1. The heights x cm of 100 boys in Year 7 at a school are summarised in the table below.

Height	$125 \leqslant x \leqslant 140$	$140 < x \leqslant 145$	$145 < x \leqslant 150$	$150 < x \leqslant 160$	$160 < x \leqslant 170$
Frequency	25	29	24	18	4

i. Estimate the number of boys who have heights of at least 155 cm.

[2]

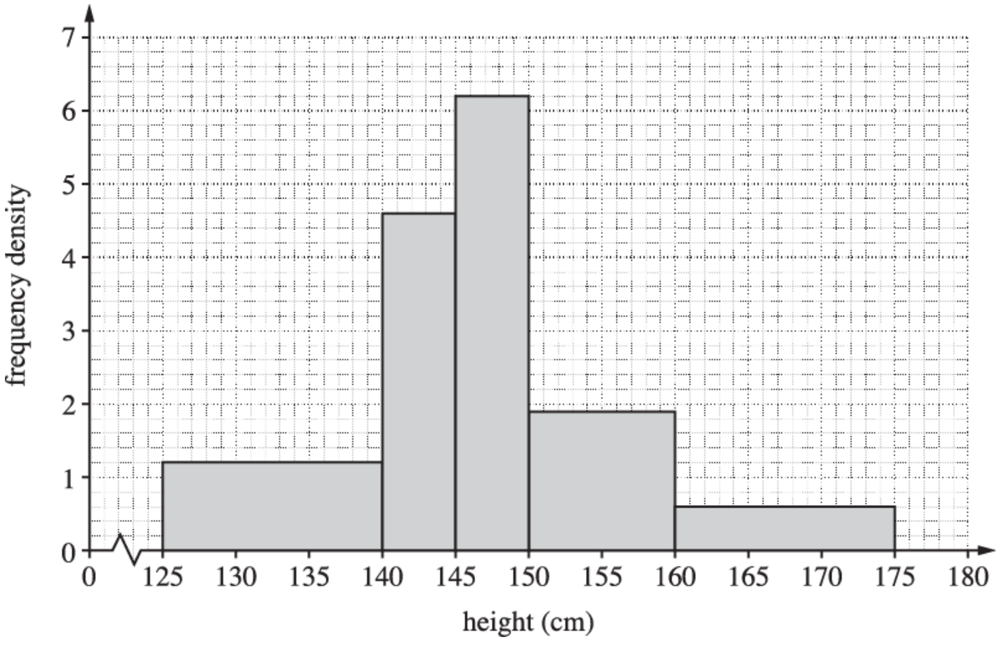
ii. Calculate an estimate of the median height of the 100 boys.

[3]

iii. Draw a histogram to illustrate the data.

[5]

The histogram below shows the heights of 100 girls in Year 7 at the same school.



iv. How many more girls than boys had heights exceeding 160 cm?

[3]

v. Calculate an estimate of the mean height of the 100 girls.

[5]

2. The stem and leaf diagram illustrates the heights in metres of 25 young oak trees.

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

Key: 4|2 represents 4.2

i. State the type of skewness of the distribution.

[1]

ii. Use your calculator to find the mean and standard deviation of these data.

[3]

iii. Determine whether there are any outliers.

[4]

3. The birth weights in kilograms of 25 female babies are shown below, in ascending order.

1.39 2.50 2.68 2.76 2.82 2.82 2.84 3.03 3.06 3.16 3.16 3.24 3.32
3.36 3.40 3.54 3.56 3.56 3.70 3.72 3.72 3.84 4.02 4.24 4.34

i. Find the median and interquartile range of these data.

[3]

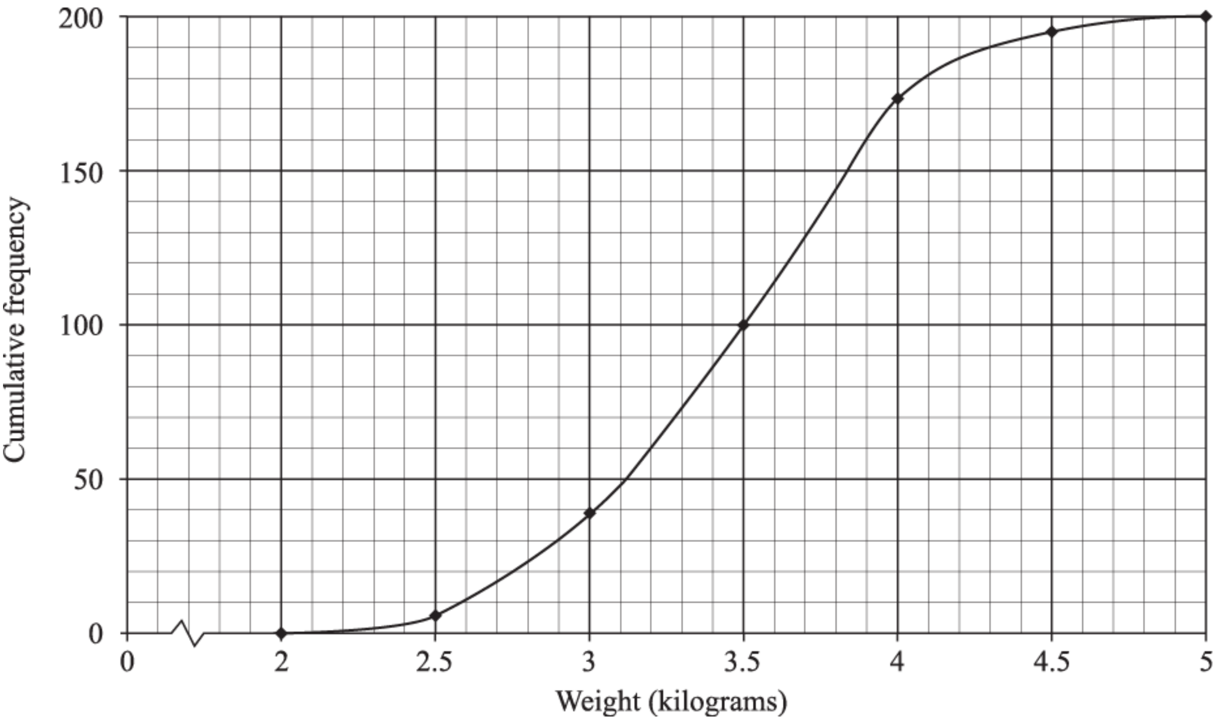
ii. Draw a box and whisker plot to illustrate the data.

[3]

iii. Show that there is exactly one outlier. Discuss whether this outlier should be removed from the data.

[4]

The cumulative frequency curve below illustrates the birth weights of 200 male babies.



iv. Find the median and interquartile range of the birth weights of the male babies.

[3]

v. Compare the weights of the female and male babies.

[2]

vi. Two of these male babies are chosen at random. Calculate an estimate of the probability that both of these babies weigh more than any of the female babies.

[3]

4. The weights, w grams, of a random sample of 60 carrots of variety A are summarised in the table below.

Weight	$30 \leqslant w < 50$	$50 \leqslant w < 60$	$60 \leqslant w < 70$	$70 \leqslant w < 80$	$80 \leqslant w < 90$
Frequency	11	10	18	14	7

i. Draw a histogram to illustrate these data.

[5]

ii. Calculate estimates of the mean and standard deviation of w.

[4]

iii. Use your answers to part **(ii)** to investigate whether there are any outliers.

[3]

The weights, x grams, of a random sample of 50 carrots of variety B are summarised as follows.

$$n = 50 \qquad \sum x = 3624.5 \qquad \sum x^2 = 265\,416$$

iv. Calculate the mean and standard deviation of x.

[3]

v. Compare the central tendency and variation of the weights of varieties A and B.

[2]

5. The ages, x years, of the senior members of a running club are summarised in the table below.

Age (x)	$20 \leqslant x < 30$	$30 \leqslant x < 40$	$40 \leqslant x < 50$	$50 \leqslant x < 60$	$60 \leqslant x < 70$	$70 \leqslant x < 80$	$80 \leqslant x < 90$
Frequency	10	30	42	23	9	5	1

i. Draw a cumulative frequency diagram to illustrate the data.

[5]

ii. Use your diagram to estimate the median and interquartile range of the data.

[3]

6. At a tourist information office the numbers of people seeking information each hour over the course of a 12-hour day are shown below.

6 25 38 39 31 18 35 31 33 15 21 28

i. Construct a sorted stem and leaf diagram to represent these data.

[3]

ii. State the type of skewness suggested by your stem and leaf diagram.

