Section Check In – 2.02 Data Presentation and Interpretation

Questions

1 A car showroom has a number of secondhand cars for sale. On a particular day the prices are recorded below. The prices are in hundreds of pounds.

35	92	113	45	30	120	98	77	68	84
65	72	53	45	99	89	78	79	30	76

(i) Calculate the mean and standard deviation for these data.

(ii) Determine whether there are any outliers.

2 The stem and leaf diagram below shows the average distance a person walks or runs or cycles each day for a random sample of 70 people.

$N = 70$
$1 \mid 3$ represents 13 km

0	0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 6 7 7 8 8 8 8 9 9
1	0 0 0 0 1 1 1 1 2 2 3 3 4 4 5 5 6 6 6 7 7 7 7 8 8 8
2	0 1 1 1 3 3 5 5
3	2 6

Find the median, the upper quartile and the lower quartile.

3 A group of 31 students undertook a mathematics project at home one night. The next day they were asked how long it took them. The results are shown in the following table, where the times given were to the nearest 10 minutes.

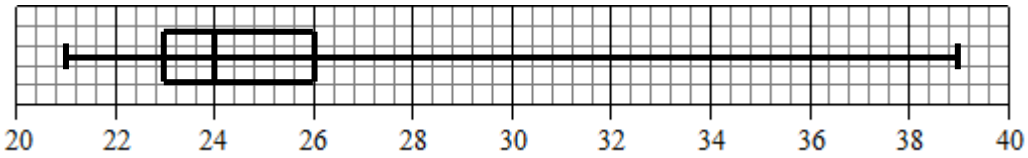
Time (to the nearest 10 minutes)	10	20	30	40	50	60	70	80	90
Frequency	5	8	8	5	3	1	0	0	1

(i) When considering average and spread for this set of data, explain whether you consider that the value 90 should be discounted, giving reasons.

(ii) Discounting this value, find the median and the quartiles of the data set.

(iii) What difference would it make to the median and quartiles of the data set if the value of 90 was included?

4 A teacher records the time she takes to drive to school each morning for a term. These data are illustrated in the box and whisker plot shown below.



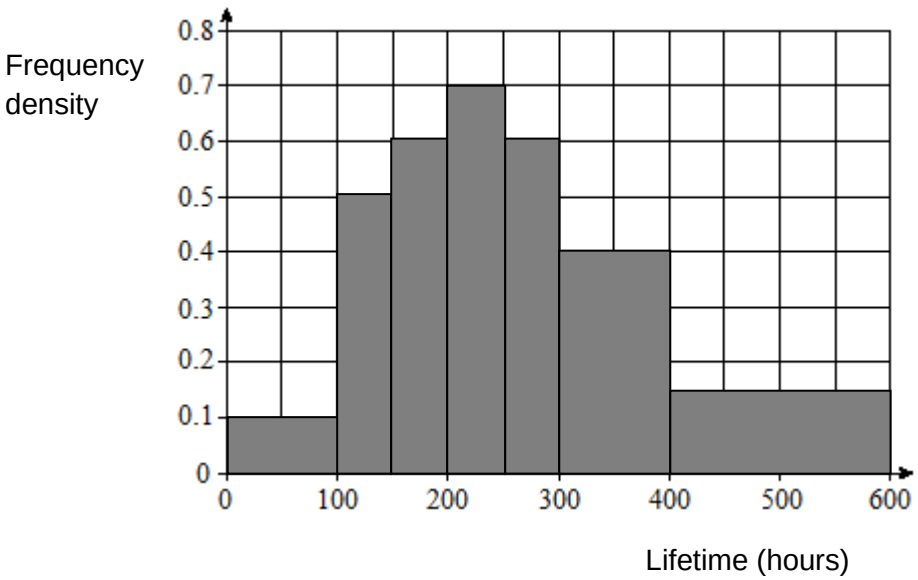
(i) Find the interquartile range and the median for these data.

(ii) Show that there are no outliers at the lower end but that there is at least one at the upper end. Explain why this might be so.

You are now given that the longest two times that she took to drive to school were 28 minutes and 39 minutes.

(iii) Explain what effect removing the 39 minutes from the set of data will have on the median and interquartile range.

5 The lifetimes in hours of 200 of a particular light bulb is shown in the histogram below.



(i) Find the number of bulbs in each class.

(ii) Estimate the median.

