A Level · Edexcel · Maths

3 hours 23 questions

2.3 Working with Data

2.3.1 Outliers & Cleaning Data / 2.3.2 Intrepreting Data

Easy (6 questions)	/37
Medium (7 questions)	/47
Hard (5 questions)	/43
Very Hard (5 questions)	/46
Total Marks	**/173**

Scan here to return to the course
or visit savemyexams.com

Easy Questions

1 (a) In a conkers competition the number of strikes required in order to smash an opponent's conker (and thus win a match) is recorded for 15 matches and are given below.

6	2	9	10	9	12	5	
8	7	5	11	9	17	8	9

Find the median, the upper and lower quartiles, and the interquartile range for the number of strikes required to smash a conker.

(3 marks)

(b) An outlier is defined as any data value that falls either more than 1.5 × (interquartile range) above the upper quartile or less than 1.5 × (interquartile range) below the lower quartile.

Identify any outliers.

(2 marks)

2 (a) A hotel manager recorded the number of towels that went missing at the end of each day for 12 days. The results are below.

2	4	1	0	3	4
3.2	9	3	2	4	5

The data value 3.2 is not an outlier but is an error.
Explain why 3.2 is an error and why it should be removed from the data set.

(2 marks)

(b) With the data value 3.2 removed, find the mean and the standard deviation for the number of towels missing at the end of each day.
You may use the summary statistics n=11, Σx=37, Σx^2=181 with the formulae $\overline{x} = \frac{\Sigma x}{n}$ and $\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\overline{x})^2}$

(3 marks)

(c) An outlier is defined as any data value lying outside of 2 standard deviations of the mean. Find any outliers in the data (still excluding 3.2) and justify whether these should be removed from the data set or not.

(2 marks)

3 (a) Joe counts the number of different species of bird visiting his garden each day for a week. The results are given below.

7 8 5 12 9 7 3

Calculate the mean number of different species of bird visiting Joe's garden.

(1 mark)

(b) Joe continues to record the number of different species of bird visiting his garden each day for the rest of the month and calculates the mean number of different species is 9.25 for the remaining 24 days.

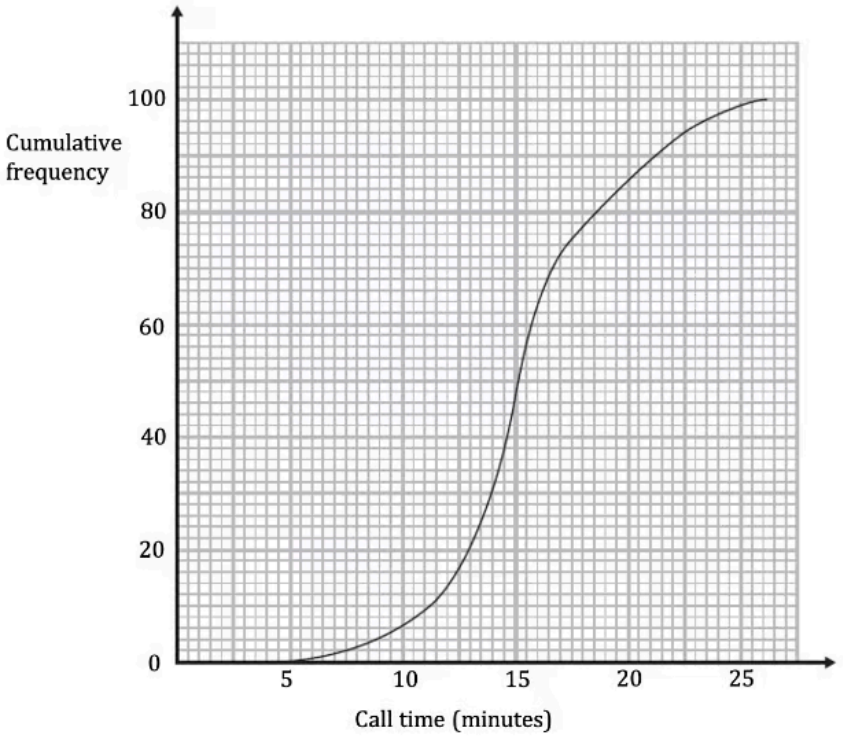
Joe says, using the data from the whole month, he would expect to see 9 different species every day. Explain whether Joe is correct. You must support your answer with clear working.