[1]

iii. For these data find the median, the mean and the mode. Comment on the usefulness of the mode in this case.

[4]

7. An online store has a total of 930 different types of women's running shoe on sale. The prices in pounds of the types of women's running shoe are summarised in the table below.

Price (£x)	$10 \leq x \leq 40$	$40 < x \leq 50$	$50 < x \leq 60$	$60 < x \leq 80$	$80 < x \leq 200$
Frequency	147	109	182	317	175

i. Calculate estimates of the mean and standard deviation of the shoe prices.

[4]

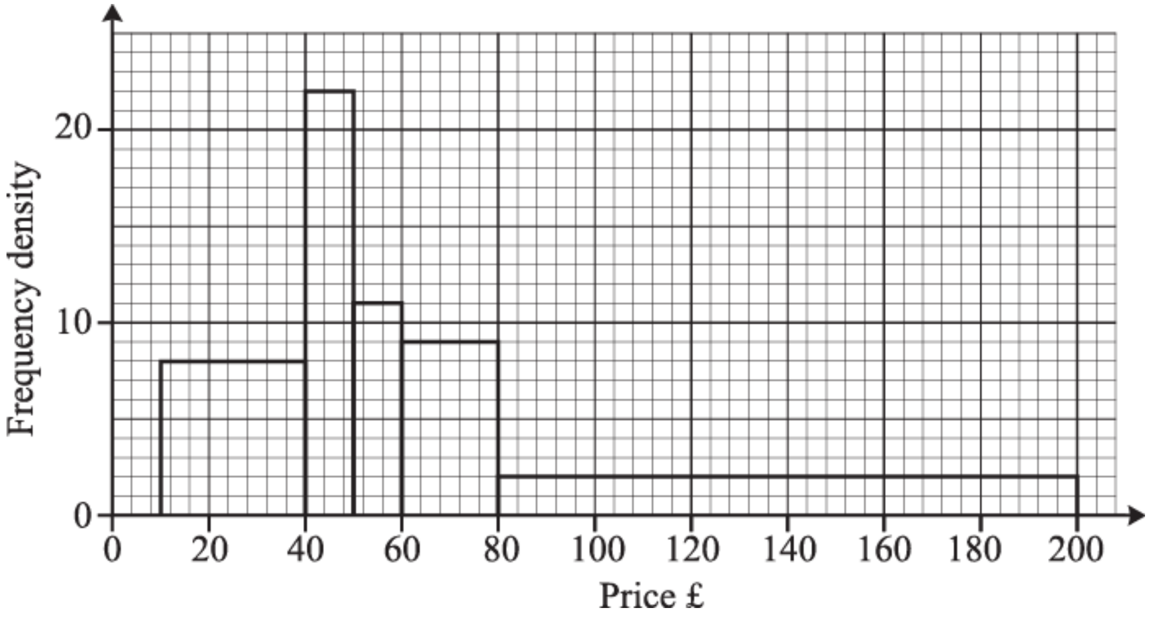
ii. Calculate an estimate of the percentage of types of shoe that cost at least £100.

[3]

iii. Draw a histogram to illustrate the data.

[5]

The corresponding histogram below shows the prices in pounds of the 990 types of men's running shoe on sale at the same online store.



iv. State the type of skewness shown by the histogram for men's running shoes.

[1]

v. Martin is investigating the percentage of types of shoe on sale at the store that cost more than £100. He believes that this percentage is greater for men's shoes than for women's shoes. Estimate the percentage for men's shoes and comment on whether you can be certain which percentage is higher.

[3]

vi. You are given that the mean and standard deviation of the prices of men's running shoes are £68.83 and £42.93 respectively. Compare the central tendency and variation of the prices of men's and women's running shoes at the store.

[2]

8. The stem and leaf diagram illustrates the weights in grams of 20 house sparrows.

25	0		
26	0	5	8
27	7	9	
28	1	4	5
29	0	0	2
30	7	7	
31	6		
32	0	4	7
33	3	3	

Key: 27 | 7 represents 27.7 grams

i. Find the median and interquartile range of the data.

[3]

ii. Determine whether there are any outliers.

[4]

9. Alison selects 10 of her male friends. For each one she measures the distance between his eyes. The distances, measured in mm, are as follows:

51 57 58 59 61 64 64 65 67 68

The mean of these data is 61.4. The sample standard deviation is 5.232, correct to 3 decimal places.

One of the friends decides he does not want his measurement to be used. Alison replaces his measurement with the measurement from another male friend. This increases the mean to 62.0 and reduces the standard deviation. Give a possible value for the measurement which has been removed and find the measurement which has replaced it. **[3]**

10. A farmer has 200 apple trees. She is investigating the masses of the crops of apples from individual trees. She decides to select a sample of these trees and find the mass of the crop for each tree.

(a) Explain how she can select a random sample of 10 different trees from the 200 trees. **[2]**

The masses of the crops from the 10 trees, measured in kg, are recorded as follows.

23.5 27.4 26.2 29.0 25.1 27.4 26.2 28.3 38.1 24.9

(b) For these data find

- the mean,
- the sample standard deviation.

[2]

(c) Show that there is one outlier at the upper end of the data. How should the farmer decide whether to use this outlier in any further analysis of the data? **[3]**

11. Fig. 9.1 shows box and whisker diagrams which summarise the birth rates per 1000 people for all the countries in three of the regions as given in the pre-release data set. The diagrams were drawn as part of an investigation comparing birth rates in different regions of the world.

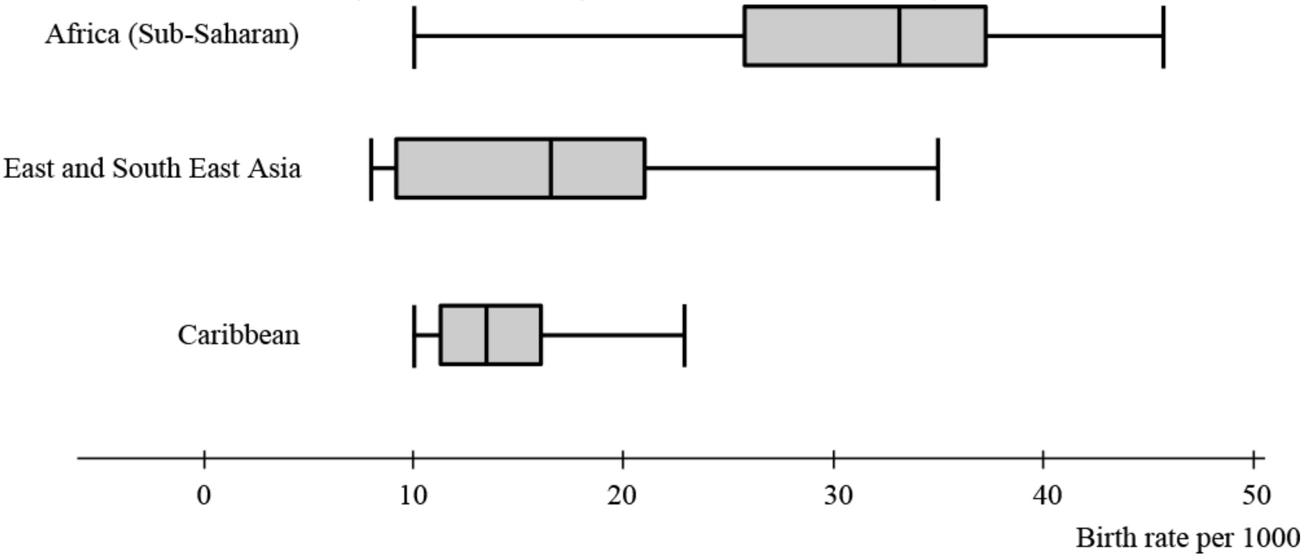


Fig. 9.1

(a) Discuss the distributions of birth rates in these regions of the world. Make three different statements. You should refer to **both** information from the box and whisker diagrams **and** your knowledge of the large data set. To access the Large Data Set please go to http:/www.ocr.org.uk/Images/308749-units-h630-and-h640-large-data-set-lds-sample-assessment-material.xls [3]

(b) The birth rates for all the countries in Australasia are shown below.

Country	Birth rate per 1000
Australia	12.19
New Zealand	13.4
Papua New Guinea	24.89

(i) Explain why the calculation below is not a correct method for finding the birth rate per 1000 for Australasia as a whole.

$$\frac{12.19+13.4+24.89}{3} \approx 16.83$$

[1]

(ii) Without doing any calculations, explain whether the birth rate per 1000 for Australasia as a whole is higher or lower than 16.83. [1]

The scatter diagram in Fig. 9.2 shows birth rate per 1000 and physicians/1000 population for all the countries in the pre-release data set.