6 Paul has investigated data relating to the sales of ice cream and the incidence of sunburn at a particular seaside town.

He draws the following conclusion:

"As ice cream sales increase so does the incidence of sunburn. Therefore the more ice cream you eat the more likely you are to get sunburnt".

Comment on this statement.

7 John and Andrew are asked to carry out a survey to find the average number of goals scored in two leagues. They each choose 10 games at random and record the total goals scored in that match.
These data are summarised in the table below.

Number of goals	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**$\geqslant$8**
John's frequency	0	2	6	2	0	0	0	0	0
Andrew's frequency	2	3	2	2	0	0	0	1	0

(i) Show that the mean for John's data is 2.

(ii) The mean for Andrew's data is also 2. Explain why the mean is not the most appropriate average to use.

(iii) Find the median value for Andrew's data.

8 A group of 30 students measured the length of their little finger in centimetres. The results are summarised in the table below.

Length (cm)	$4.0 < x \leq 4.5$	$4.5 < x \leq 5.0$	$5.0 < x \leq 5.5$	$5.5 < x \leq 6.0$	$6.0 < x \leq 6.5$
Frequency	2	10	9	8	1

(i) Find an estimate of the mean and standard deviation of these data.

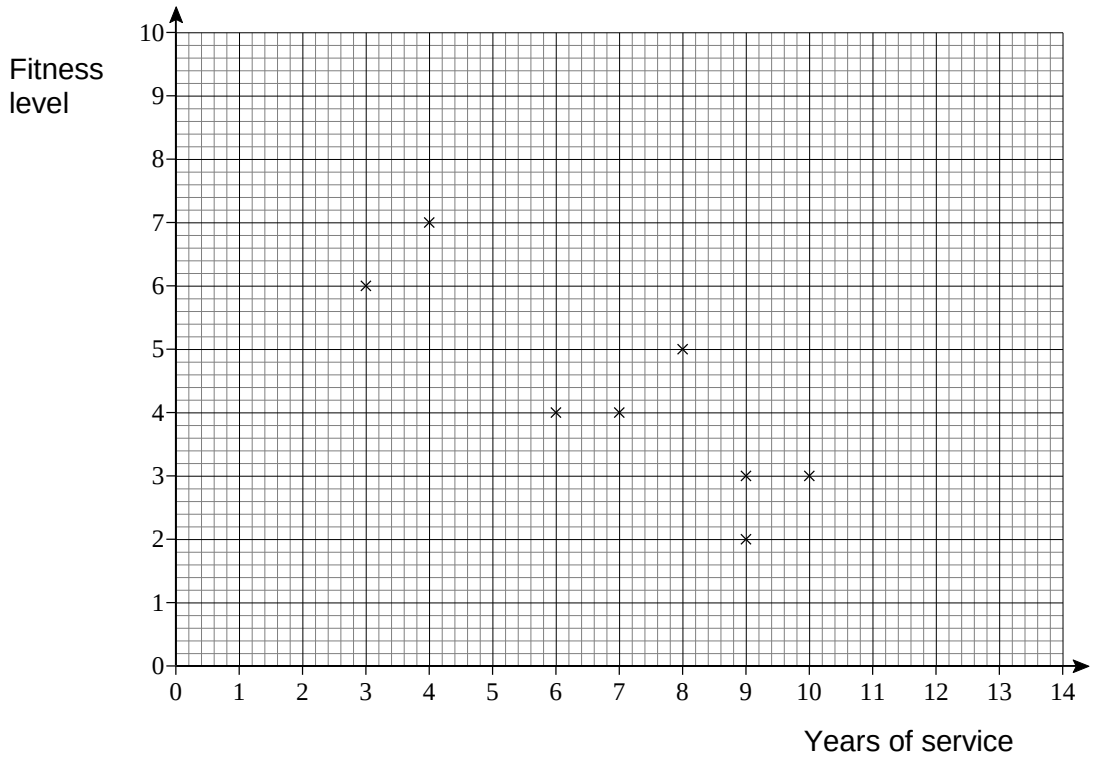
(ii) Determine whether there are any outliers. Explain your answer.

(iii) Why are these values only estimates?

9 A company tests a random sample of 10 employees in order to determine whether their fitness decreases with increasing years of service. The data for the 10 employees are shown in the table below. The level of fitness is determined on a scale of $0 - 10$ by a test carried out one afternoon.