(3 marks)

(c) Later, Joe notices that one of the values in his data is 8.8. Explain why this must be an error and justify whether you think this value should be removed from the data set or not.

(2 marks)

4 (a) The cumulative frequency diagram below shows the length of 100 phone calls, in minutes, made to a computer help centre for one morning.



(i) Use the cumulative frequency graph to estimate the 10^{th} and 90^{th} percentiles.

(ii) Find the 10^{th} to 90^{th} interpercentile range.

(3 marks)

(b) In the afternoon, on the same day, the length of another 100 phone calls to the computer help centre were recorded. The median length of these calls was 15 minutes and the 10^{th} to 90^{th} interpercentile range was 18 minutes.

Compare the location (median) and spread (interpercentile range) of the calls in the morning and the afternoon.

(3 marks)

5 (a) Two geologists are measuring the size of rocks found on a beach in front of a cliff. The geologists record the greatest length, in millimetres, of each rock they find at distances of 5 m and 25 m from the base of the cliff. They randomly choose 20 rocks at each distance. Their results are summarised in the table below.

Distance from cliff base	5 m	25 m
Number of rocks, n	20	20
Σx	3885	2220
S_{xx}	369 513.75	287 580

Using the formulae $\bar{x} = \frac{\Sigma x}{n}$ and $\sigma = \sqrt{\frac{S_{xx}}{n}}$, find the mean and standard deviation for the size of rocks at both 5 m and 25 m from the base of the cliff.

(3 marks)

(b) Compare the location (mean) and spread (standard deviation) of the size of rocks at 5 m and 25 m from the base of the cliff.

(2 marks)

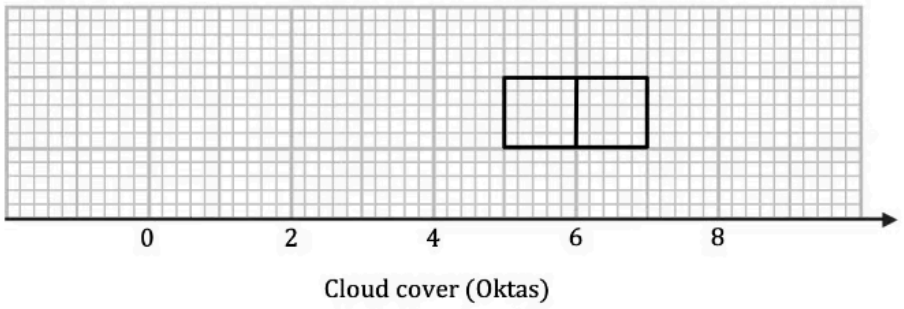
(c) In this instance, an outlier is determined to be any data value that lies outside one standard deviation of the mean ($\bar{x} \pm \sigma$).

(i) Find the smallest rock that is ***not*** an outlier at 5 m from the base of the cliff.

(ii) Briefly explain why there cannot be any ***small*** rock outliers at 25 m from the base of the cliff.

(2 marks)

6 (a) The incomplete box plot below shows data from the large data set regarding cloud cover between May and October 2015 in Cambourne. Cloud cover is measured in Oktas on a scale from 0 (no cloud cover) to 8 (full cloud cover).



Find the interquartile range.

(1 mark)

(b) An outlier is defined as any data value that falls either more than 1.5 × (interquartile range) above the upper quartile or less than 1.5 × (interquartile range) below the lower quartile.

(i) Find the boundaries (fences) at which outliers are defined.

(ii) Explain why, using your knowledge of how cloud cover is measured in the large data set, there cannot be any high valued outliers.

(3 marks)

(c) Complete the box plot given that, where appropriate, the maximum and minimum values should be located at the boundaries (fences) at which outliers are defined. (You are not required to mark any outliers on the box plot.)

(2 marks)

Medium Questions

1 (a) The lengths of unicorn horns are measured in cm. For a group of adult unicorns, the lower quartile was 87 cm and the upper quartile was 123 cm. For a group of adolescent unicorns, the lower quartile was 33 cm and the upper quartile was 55 cm.

An outlier is an observation that falls either more than 1.5 × (interquartile range) above the upper quartile or less than 1.5 × (interquartile range) below the lower quartile.

Which of the following adult unicorn horn lengths would be considered outliers?