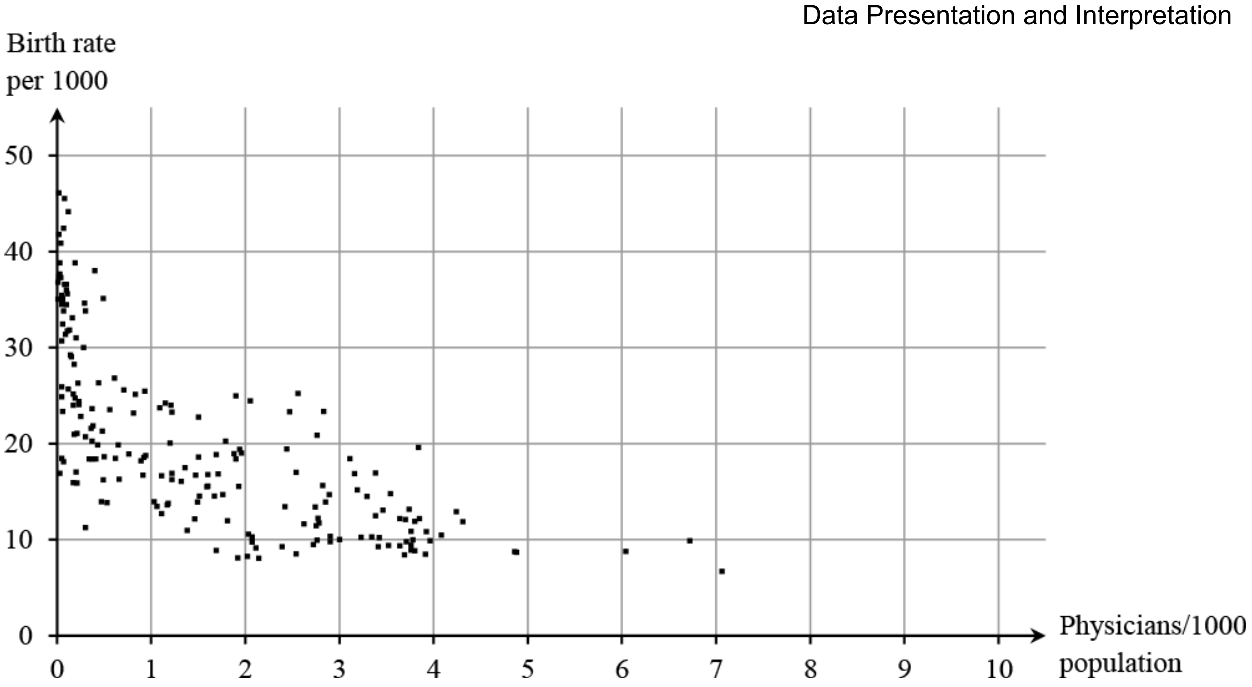


Fig. 9.2

(c) Describe the correlation in the scatter diagram. [1]

(d) Discuss briefly whether the scatter diagram shows that high birth rates would be reduced by increasing the number of physicians in a country. [1]

12. The maximum daytime temperature was recorded on each day in May 2016 at a weather station in Canada. The data are shown in the stem-and-leaf diagram below.

0	3
5	
10	0 0 1 3 3
15	0 1 1 2 3 4
20	0 0 1 2 2 3 3 4
25	0 0 1 2 2 2
30	0 0 2 2 3

key: 15 | 1 represents a temperature of 16 °C

(a) Describe the shape of the distribution. [1]

(b) Find the interquartile range. [2]

(c) Hence determine whether 3 is an outlier. [2]

13. A recruitment company advertises vacancies on their website. Information on the salaries for 36 of these vacancies is given in Fig. 7. The data have been grouped.

Salary in thousands of pounds	20 –	25 –	30 –	35 –	40 –	45 –	50 – 55
Number of vacancies	3	6	6	12	3	3	3

Fig. 7

(a) For these salaries, calculate estimates of

- the mean,
- the sample standard deviation.

Give your answers to the nearest pound. [4]

(b) Explain why your values are only estimates. [1]

(c) Give a reason why it would not be appropriate to use the mean calculated in part (a) as an estimate of the mean salary for all vacancies in the country. [1]

(d) Another vacancy has an annual salary of £52 573. This was not included in the table. Without further calculation, state how the mean salary would be affected if it were to be recalculated including this value. [1]

14. The managing director of an international internet communications company wishes to investigate the number of internet users and the number of mobile phone users for different

countries in Eastern Europe. He wishes to identify a country with the potential to increase the number of internet users so that he can consider investing in that country.

Fig. 11.1 shows all the countries of Eastern Europe.

Country (in Eastern Europe)	Mobile phone subscribers (millions)	Internet users (millions)
Albania	3.50	1.30
Belarus	10.68	2.64
Bosnia and Herzegovina	3.35	1.42
Bulgaria	10.78	3.40
Croatia	4.97	2.23
Crezh Republic	12.97	6.68
Estonia	2.07	0.97
Hungary	11.58	6.18
Kosovo	0.56	missing
Moldova	4.08	1.33
Montenegro	1.13	0.28
Poland	50.84	22.45
Romania	22.70	7.79
Serbia	9.14	4.11
Slovakia	6.10	4.06
Slovenia	2.25	1.30
Ukraine	59.34	7.77

Source: CIA World Factbook

Fig. 11.1

(a) Are the countries in Fig. 11.1 a sample or a population? Explain your answer. [1]

(b) These data have been used to construct the scatter diagram in Fig. 11.2. Use your knowledge of the large data set to comment on the correlation in the scatter diagram. [2]

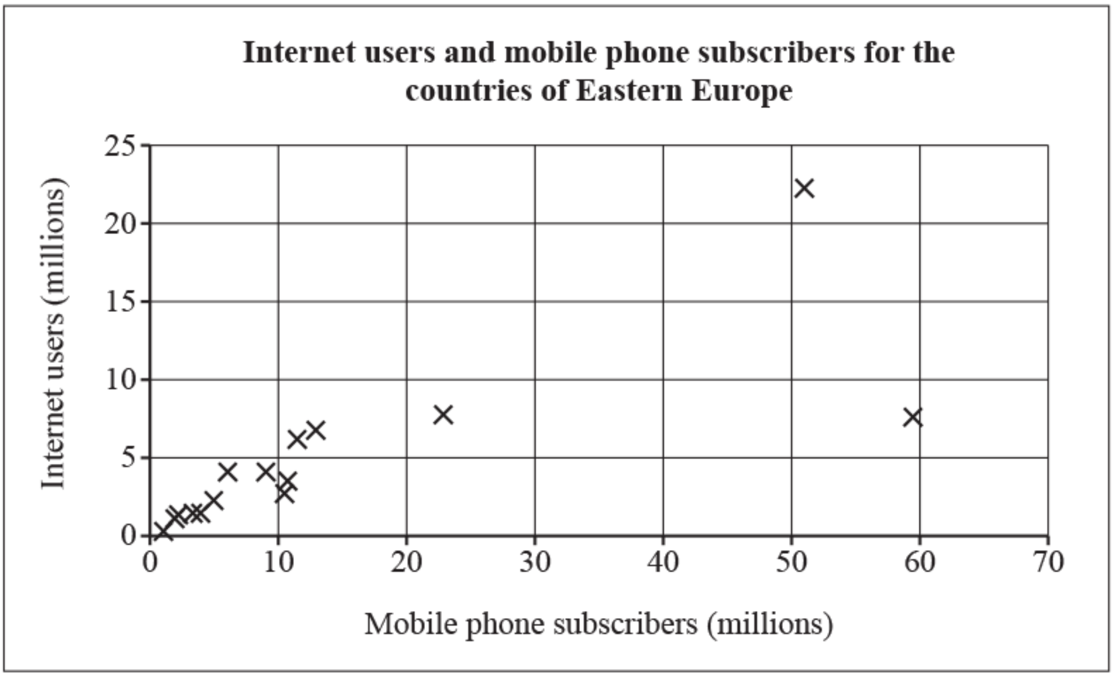


Fig. 11.2

(c) Use the scatter diagram to identify a country that appears to have high potential to increase the number of internet users. Give a reason for your choice. [2]

(d) Having decided on the country he wishes to invest in, the director will select a sample of 20 marketing consultants from that country to contact for information and advice. He has found a website listing 600 marketing consultants. Give clear instructions for the director on how to select a simple random sample of 20 marketing consultants from the 600. [3]