	A	B	C	D	E	F	G	H	I	J
Years of service	3	4	6	7	8	9	9	10	12	14
Fitness level	6	7	4	4	5	3	2	3	2	1

The data for 8 employees are plotted on a graph as shown below.



(i) On a copy of the graph plot the data for the remaining two employees (I and J).

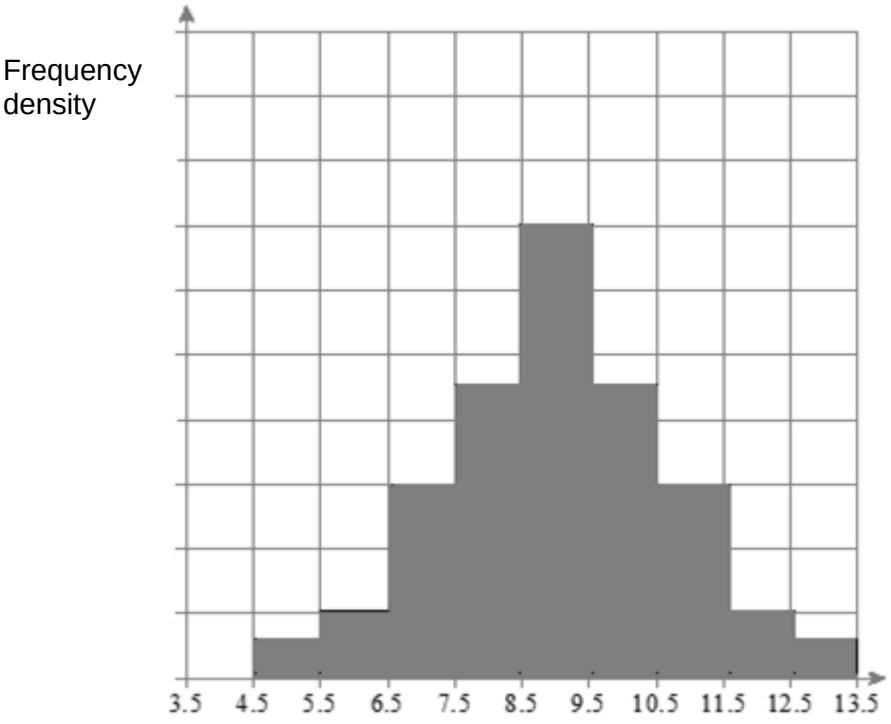
(ii) Draw a line of best fit for the data.

(iii) (a) Suggest a level of fitness for an employee who has worked for the company for 2 years.

(b) Give two reasons why this may not be a reliable value.

10 Philippa drives to and from work, usually in the rush hour.
She wishes to judge the wisdom of driving to and from work earlier or later than usual.

(i) She draws a histogram, as shown below, to model the number of minutes spent stationary on each day when she leaves home at 0800 in the morning and work at 1700 in the evening. The data has been collected at the same times each day for 50 days.
The vertical axis has a correct even scale but the numbers are not shown.



(a) State two features of this histogram.

(b) State an estimate mean time for this model and calculate an estimate for the standard deviation of the times.

(ii) For the next 50 days she leaves home at 0700 and work at 1600. She notes the time spent stationary each day. This time the mean number of minutes spent stationary is 6 minutes and the standard deviation is 0.9.
Comment on these figures in comparison to the model.

Extension

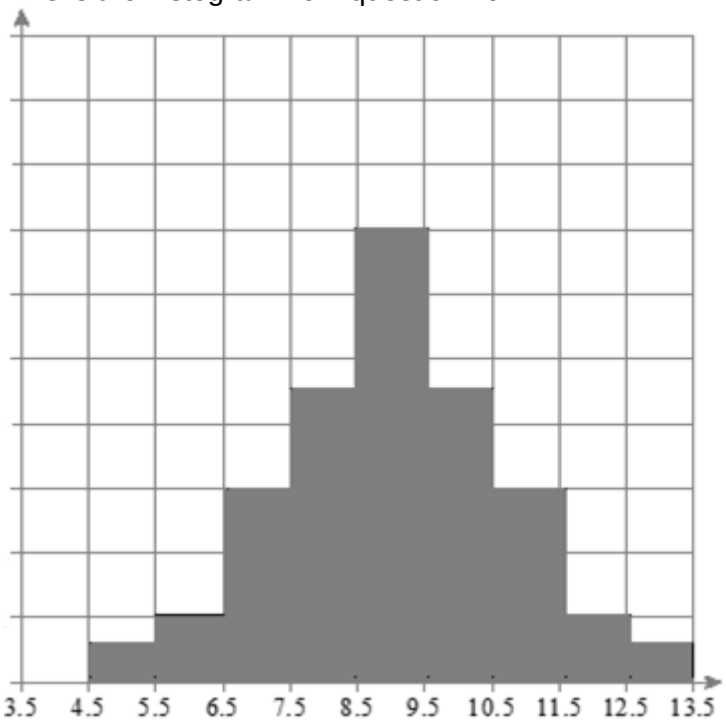
1 Aswin collected 100 pebbles from a beach and weighed them. For these masses, m grams, he calculated the following values.