32 cm 96 cm 123 cm 188 cm

(2 marks)

(b) Which of the following adolescent unicorn horn lengths would be considered outliers?

12 cm 52 cm 86 cm 108 cm

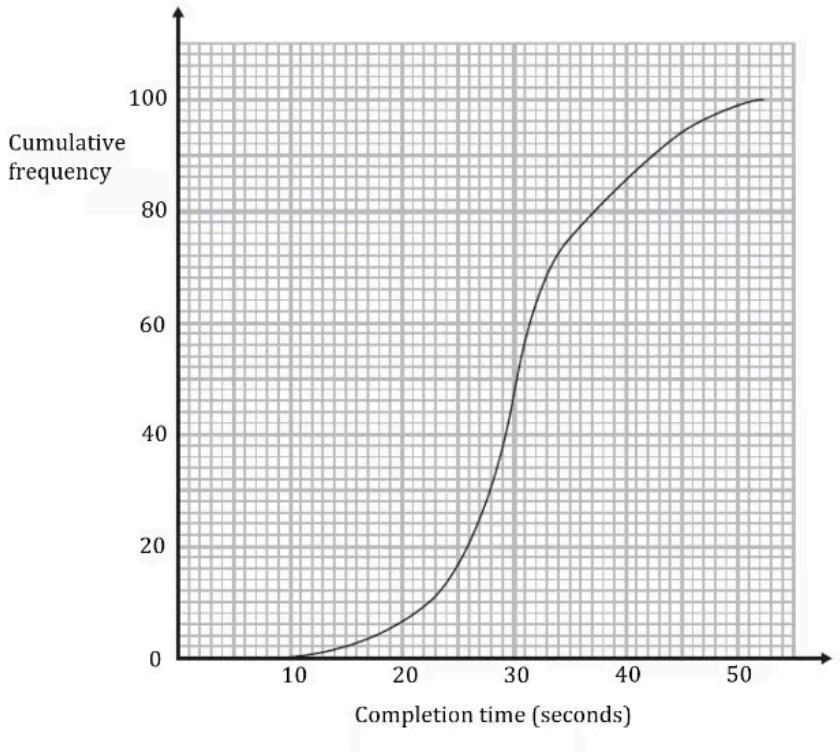
(2 marks)

(c) (i) State the smallest length an adult unicorn horn can be without being considered an outlier.

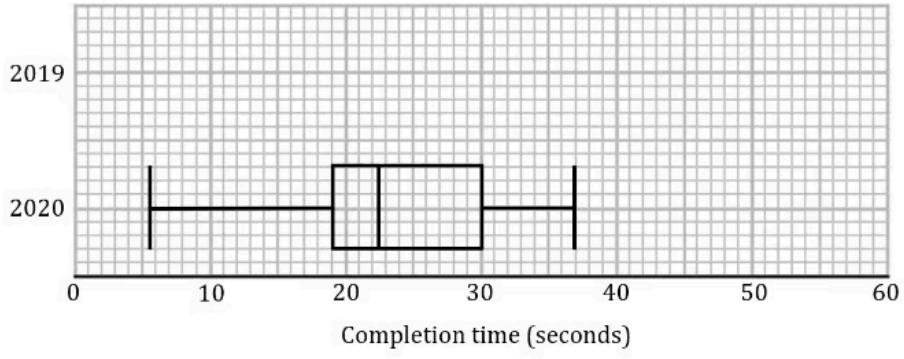
(ii) State the smallest length an adolescent unicorn horn can be without being considered an outlier.

(2 marks)

2 (a) The cumulative frequency diagram below shows completion times for 100 competitors at the 2019 Rubik's cube championships. The quickest completion time was 9.8 seconds and the slowest time was 52.4 seconds.



The grid below shows a box plot of the 2020 championship data. Draw a box plot on the grid to represent the 2019 championship data.



(4 marks)

(b) (i) Compare the distribution of completion times for the 2019 and 2020 championships.

(ii) Given that the 2020 championships happened after the global pandemic, during which many competitors spent months at home, interpret your findings from part (b)(i).

(3 marks)

3 Students at two Karate Schools, Miyagi Dojo and Cobra Kicks, measured the force of a particular style of hit. Summary statistics for the force, in newtons, with which the students could hit are shown in the table below:

	n	Σx	Σx^2
Miyagi Dojo	12	21873	41532545
Cobra Kicks	17	29520	52330890

(i) Calculate the mean and standard deviation for the forces with which the students could hit.