15. The head of sales of a large company presented Fig. 3 to the board of directors as part of his end-of-year report.

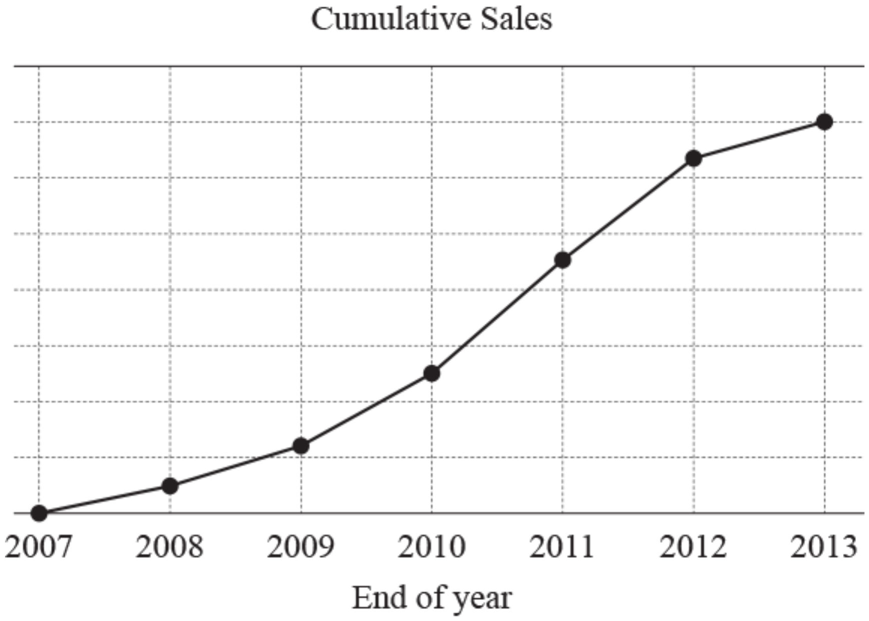


Fig. 3

(a) What key feature is missing from this graph? [1]

One director comments that the diagram shows increasing year-on-year sales.

(b) Explain whether the director is correct. [1]

16.

(a) (i) The pre-release data shows that the total population of the 239 countries in the world in 2016 was 7174654290. The populations of a sample of 10 countries are given in Fig. 9.1.

Country	Population
Tuvalu	10 782
Equatorial Guinea	722 254
Somalia	10 428 043
Denmark	5 569 077
Burma	55 746 253
Norway	5 147 792
Botswana	2 155 784
Rwanda	12 337 138
Sint Maarten	39 689
Swaziland	1 419 623

Fig. 9.1

Show that the mean population per country for the whole world is much larger than the mean population per country for this sample. [3]

(ii) Rebecca takes a large number of different samples of 10 countries. She finds that the mean population per country is usually smaller for the sample than it is for the whole world. Explain whether this suggests that the sampling was not random. [2]

(b) Fig. 9.2 shows data for Norway.

Country	Population	GDP per capita (US$)	Health expenditure (% of GDP)
Norway	5 147 792	55 400	9.1

Fig. 9.2

Calculate Norway’s health expenditure per person in US$. [2]

(c) As part of an investigation into the factors which might be associated with life expectancy, the scatter diagrams in Fig. 9.3 are drawn. The line of best fit and the corresponding value of the square of the correlation coefficient are also shown for each scatter diagram.

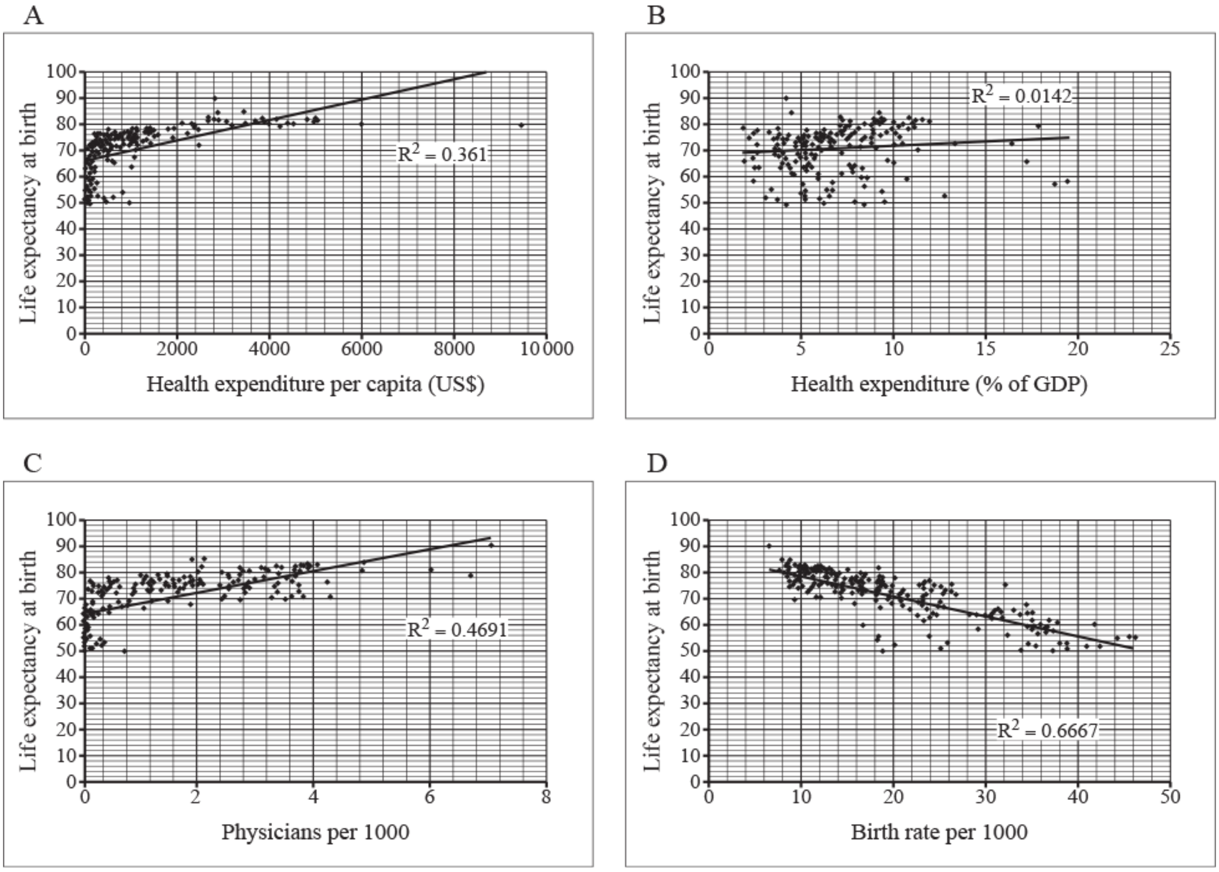


Fig. 9.3

(i) Which of the four factors appears to have the strongest positive association with life expectancy at birth? Give a reason for your answer. [2]

(ii) Explain why the line of best fit in scatter diagram B is not a good model for the relationship between the two variables. [1]

17. The numbers of units of electricity, x kWh (kilowatt-hours), used by 50 customers of an energy firm in a period of one month are summarised as follows.

$$\Sigma x = 17\ 100 \qquad \Sigma x^2 = 6115\ 108$$

(i) Calculate the mean and standard deviation of x. [3]

(ii) The cost, £y, of the electricity used by each customer is given by the formula $y = 0.108x + 7.2$. Use your answers to part (i) to deduce the mean and standard deviation of the costs of the electricity used by these customers. [3]

18. The table below shows the maximum daily level of the pollutant nitrogen dioxide in Marylebone Road in London in 2015. The levels are measured in micrograms per cubic metre (µg / m^3). There were 7 days where no figures were available.

