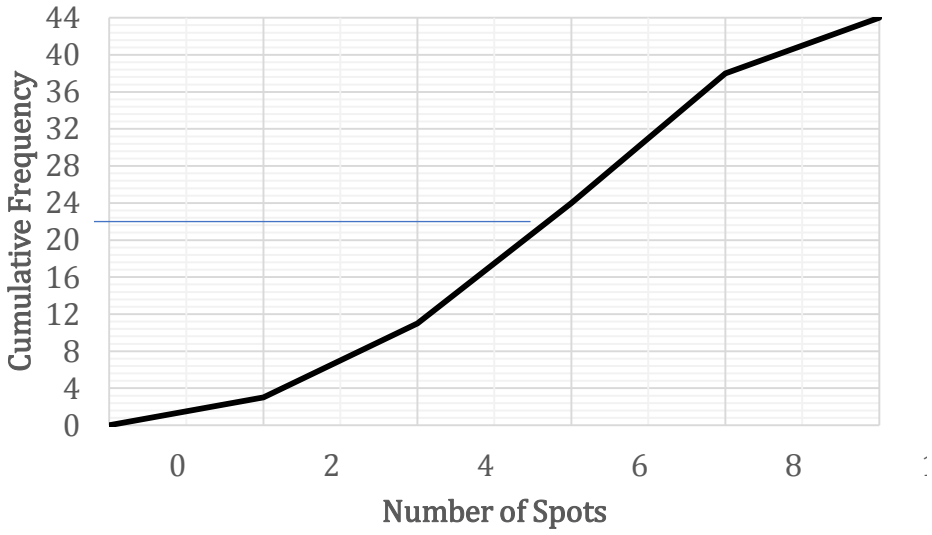
Midpoint of interval (m)	1	3	5	7	9
Frequency (f)	3	8	13	14	6
mf	3	24	65	98	54

[1 mark for correct formula and execution]

$$\mu = \frac{\Sigma mf}{\Sigma f} = \frac{244}{44} = 5.55\ (to\ 2dp)$$

ii) Draw an appropriate chart for the data, and use this chart to obtain an estimate for the median.

[1 mark for cumulative frequency correct, 1 mark for ends of intervals- 2 max]



[1 mark for line drawn in a 22.5th point, 1 mark for correct answer of spots – 2max]

3) Salaries of 30 people at a company are normally distributed, with a mean of 25,000 and

standard deviation of 2000. For each of the following scenarios state what will happen if:

i) One person, with a salary 10^2 larger than the mean is added.

[1 mark]

The mean will increase, the median will stay the same.

[1 mark]

The data will become positively skewed.

ii) One person, with a salary of 10^0 larger than the mean is added.

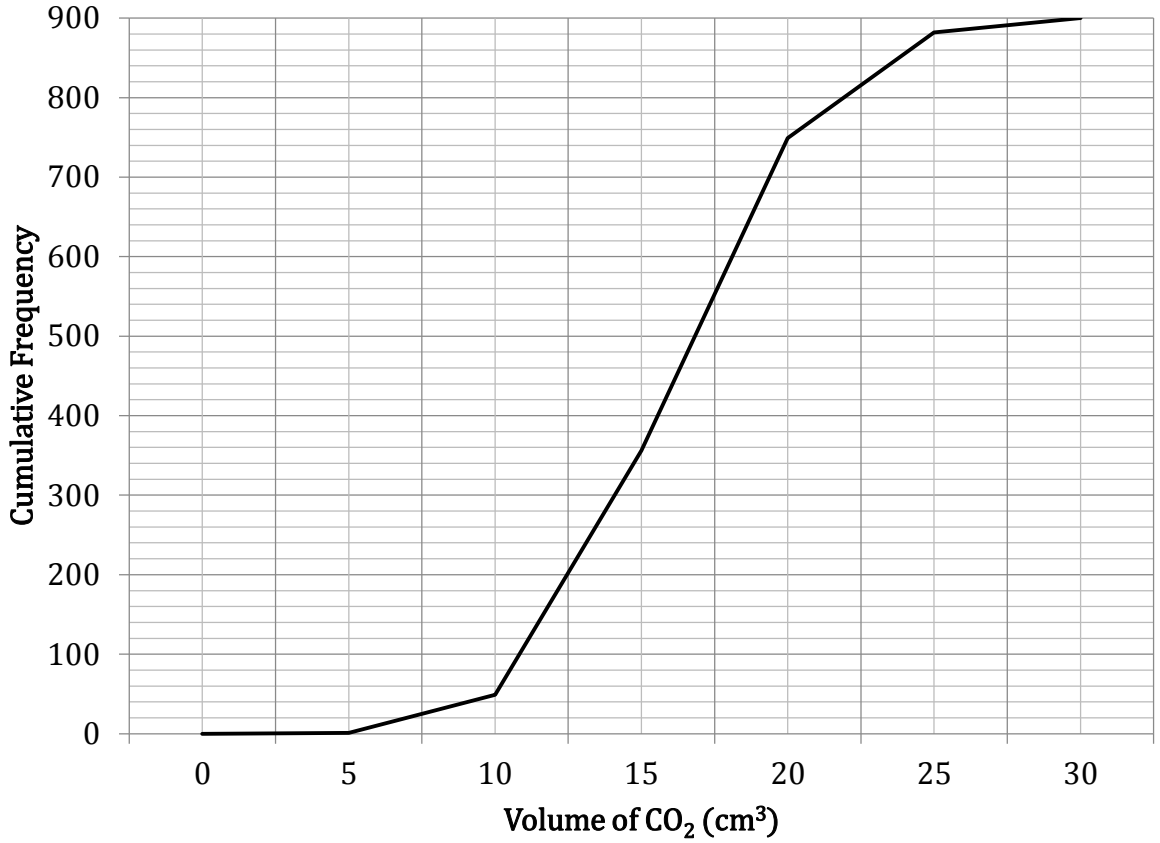
[1 mark]

Spotting $10^0 = 1$.

[1 mark]

The mean, standard deviation and median will stay the same.

4) The volume of $CO2$ a machine produces was measured 900 times in June. These volumes are displayed in the cumulative frequency graph below.



i) Estimate the median and interquartile range of the data [2]

ii) Identify the outliers, if there are any. [4]

iii) Discuss the whether these outliers should be removed. [1]

iv) On the next six occasions, the machine reported negative values. Discuss how adding these into the data would alter the graph. [1]