(ii) Compare the distributions for the two Karate Schools.

(7 marks)

4 (a) The heights, in metres, of a flock of 20 flamingos are recorded and shown below:

0.4	0.9	1.0	1.0	1.2	1.2	1.2	1.2	1.2	1.2
1.3	1.3	1.3	1.4	1.4	1.4	1.4	1.5	1.5	1.6

An outlier is an observation that falls either more than 1.5 $\times$ (interquartile range) above the upper quartile or less than 1.5 $\times$ (interquartile range) below the lower quartile.

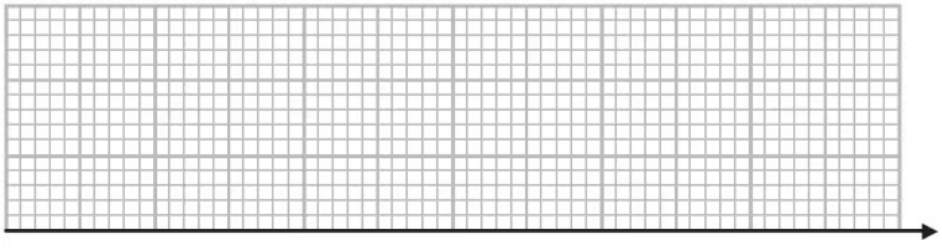
(i) Find the values of Q_1, Q_2 and Q_3.

(ii) Find the interquartile range.

(iii) Identify any outliers.

(4 marks)

(b) Using your answers to part (a), draw a box plot for the data.



(3 marks)

5 (a) The number of daily Covid-19 vaccinations reported by one vaccination centre over a 14-day period are given below:

237	264	308	313	319	352	378
378	405	421	428	450	465	583

Given that $\Sigma x = 5301$ and $\Sigma x^2 = 2\,113\,195$, calculate the mean and standard deviation for the number of daily vaccinations.

(3 marks)

(b) An outlier is an observation which lies more than ± 2 standard deviations away from the mean.

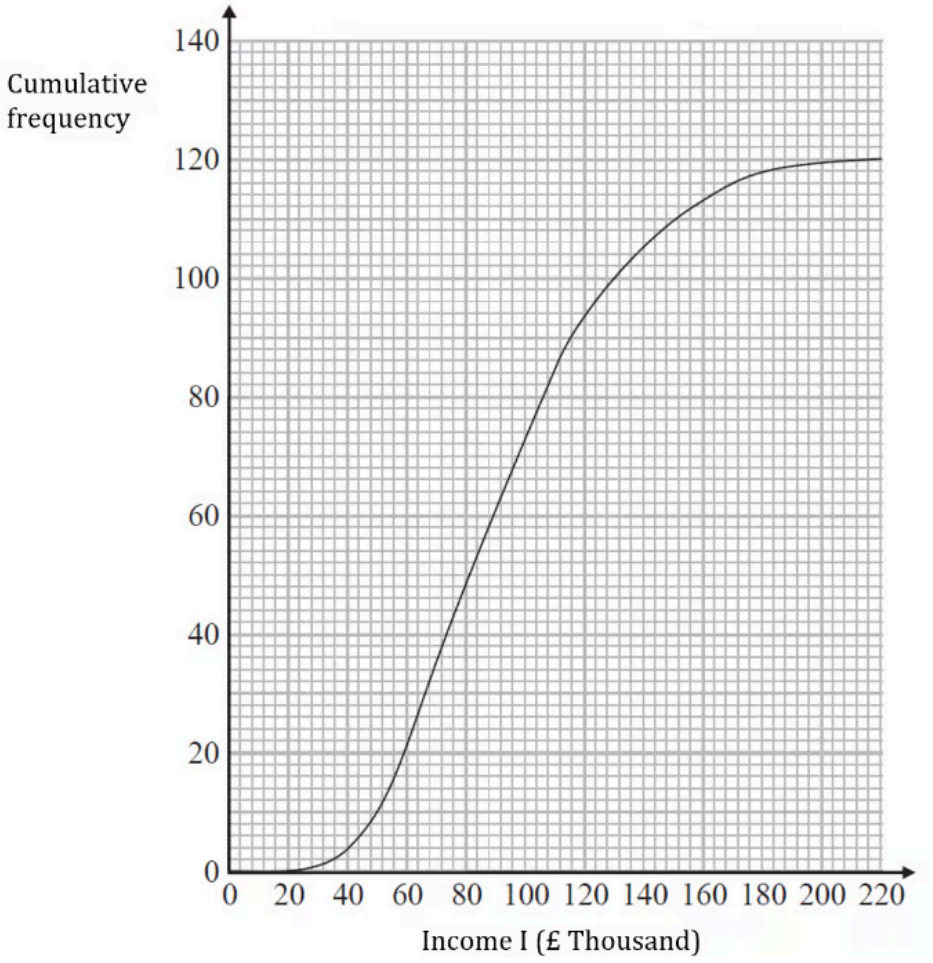
Identify any outliers for this data.