$$\sum m = 13\,790, \quad \sum m^2 = 1\,925\,650.$$

(i) Calculate the mean and variance of these masses.

(ii) Aswin decided that the one with the greatest mass, 260 grams, was not typical and so decided to remove this pebble from his collection. Find the mean and variance of the remaining 99 pebbles.

2 This is the histogram from question 10.



Sketch a histogram with the same mean but a larger standard deviation.

3. When drawing graphs for the data in your large data set, software will sometimes give a graph that does not display the data well. Choose a suitable graph for an aspect of the data and write instructions for how to choose options that will show the data clearly. Your instructions should be clear enough for someone else to follow.

Worked solutions

1 (i) Mean $= 72.4$, standard deviation $= 25.62$ **BC**

(ii) $72.4 + 2 \times 25.62 = 123.64$ and $72.4 - 2 \times 25.62 = 21.16$ so there are no outliers.

2 Location of the median $= \frac{(70+1)}{2} = 35.5$ so it lies between the 35^{th} and 36^{th} value.

Median $= 10$ km

Location of the UQ $= \frac{(3 \times 70 + 1)}{4} = 52.75$ so it lies between the 52^{nd} and 53^{rd} value.

UQ $= 16$ km

Location of the LQ $= \frac{(70+1)}{4} = 17.75$ so it lies between the 17^{th} and 18^{th} value.

LQ $= 4$ km

3 (i) Whether or not the value is included depends on the purpose for which the average and spread are being calculated; if the purpose is to get an idea of how long the project would take for a typical student, the value should be excluded as the student does not seem to be typical. Outliers like this one could have come from a false statement from the student or it might be because they have been very slow.

(ii)

Time (to the nearest 10 minutes)	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	~~**90**~~
Frequency	5	8	8	5	3	1	0	0	~~1~~
Cumulative frequency	5	13	21	26	29	30	30	30	

Location of the median = $= \frac{(30+1)}{2} = 15.5$ so it lies between the 15^{th} and 16^{th} value.

Cumulative frequency to 20 minutes is 13 and for 30 minutes it is 21 so the median $= 30$

Location of the LQ $= \frac{(30+1)}{4} = 7.75$ so it lies between the 7^{th} and 8^{th} value.

Cumulative frequency to 10 minutes is 5 and for 20 minutes it is 13 so the LQ $= 20$

Location of the UQ = $= \frac{(3 \times 30 + 1)}{4} = 22.75$ so it lies between the 22^{nd} and 23^{rd} value.

Cumulative frequency to 30 minutes is 21 and for 40 minutes it is 26 so the UQ $= 40$

(iii) The values would be the same, the median and quartiles are not affected by outliers (they could have been different since there is an extra data value and this affects where the median and quartiles are located but it happens that all three values are the same).

4 (i) IQR $= 26 - 23 = 3$, median $= 24$

(ii) Outliers above $26 + 1.5 \times 3 = 30.5$. Since the top value is 39 there is at least one outlier at the top end of the distribution.
Outliers below $23 - 1.5 \times 3 = 18.5$. Since the lowest value of the distribution is 20 there are no outliers at the bottom end of the distribution.
Traffic delays and congestion on certain mornings could be a possible explanation for outliers at the top end and the non-existence of outliers at the bottom end would be because the teacher is unable to travel faster than the speed limit to reach their place of work in a faster time.

(iii) Removing the value of 39 means that the top value is now 38.There will be one less data value so the median and quartiles will be located in different places but they are unlikely to change by much. That means that there will be no outliers.