Pollutant level (x µg/m^3)	$40 \leqslant x < 80$	$80 \leqslant x < 120$	$120 \leqslant x < 140$	$140 \leqslant x < 180$	$180 \leqslant x < 220$	$220 \leqslant x \leqslant 300$
Frequency	29	74	52	129	64	10

(i) Draw a cumulative frequency diagram to illustrate the data. **[5]**

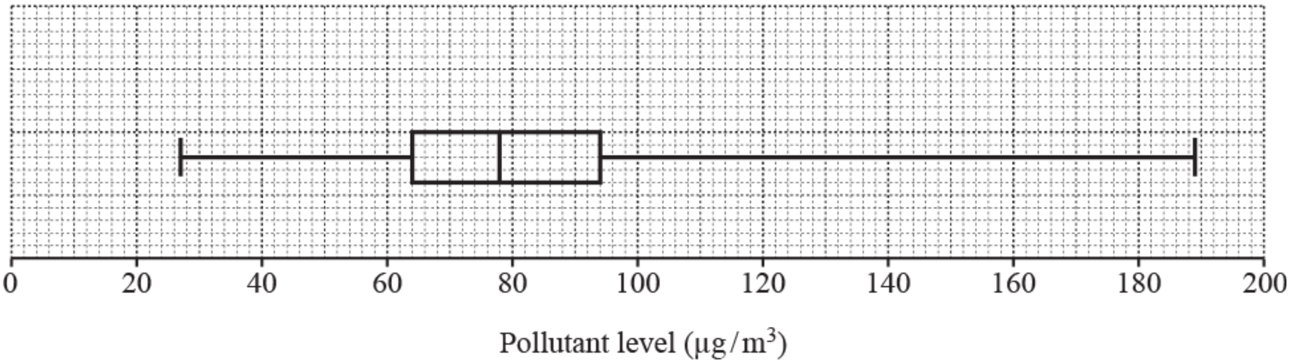
(ii) Levels of nitrogen dioxide below 200 are classified as low. Estimate the proportion of days on which the level was low. [2]

(iii) Use your diagram to estimate the median and interquartile range of the data. [3]

(iv) For each end of the distribution, explain whether outliers definitely exist, may possibly exist or definitely do not exist. [4]

(v) Draw a box and whisker plot to illustrate the data. [3]

The box and whisker plot below shows similar data for a roadside location in Tower Hamlets in London.



(vi) Compare the skewness of the data from the two locations. [2]

19. Doug has a list of times taken by competitors in a ‘fun run’. He has grouped the data and calculated the frequency densities in order to draw a histogram to represent the information. Some of the data are presented in Fig. 2.

Time in minutes	15 –	20 –	25 –	35 –	45 – 60
Number of runners	12	23	59	71	
Frequency density	2.4		5.9	7.1	1.4

Fig. 2

(a) Write down the missing values in Fig. 2 above. **[2]**

(b) Doug labels the horizontal axis on the histogram ‘time in minutes’ and the vertical axis ‘number of minutes per runner’. State which one of these labels is incorrect and write down a correct version. **[1]**

20. Rose and Emma each wear a device that records the number of steps they take in a day. All the results for a 7-day period are given in Fig. 7.

Day	1	2	3	4	5	6	7
Rose	10 014	11 262	10 149	9361	9708	9921	10 369
Emma	9204	9913	8741	10 015	10 261	7391	10 856

Fig. 7

The 7-day mean is the mean number of steps taken in the last 7 days. The 7-day mean for Rose is 10 112.

(a) Calculate the 7-day mean for Emma. **[1]**

At the end of day 8 a new 7-day mean is calculated by including the number of steps taken on day 8 and omitting the number of steps taken on day 1. On day 8 Rose takes 10 259 steps.

(b) Determine the number of steps Emma must take on day 8 so that her 7-day mean at the end of day 8 is the same as for Rose. **[4]**

In fact, over a long period of time, the mean of the number of steps per day that Emma takes is 10 341 and the standard deviation is 948.

(c) Determine whether the number of steps Emma needs to take on day 8 so that her 7-day mean is the same as that for Rose in part **(b)** is unusually high. **[3]**

21. The pre-release material contains data concerning the death rate per thousand people and the birth rate per thousand people in all the countries of the world. The diagram in Fig. 11.1 was generated using a spreadsheet and summarises the birth rates for all the countries in Africa.

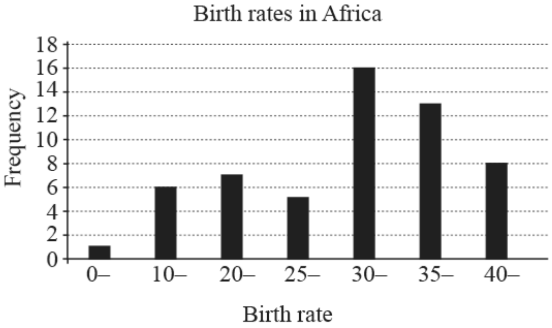


Fig. 11.1

(a) Identify **two** respects in which the presentation of the data is incorrect. [2]

Fig. 11.2 shows a scatter diagram of death rate, y, against birth rate, x, for a sample of 55 countries, all of which are in Africa. A line of best fit has also been drawn.

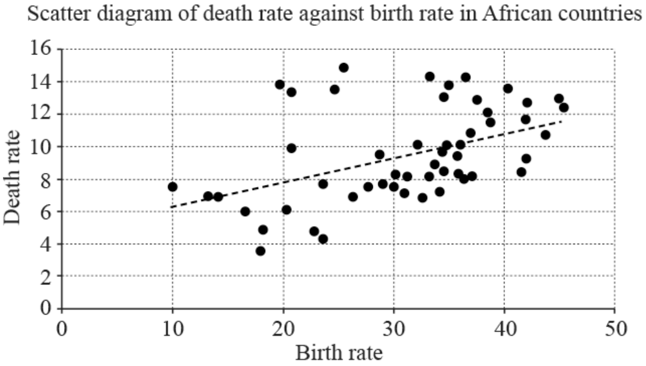


Fig. 11.2

The equation of the line of best fit is $y = 0.15x + 4.72$.

(b) (i) What does the diagram suggest about the relationship between death rate and birth rate? [1]

(ii) The birth rate in Togo is recorded as 34.13 per thousand, but the data on death rate has been lost. Use the equation of the line of best fit to estimate the death rate in Togo. [1]

(iii) Explain why it would not be sensible to use the equation of the line of best fit to estimate the death rate in a country where the birth rate is 5.5 per thousand. [1]

(iv) Explain why it would not be sensible to use the equation of the line of best fit to estimate the death rate in a Caribbean country where the birth rate is known. [1]

(v) Explain why it is unlikely that the sample is random. [1]

Including Togo there were 56 items available for selection.

(c) Describe how a sample of size 14 from this data could be generated for further analysis using systematic sampling. [2]

22. A survey of the number of cars per household in a certain village generated the data in Fig. 4.

Number of cars	0	1	2	3	4
Number of households	8	22	31	27	7

Fig. 4

(a) Calculate the mean number of cars per household. [1]

(b) Calculate the standard deviation of the number of cars per household. [1]

23. At the end of each school term at North End College all the science classes in year 10 are given a test. The marks out of 100 achieved by members of set 1 are shown in Fig. 9.

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

Key 5 | 2 represents a mark of 52

Fig. 9

(a) Describe the shape of the distribution. [1]

(b) The teacher for set 1 claimed that a typical student in his class achieved a mark of 95. How did he justify this statement? [1]

(c) Another teacher said that the average mark in set 1 is 76. How did she justify this statement? [1]

Benson's mark in the test is 35. If the mark achieved by any student is an outlier in the lower tail of the distribution, the student is moved down to set 2.

(d) Determine whether Benson is moved down to set 2. [2]

24. The stem–and–leaf diagram in Fig. 7 shows the numbers of customers at a village post office on the days it was open in March 2017.

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

Key 4|0 represents 40 customers

Fig. 7

(a) Describe the shape of the distribution. [1]

(b) Draw a box plot to represent the data. [4]

25. At the start of the January term year 11 students at Amplesides College sat mock examinations in GCSE mathematics papers 4 and 5. A teacher collected the results and presented them in a scatter diagram, which is shown in Fig. 10. A line of best fit has been added.