(2 marks)

(c) By removing any outliers identified in part (b), clean the data and recalculate the mean and standard deviation.

(3 marks)

6 The cumulative frequency diagram below shows the distribution of income of 120 managers across a supermarket chain.



The income of a sample of 120 other employees across the supermarket chain are recorded in the table below.

Income I (£ Thousand)	Frequency
$0 \leq I < 20$	34
$20 \leq I < 40$	28
$40 \leq I < 60$	27
$60 \leq I < 80$	17
$80 \leq I < 100$	10
$100 \leq I < 120$	4

On the grid above, draw a cumulative frequency graph to show the data for the other employees and compare the income of managers and other employees.

(7 marks)

7 (a) Summary statistics from the large data set for the daily mean windspeed (knots) measured in Heathrow throughout October 1987 and October 2015 are given in the table below.

	Min	**Max**	**Median**	Σx	Σx^2
1987	2	16	5	185	1401
2015	3	10	6	197	1357

Calculate the mean of the daily mean windspeeds for each of the two years.

(2 marks)

(b) The standard deviation for 2015 was 1.84.

Calculate the standard deviation for 1987 and compare the daily mean windspeeds for each of the two years.

(3 marks)

Hard Questions

1 (a) As part of an experiment, 15 maths teachers are asked to solve a riddle and their times, in minutes, are recorded:

8	12	19	20	20
21	22	23	23	23
25	26	27	37	39

An outlier is an observation which lies more than ± 2 standard deviations away from the mean.

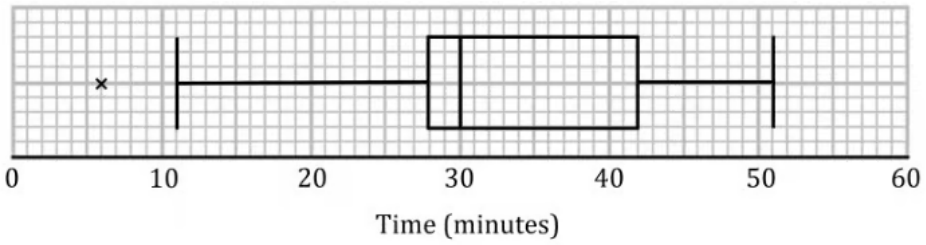
Show that there is exactly one outlier.

(4 marks)

(b) State, with a reason, whether the mean or the median would be the most suitable measure of central tendency for these data.

(2 marks)

(c) 15 history teachers also completed the riddle; their times are shown below in the box plot:



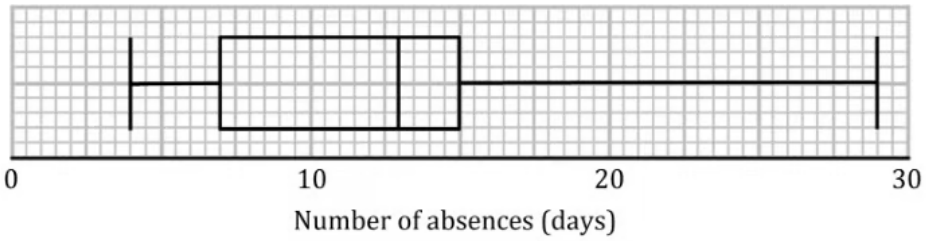
Explain what the cross (×) represents on the box plot above. Interpret this in context.

(2 marks)

(d) By comparing the distributions of times taken to complete the riddle, decide which set of teachers were faster at solving the riddle.