5 (i)

Lifetime (hours)	(i) Frequency	(ii) Cumulative frequency
$0 - 100$	$100 \times 0.1 = 10$	10
$100 - 150$	$50 \times 0.5 = 25$	35
$150 - 200$	$50 \times 0.6 = 30$	65
$200 - 250$	$50 \times 0.7 = 35$	100
$250 - 300$	$50 \times 0.6 = 30$	130
$300 - 400$	$100 \times 0.4 = 40$	140
$400 - 600$	$200 \times 0.15 = 30$	170
		200

(ii) Location of the median $= \frac{(200+1)}{2} = 100.5$ so it lies between the 100^{th} and 101^{st} value.
The cumulative frequency for interval $200 - 250$ is 50% so the median value is the endpoint of this interval. Median $= 250$ hours

6 One does not cause the other; both are likely to depend on a third factor, the sun.

7 (i) Without doing any calculations it can be seen that the distribution is symmetric about 2.

(ii) The value for the mean would have been inflated by the single high value.

(iii) Location of the median $= \frac{(10+1)}{2} = 5.5$ so it lies between the 5^{th} and 6^{th} value.
Cumulative frequency to 1 goal is 5 and for 2 goals it is 7 so the median is 1.5.

8 (i) Using the mid-points: 4.25, 4.75, 5.25, 5.75 and 6.25, the mean $= 5.18$ and the standard deviation $= 0.496$ **BC**

(ii) Upper boundary for outliers is $5.18 + 2 \times 0.496 = 6.17$. The item in the top group could be greater than this so there could be an outlier.
Lower boundary for outliers is $5.18 - 2 \times 0.496 = 4.19$. Both of the items in the lower group could be less than this so there could be 2 outliers.

(iii) The values are only estimates because we do not have raw data and so assume a mid-interval value for all members of each group.

9 (i) Plot the two points $(12, 2)$ and $(14, 1)$.

(ii) The line has intercept on y-axis $7.6 – 8.0$ and gradient approximately $–0.5$

(iii) (a) 7 years

(b) Reasons include: sample is too small; for each person there was only one test one afternoon and there would be many factors affecting the result for each person; and people with the same number of years of service can have different fitness levels.

10 (i) (a) Symmetric, bell shaped

(b) Estimate mean time is 9 minutes (from the symmetry of the histogram).
50 days gives a frequency of 100 (data is collected twice per day). Using the symmetry of the histogram and the fact the bars are equal width of 1 minute, the frequency distribution is as follows:

Mid-point	Frequency
5	$0.5 \times 4 = 2$
6	$1 \times 4 = 4$
7	$3 \times 4 = 12$
8	$4.5 \times 4 = 18$
9	$7 \times 4 = 28$
10	$4.5 \times 4 = 18$
11	$3 \times 4 = 12$
12	$1 \times 4 = 4$
13	$0.5 \times 4 = 2$
Total frequency =	$25 \times 4 = 100$

Estimate standard deviation is 1.637 **BC**

(ii) The mean time is less. Given that it is likely that traffic lights will affect her journey in the same way it is most likely that there will be less traffic congestion resulting in the mean being lower. Each day is rather more consistent meaning that the times are more "bunched" towards the mean.

Extension

1 (i) $\sum m = 13\,790, \quad \sum m^2 = 1\,925\,650$

$$\Rightarrow \bar{m} = \frac{13790}{100} = 137.9$$

$$\sigma^2 = \frac{\sum m^2}{n} - \bar{m}^2 = 19256.5 - 19016.41 = 240.09$$

(ii) $\sum m = 13\,790 - 260 = 13530, \quad \sum m^2 = 1\,925\,650 - 260^2 = 1858050$

$$\Rightarrow \bar{m} = \frac{13530}{99} = 136.67$$

$$\sigma^2 = \frac{\sum m^2}{n} - \bar{m}^2 = 18768.18 - 18677.77 = 90.40$$

2 A histogram with the same line of symmetry but more data in the first and last bars will achieve this.

3. Student's own work.

We'd like to know your view on the resources we produce. By clicking on 'Like' or 'Dislike' you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click 'Send'. Thank you.

Whether you already offer OCR qualifications, are new to OCR, or are considering switching from your current provider/awarding organisation, you can request more information by completing the Expression of Interest form which can be found here: www.ocr.org.uk/expression-of-interest

Looking for a resource? There is now a quick and easy search tool to help find free resources for your qualification: www.ocr.org.uk/i-want-to/find-resources/