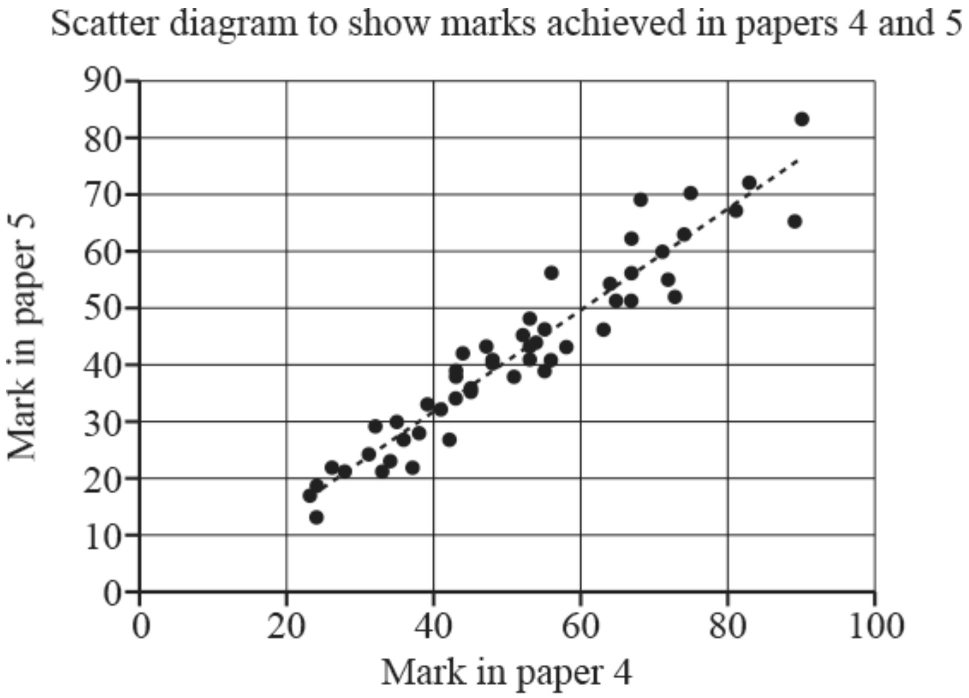


Fig. 10

The correlation coefficient for the data is 0.9566.

(a) Give **two** reasons why it is reasonable to model the relationship between the mark achieved in paper 4 and the mark achieved in paper 5 by a straight line.

[2]

The equation of the line of best fit is $y = 0.89x - 3.76$, where y is the mark achieved in paper 5 and x is the mark achieved in paper 4.

(b) Tina achieved a mark of 83 in paper 4, but was absent for paper 5. Calculate an estimate of the mark she would have achieved in paper 5.

[1]

(c) Dave was absent for paper 4. He achieved a mark of 8 in paper 5. Calculate an estimate of the mark Dave would have achieved in paper 4.

[2]

(d) Explain why the estimate of Tina's mark for paper 5 is more reliable than the estimate of Dave's mark for paper 4. [1]

26. The spreadsheet output in Fig. 12.1 gives some information about all the countries that won more than 10 gold medals in the London 2012 Olympic Games.

A	B	C	D	E	F	G	H
1	Country	population	GDP per capita (US$)	Total GDP (US$)	Gold medals	Silver medals	Bronze medals
2	United States	318 892 103	52 800	1.68×10^{13}	46	29	29
3	China	1 355 692 576	9800	1.33×10^{13}	38	27	22
4	United Kingdom	63 742 977	37 300	2.38×10^{12}	29	17	19
5	Russia	142 470 272	18 100	2.58×10^{12}	24	25	33
6	Korea, South	49 039 986	33 200	1.63×10^{12}	13	8	7
7	France	66 259 012	35 700	2.37×10^{12}	11	11	12
8	Germany	80 996 685	39 500	3.20×10^{12}	11	19	14

Fig. 12.1

(a) Give a spreadsheet formula for calculating the value in cell D2 using other cell values in Fig. 12.1. [1]

The statistics in Fig. 12.2 are for all the countries in the pre-release data.

	Population	GDP per capita (US$)	Total GDP (US$)
Lower Quartile	4.587×10^{5}	4525	2.09×10^{9}
Median	5.623×10^{6}	13750	7.73×10^{10}
Upper Quartile	2.116×10^{7}	31750	6.72×10^{10}

Fig. 12.2

(b) Explain whether or not the statements below are consistent with the information given in Fig. 12.1 and Fig 12.2.

Statement A
Countries with larger populations are more likely to win Olympic gold medals.

Statement B
Countries with larger total GDP are more likely to win Olympic gold medals. [3]

There were approximately 10 500 Olympic competitors in 2012. The population of the world was approximately
7 000 000 000 in 2012.

A geography student assumed that Olympic competitors are randomly and uniformly scattered across the population of the world. The student calculated that the population of a country with two Olympic competitors would be approximately 1 300 000.

(c) Use your knowledge of the large data set to explain whether the geography student's assumption is realistic. [1]

27. Qasim has just opened a café in Burnton. He decides to conduct some market research on his customers.

One Monday morning at 11 am he asked every customer in the café to fill out a questionnaire which included asking for the customer's age. The results he collected are shown in the stem-and-leaf diagram in Fig. 7.1.

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

Key 2 | 0 represents an age of 20.

Fig. 7.1

(a) What name is given to the sampling method used by Qasim? [1]

(b) Explain why the sample does not represent a simple random sample of all the customers who use Qasim's café. [1]

(c) Describe the shape of the distribution of the ages in the sample. [1]

(d) For the data in Fig. 7.1 find

- the median,
- the interquartile range.

[3]

Kai works at Qasim's café. He believes that Qasim's sample data may not be representative of all Qasim's customers. One week he asks every customer to fill in the questionnaire. Summary statistics for the customers who filled in the questionnaire are shown in Fig. 7.2.

Number of respondents	237
Age of youngest person	7
Lower quartile	27
Median	31
Upper quartile	39
Age of oldest person	81

Fig. 7.2

(e) Comment on whether the statistics in Fig. 7.2 provide any evidence to support Kai's belief. [2]

28. The large data set (LDS1) provides information on average life expectancy at birth for countries of the world.
Fig. 13.1 shows the entry for South Sudan, in Africa.

Country	life expectancy at birth
South Sudan	#N /A

Fig. 13.1

No data concerning average life expectancy at birth is available for South Sudan.

(a) Explain why the spreadsheet entry is #N /A instead of simply being left blank. [1]

Summary statistics for the values given for average life expectancy at birth for all the countries in Africa apart from South Sudan were generated using software. These are shown in Fig. 13.2.

Mean	61.21418
Standard Deviation (s)	7.83837
Lowest Score	49.81
Highest Score	79.36
Distribution Range	29.55
Total Number of Scores	55
Number of Distinct Scores	54
Lowest Class Value	45
Highest Class Value	79.99
Number of Classes	7
Class Range	5

Fig. 13.2

(b) Explain why the mean of 61.21418 may not represent a good estimate of the average life expectancy at birth of all people in Africa. [1]

Fig. 13.3 shows a frequency diagram of average life expectancies of countries in Africa, excluding South Sudan.

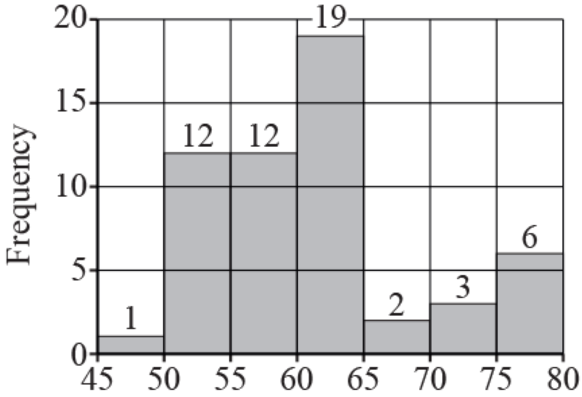


Fig. 13.3

(c) Explain why it would be incorrect to call the diagram in Fig. 13.3 a histogram. [1]

(d) Draw a histogram to represent the data, using class boundaries at 45, 55, 60, 65 and 80. [3]

Fig. 13.4 shows a scatter diagram of median age in years against average life expectancy for countries in Africa excluding South Sudan. This was generated using a spreadsheet.

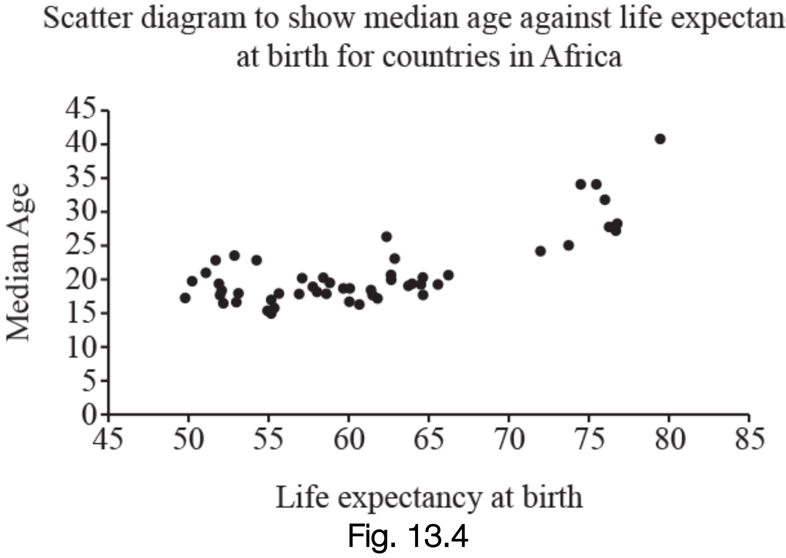
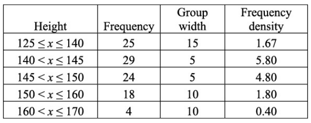


Fig. 13.4

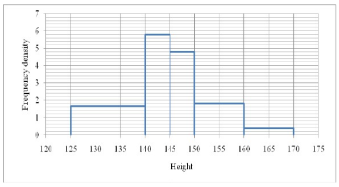
The median age for the population of South Sudan is given as 17.