(4 marks)

2 (a) Hugo, a newly appointed HR administrator for a company, has been asked to investigate the number of absences within the IT department. The department contains 23 employees, and the box plot below summarises the data for the number of days that individual employees were absent during the previous quarter.



An outlier is an observation that falls either more than 1.5 (interquartile range) above the upper quartile or less than 1.5 (interquartile range) below the lower quartile.

Show that these data have an outlier, and state its value.

(3 marks)

(b) For the 23 employees within the department, Hugo has the summary statistics:

$$\Sigma x = 286 \text{ and } \Sigma x^2 = 4238$$

Hugo investigates the employee corresponding to the outlier value found in part (a) and discovers that this employee had a long-term illness. Hugo decides not to include that value in the data for the department.

Assuming that there are no other outliers, calculate the mean and standard deviation of the number of days absent for the remaining employees.

(4 marks)

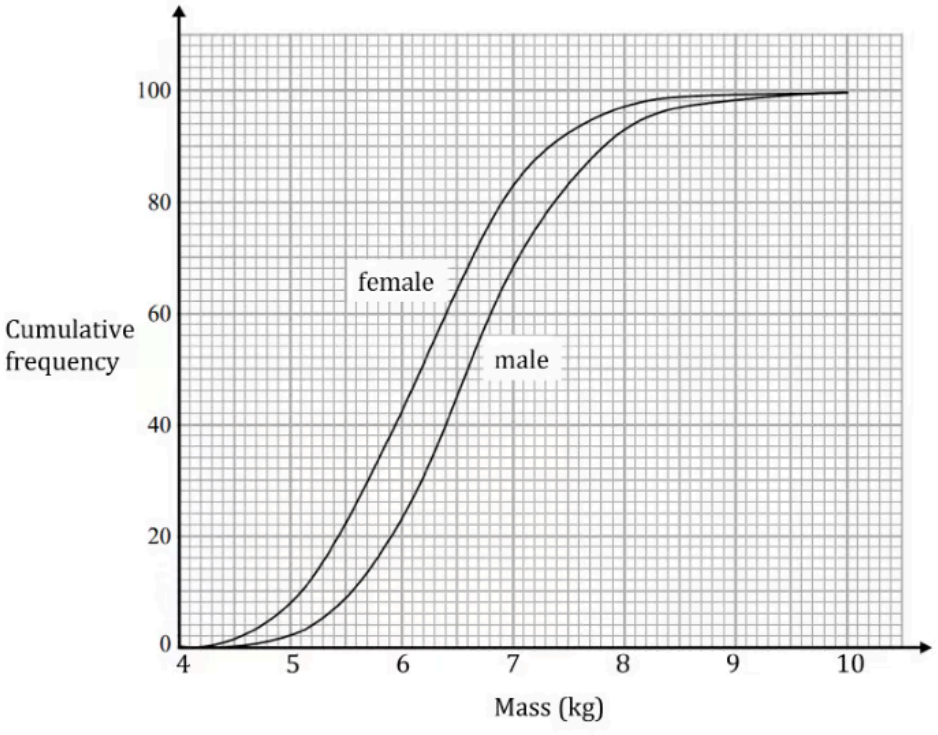
3 (a) Sam, a zoologist, is a member of a group researching the masses of gentoo penguins. The research group takes a sample of 100 male and 100 female penguins and records their masses.

An outlier is an observation that falls either more than 1.5 × (interquartile range) above the upper quartile or less than 1.5 × (interquartile range) below the lower quartile.

Given that values are outliers if they are less than 4.2kg or more than 8.5kg, calculate the upper and lower quartiles for the mass of the 200 gentoo penguins.

(4 marks)

(b) Casey is another member of Sam's research group. She believes that the masses of male and female gentoo penguins follow different distributions. The cumulative frequency graphs below show the masses of the male and female gentoo penguins in the sample.



By calculating a measure of central tendency and a measure of variation, compare the two distributions.

(5 marks)

4 (a) Ms Chew is an accountant who is examining the length of time it takes her to complete jobs for her clients. Ms Chew looks at her spreadsheet and lists the number of hours it took her to complete her last 12 jobs:

9 2 - 6 5 2 - 6 21 5 4 8

'-' represents a job for which the length of time taken was not recorded.

An outlier is an observation which lies more than ± 2 standard deviations away from the mean.

By first cleaning the data, show that 21 is the only outlier.

(4 marks)

(b) Ms Chew looks at her handwritten records and finds that the value 21 was typed into the spreadsheet incorrectly. It should have been 12.

Without further calculations, explain the effect this would have on the:

(i) mean

(ii) standard deviation

(iii) median.

(3 marks)

5 (a) David and Bowey are planning a trip in June to Beijing, Jacksonville or Perth. The temperature of the city and the atmospheric pressure will be deciding factors, so they investigate these three cities using all of the data for June 2015 from the large data set.

Using all of the days in June 2015, the following summary statistics for the daily mean air temperatures (t °C) and the daily mean pressure (p hPa) are calculated:

	Daily mean air temperature		**Daily Mean Pressure**	
	$\bar{t}$	σ_t	$\bar{p}$	σ_p
Beijing	a	b	1004	3.81
Jacksonville	26.4	1.80	1017	1.88
Perth	14.8	2.37	1021	5.63

David also has the following information for Beijing in June:

$$\sum t = 741.8 \text{ and } \sum t^2 = 18513.2$$

Calculate the values of a and b.

(3 marks)

(b) David suffers from headaches when the atmospheric pressure changes quickly so he would like to choose a city where the pressure does not vary a lot. Additionally, Bowey does not like it when the temperature is higher than 30 °C. It is known that all the temperatures for Beijing in June 2015 were within 2 standard deviations of the mean, whereas in Jacksonville there were temperatures that were higher than the mean by more than 2 standard deviations.

Suggest a city which both David and Bowey would be happy to visit. Give reasons for your answer.

(5 marks)

Very Hard Questions

1 (a) Marya is consistently late for work. David, Marya's boss, records the number of minutes that she is late during the next six days. David calculates the mean is 18 minutes and the variance is 210 minutes^2. On one of the six days, Marya was 50 minutes late.

Show that 50 is an outlier, using the definition that outliers are more than 2 standard deviations away from the mean.

(2 marks)

(b) Marya states that the 50 minutes should not be included as it is an outlier.

(i) Give a reason why Marya wants the 50 minutes to be excluded from the data set.

(ii) Give a reason why David wants the 50 minutes to be included in the data set.

(2 marks)

(c) Marya tells David that she was 50 minutes late that day due to a road accident, she shows David the traffic report as evidence.

David agrees to remove the 50 from the dataset, calculate the new mean and standard deviation for the remaining values.

(5 marks)

2 For each scenario state, with a reason, whether the identified outlier should be included or excluded in the data set.

(i) Alice is collecting the ages of children in a school classroom. The outlier is the age of 29.

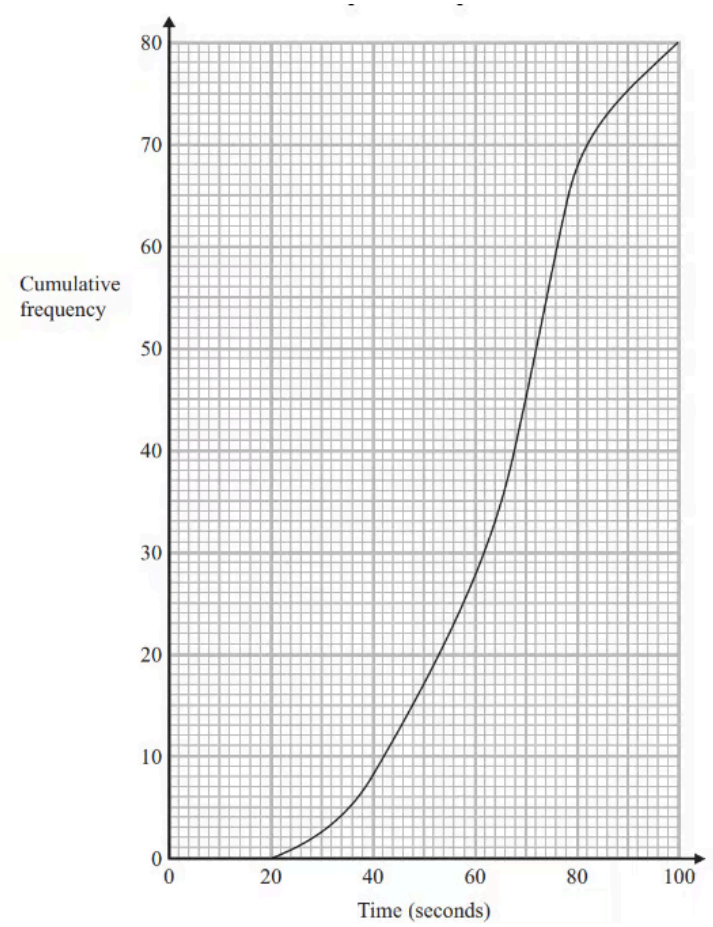
(ii) Benji records the times taken for some athletes to run a mile. The outlier is the time of 7 seconds.