(e) With reference to Fig. 13.4, comment on whether it might be possible to obtain a reliable estimate of the average life expectancy at birth for South Sudan. [2]

END OF QUESTION paper

<table>
<tr><th colspan="3">Question</th><th>Answer/Indicative content</th><th>Marks</th><th colspan="2">Part marks and guidance</th></tr>
<tr><td>1</td><td></td><td>i</td><td>4 + ½ of 18 = 4 + 9 = 13</td><td>M1</td><td>For ½ of 18</td><td></td></tr>
<tr><td></td><td></td><td>i</td><td></td><td>A1</td><td>cao
Examiner's Comments
On the whole, this question was answered well. The most common incorrect answer was 22, which was seen fairly frequently. A small minority of candidates wrote 9 + 4, but then calculated incorrectly (both 11 and 12 seen).</td><td>13/100 gets M1A0</td></tr>
<tr><td></td><td></td><td>ii</td><td>(Median) = 50.5th value</td><td>M1</td><td>For 50.5 seen</td><td>SC2 for use of 50th value leading to Est = 140 + (25/29 × 5) = 144.3 (SC1 if over-specified)</td></tr>
<tr><td></td><td></td><td>ii</td><td>$\text{Est} = 140 + \left(\frac{25.5}{29}\right) \times 5 \quad \text{or} = 140 + \left(\frac{50.5 - 25}{54 - 25}\right) \times 5$</td><td>M1</td><td>For attempt to find this value</td><td>$\text{or Est} = 145 - \left(\frac{3.5}{29}\right) \times 5 = 144.4$</td></tr>
<tr><td></td><td></td><td>ii</td><td>= 144.4</td><td>A1</td><td>**Examiner's Comments**
Only about 10% of candidates produced totally correct answers. Many scored SC2 for finding the 50th, rather than 50.5th, value. Those that did state that they were looking for 50.5th value often just gave the mid-value, rather than using interpolation. Many candidates lost a mark due to over-specification.</td><td>NB no marks for mean = 144.35
NB Watch for over-specification</td></tr>
<tr><td></td><td></td><td>iii</td><td>
<table>
<tr><th>Height</th><th>Frequency</th><th>Group width</th><th>Frequency density</th></tr>
<tr><td>$125 \leq x \leq 140$</td><td>25</td><td>15</td><td>1.67</td></tr>
<tr><td>$140 < x \leq 145$</td><td>29</td><td>5</td><td>5.80</td></tr>
<tr><td>$145 < x \leq 150$</td><td>24</td><td>5</td><td>4.80</td></tr>
<tr><td>$150 < x \leq 160$</td><td>18</td><td>10</td><td>1.80</td></tr>
<tr><td>$160 < x \leq 170$</td><td>4</td><td>10</td><td>0.40</td></tr>
</table>

**</td><td>M1</td><td>For fd's - at least 3 correct</td><td>M1 can be also be gained from freq per 10 – 16.7, 58, 48, 18, 4 (at least 3 correct) or freq per 5 – 8.35, 29, 24, 9, 2 for all correct.</td></tr>
</table>

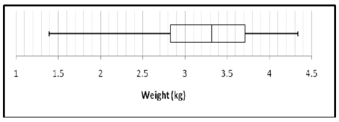
Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii		A1	Accept any suitable unit for fd such as eg freq per cm. correct to at least one dp allow 1.66 but not 1.6 for first fd	If fd not explicitly given, M1 A1 can be gained from all heights correct (within one square) on histogram (and M1A0 if at least 3 correct)
		iii		G1	linear scales on both axes and label on vertical axis	Linear scale and label on vertical axis IN RELATION to first M1 mark ie fd or frequency density or if relevant freq / 10, etc (NOT eg fd / 10). However allow scale given as fd × 10, or similar Accept f/w or f/cw (freq / width or freq / class width) Can also be gained from an accurate key G0 if correct label but not fd's.
		iii		W1	width of bars	Must be drawn at 125, 140 etc NOT 124.5 or 125.5 etc NO GAPS ALLOWED Must have linear scale. No inequality labels on their own such as 125≤S<140, etc but allow if a clear horizontal linear scale is also given. Ignore horizontal label.

<table>
<tr><th colspan="3">Question</th><th>Answer/Indicative content</th><th>Marks</th><th colspan="2">Part marks and guidance</th></tr>
<tr><td></td><td></td><td>iii</td><td></td><td>H1</td><td>height of bars
<b>Examiner's Comments</b>
The histogram was generally completed rather better than in previous years. Most candidates were able to calculate frequency densities correctly, and they also usually labelled the axes correctly. A fairly common error was to round the first frequency density down to 1.6 rather than to 1.7. Some made errors with careless drawing of bars, making slips with incorrect heights.</td><td>Height of bars – must be linear vertical scale.
FT of heights dep on at least 3 heights correct and all must agree with their fds If fds not given and at least 3 heights correct then max M1A0G1W1H0 Allow restart with correct heights if given fd wrong (for last three marks only)</td></tr>
<tr><td></td><td></td><td>iv</td><td>4 boys</td><td>3</td><td></td><td></td></tr>
<tr><td></td><td></td><td>iv</td><td>0.6 × 15</td><td>M1</td><td>For 0.6 × 15</td><td>Or 45 × 0.2 = 9 (number of squares and 0.2 per square)</td></tr>
<tr><td></td><td></td><td>iv</td><td>= 9 girls</td><td>A1</td><td>For 9 girls</td><td></td></tr>
<tr><td></td><td></td><td>iv</td><td>So 5 more girls</td><td>A1</td><td>cao
<b>Examiner's Comments</b>
Roughly 90% of candidates scored full marks here.</td><td></td></tr>
<tr><td></td><td></td><td>v</td><td>Frequencies and midpoints for girls are<table><tr><td>Height</td><td>132.5</td><td>142.5</td><td>147.5</td><td>155</td><td>167.5</td></tr><tr><td>Frequency</td><td>18</td><td>23</td><td>31</td><td>19</td><td>9</td></tr></table></td><td>B1</td><td>For at least three frequencies correct</td><td></td></tr>
<tr><td></td><td></td><td>v</td><td></td><td>B1</td><td>At least three midpoints correct</td><td>No further marks if not using midpoints</td></tr>
<tr><td></td><td></td><td>v</td><td>So mean =</td><td>M1</td><td>For attempt at Σ<i>xf</i></td><td>For sight of at least 3 <i>xf</i> pairs</td></tr>
<tr><td></td><td></td><td>v</td><td>$\frac{(132.5\times18)+(142.5\times23)+(147.5\times31)+(155\times19)+(167.5\times9)}{100}$
$=\frac{(2385)+(3277.5)+(4572.5)+(2945)+(1507.5)}{100}$</td><td>M1* Dep on M1</td><td>For division by 100</td><td></td></tr>
</table>

Question			Answer/Indicative content	Marks	Part marks and guidance	
		v		A1	Cao NB Watch for over-specification **Examiner's Comments** Many candidates did everything correctly but gave the final answer as 146.875 or 146.88 thus losing the final mark for over-specifying. The scheme allowed for a slip in both the frequencies and the mid-points and candidates were still able to gain 4 marks. The most common error was giving the final mid-point as 165 rather than 167.5.	Allow answer 146.9 or 147 but not 150 NB Accept answers seen without working (from calculator) Use of 'not quite right' midpoints such as 132.49 or 132.51 etc can get B1B0M1M1A0
			Total	18		