(iii) Carlos is collecting data on the number of hours of sunlight per day for the city, Burrow, located in the north of the North America. The outlier is the value of 23.4 hours.

(iv) Daisy is collecting data on the heights of cows; the median height is 161cm. The outlier is the height 189cm.

(8 marks)

3 (a) The cumulative frequency graph below shows the information about the lengths of time taken for 80 students to run a lap of the sports hall.



Complete the table below:

Time (*t* seconds)	20 < t ≤ 40	40 < t ≤ 60	60 < t ≤ 80	80 < t ≤ 100
Frequency	8			

(3 marks)

(b) Hence estimate the mean and the standard deviation of the times.

(3 marks)

(c) Given that the fastest time was 21 seconds and the slowest time was 100 seconds, show that these values are outliers using the definition that an outlier is more than 2 standard deviations away from the mean.

(3 marks)

4 (a) Tim has just moved to a new town and is trying to choose a doctor's surgery to join, HealthHut or FitFirst. He wants to register with the one where patients get seen faster. He takes of sample of 150 patients from HealthHut and calculates the range of waiting times as 45 minutes and the variance as 121 minutes^2.

An outlier is defined as a value which is more than 2 standard deviations away from the mean.

Prove that the sample contains an outlier.

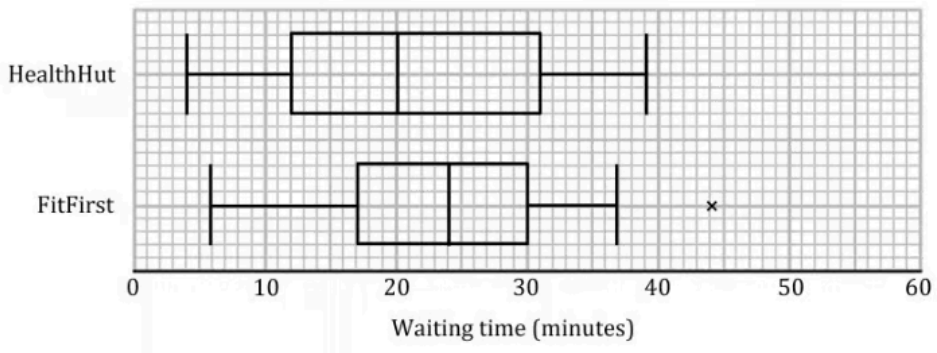
(3 marks)

(b) Tim finds out that the outlier is a valid piece of data and decides to keep the value in his sample.

Which pair of statistical measures would be more appropriate to use when using the sample to compare the doctor's surgeries: the mean and standard deviation or the median and interquartile range? Give a reason for your answer.

(2 marks)

(c) The box plots below show the waiting times for the two surgeries.



Given that there is only one outlier for HealthHut, label it on the box plot with a cross (×).

(1 mark)

(d) Compare the two distributions of waiting times in context.

(4 marks)

5 (a) Ororo, a meteorologist, is investigating the great storm of 1987 which devastated the south of England. Ororo would like to compare the daily maximum gust in Hurn during the months of October 1987 and October 2015.

Using your knowledge of the large data set

(i) suggest one other city from the large data set that Ororo could use to investigate the great storm of 1987

(ii) state the units that are used in large data set to measure the daily maximum gust.

(2 marks)

(b) Ororo calculates the following summary statistics for the daily maximum gust in Hurn using the available data for October:

	Number of available days	**Maximum value**	Σx	S_{xx}
1987	25	61	665	3462
2015	31	27	586	612.56

An outlier is defined as a value which is more than 2 standard deviations away from the mean.

(i) Show that the maximum value in 1987 is an outlier.

(ii) Give a reason why Ororo should include the outlier when comparing the data from the two years.

(4 marks)

(c) Compare the daily maximum gust in Hurn for October 1987 and October 2015.

(4 marks)