Question			Answer/Indicative content	Marks	Part marks and guidance	
2		i	Positive	B1	CAO **Examiner's Comments** Approximately 95% of candidates scored this mark.	
		ii	Mean = 5.064 allow 5.1 with working 126.6/25 or 5.06 without	B1		
		ii	SD = 1.324 allow 1.3 with working or 1.32 without	B2	Allow B1 for RMSD = 1.297 or var = 1.753 or MSD = 1.683 **Examiner's Comments** There were many fully correct answers. Most used the relevant formulas rather than using the built in functions on their calculators. A few candidates found the variance or the rmsd, and these gained a method mark. The most common error was not to use the key and thus get answers ten times too high. This error was severely penalised, but full marks were allowed for a follow through.	Also allow B1 for S*xx* = 42.08 or for Σx^2 = 683 SC1 for both mean = 50.64 and SD = 13.24 (even if over-specified)
		iii	$\bar{x} - 2s = 5.064 - 2\times 1.324 = 2.416$	B1FT	FT their mean and sd	For use of quartiles and IQR Q_1 = 3.95; Q_3 = 6.0; IQR = 2.05 3.95 – 1.5(2.05) gets M1 Allow other sensible definitions of quartiles
		iii	$\bar{x} + 2s = 5.064 + 2\times 1.324 = 7.712$	M1	for $\bar{x} + 2s$ but withhold final E mark if their limits mean that there are no outliers.	6.0 + 1.5(2.05) gets M1
		iii		A1FT	For upper limit	Limits 0.875 and 9.075

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	So there is an outlier.	E1	Incorrect statement such as 7.6 and 8.1 are outliers gets E0 Do not award E1 if calculation error in upper limit **Examiner's Comments** The limits for outliers were widely known and correctly used by most candidates. Even those with incorrect mean and standard deviation were able to gain 3 or all 4 marks if they followed through correctly. Some candidates used the quartiles method, despite often having got part (ii) correct, but some of these made errors, losing some if not all of the marks.	So there are no outliers NB do not penalise over-specification here as not the final answer but just used for comparison. FT from SC1
			Total	8		

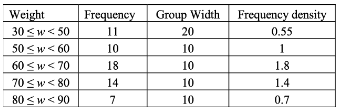
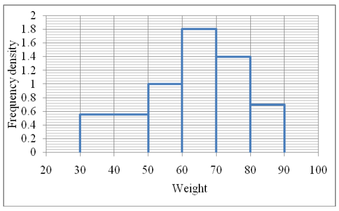
Question			Answer/Indicative content	Marks	Part marks and guidance	
3		i	Median = 3.32 kg	B1		
		i	Q1 (= 6.5th value) = 2.83 Q3 (= 19.5th value) = 3.71	B1	For Q1 or Q3	For Q1 allow 2.82 to 2.84
		i	Inter-quartile range = 3.71 – 2.83 = 0.88	B1	For IQR dep on both quartiles correct Examiner's Comments Most candidates successfully found the median, although instead of the 13th value some found average of the 12th and 13th values. However, candidates were less successful in finding the interquartile range. The lower quartile was usually found correctly, but the upper quartile was more frequently wrong, with an answer of 3.665 being the most common error. Occasionally candidates did not subtract to find the interquartile range, but instead some found the midpoint of their quartiles.	For Q3 allow 3.70 to 3.72 If no quartiles given allow B0B1 for IQR in range 0.86 to 0.90
		ii		G1	For reasonably linear scale shown.	Dep on attempt at box and whisker plot with at least a box and one whisker. Condone lack of label.
		ii		G1	For boxes in approximately correct positions, with median just to right of centre	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii		G1	For whiskers in approximately correct positions in proportion to the box FT their median and quartiles if sensible – guidance above is only for correct values <u>Examiner's Comments</u> The response to this question was very disappointing. Perhaps because they were faced with a blank space rather than graph paper, most candidates thought that accuracy was not required. Very few had a scale and some of those that did failed to make it linear. Some candidates simply sketched a box and whisker plot and then labelled the diagram with the relevant values. This did not gain marks as the question clearly instructs candidates to 'Draw a box and whisker plot…'. It seems likely that many candidates either did not have, or did not think to use a ruler. Far too many freehand diagrams were seen, with the sizes of the box and whiskers and the position of the median not in proportion.	Do not award unless RH whisker significantly shorter than LH whisker Allow LH whisker going to 2.5 and outlier marked at 1.39
		iii	Lower limit 2.83 – (1.5 × 0.88) = 1.51	B1	For 1.51 FT	Any use of median ± 1.5 × IQR scores B0 B0 E0 No marks for ± 2 or 3 × IQR In this part FT their values from (i)or (ii) if sensibly obtained but not from location ie 6.5, 19.5

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	Upper limit 3.71 + (1.5 × 0.88) = 5.03	B1	For 5.03 FT	Do not penalise over-specification as not the final answer
		iii	Exactly one baby weighs less than 1.51 kg and none weigh over 5.03 kg so there is exactly one outlier.	E1*	Dep on their 1.51 and 5.03	Do not allow unless FT leads to upper limit above 4.34 and lower limit between 1.39 and 2.50
		iii	‘Nothing to suggest that this baby is not a genuine data value so she should not be excluded’ or ‘This baby is premature and therefore should be excluded’.	E1* Dep	Any sensible comment in context <u>Examiner's Comments</u> Many candidates correctly found the upper and lower limits for the outliers. The most common misconception was that outliers were calculated using median ± 1.5×IQR, although many other errors were also seen. A few candidates attempted to use the mean and standard deviation, and if they got both of these correct, full marks were available, but unfortunately one or other of the two statistics was usually incorrect. It was necessary to check both limits to show that there was only one outlier, but some candidates ignored the upper limit. Many candidates failed to give an explanation in context regarding the outlier, though those that did often made a valid point about premature babies.	For use of mean ± 2sd allow B1 For 3.27 + 2 × 0.62= 4.51 B1 For 3.27 - 2 × 0.62= 2.03 Then E1E1 as per scheme
		iv	Median = 3.5 kg	B1		
		iv	Q1 = 50th value = 3.12 Q3 = 150th value = 3.84	B1	For Q1 or Q3	For Q1 allow 3.11 to 3.13 For Q3 allow 3.83 to 3.85

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iv	Inter-quartile range = 3.84 – 3.12 = 0.72	B1	For IQR FT their quartiles **Examiner's Comments** As in part (i), the median was usually found correctly, but some candidates lost a mark due to inaccurate reading of the scales in finding the quartiles.	Dep on both quartiles correct If no quartiles given allow B0B1 for IQR in range 0.70 to 0.74
		v	Female babies have lower weight than male babies on the whole	E1	Allow 'on average' or	Do not allow lower median
		v		FT	similar in place of 'on the whole'	
		v	Female babies have higher weight variation than male babies	E1	Allow 'more spread' or	Do not allow higher IQR, but SC1 for
		v		FT	similar but not 'higher range' Condone less consistent **Examiner's Comments** Only about one third of candidates scored both marks. Credit was given to those candidates who could only compare medians and interquartile ranges without an explanation of what they meant. Candidates who just said 'boys are heavier' failed to get credit without a comment such as 'generally' or 'on average' or 'tend to be'. Similarly 'more consistent' or 'vary less' or 'less spread' gained credit for interquartile range – 'smaller range' was not awarded credit.	both lower median and higher IQR, making clear which is which

Question			Answer/Indicative content	Marks	Part marks and guidance	
		vi	Male babies must weigh more than 4.34 kg			
		vi	Approx 10 male babies weigh more than this.	M1*	For 10 or 9 or 8	Or 200 – 190, 200 –191 or 200 –192
		vi	Probability $= \frac{10}{200} \times \frac{9}{199} = \frac{90}{39800} = \frac{9}{3980} = 0.00226$ or $\frac{9}{200} \times \frac{8}{199} = \frac{72}{39800} = 0.00181$	M1* dep	For first fraction multiplied by any other different fraction (Not a binomial probability)	Allow any of these answers For spurious factors, eg 2 × correct answer allow M1M1A0
		vi	or $\frac{8}{200} \times \frac{7}{199} = \frac{56}{39800} = \frac{7}{4975} = 0.00141$	A1	CAO Allow their answer to min of 2 sf <u>Examiner's Comments</u> This part discriminated very well between the higher-scoring candidates. Many candidates realised that approximately 10 male babies weighed more than 4.34 kg. Unfortunately many then did not know how to proceed, often squaring 0.05 (10/200) rather than multiplying by 9/199. Those candidates who misread the scale but knew how to proceed could gain a Special Case mark. A significant number of candidates missed out this part altogether.	SC1 for *n*/200 × (*n*–1)/199 **NOTE RE OVER-SPECIFICATION OF ANSWERS** If answers are grossly over-specified, deduct the final answer mark in every case. Probabilities should also be rounded to a sensible degree of accuracy. In general final non probability answers should not be given to more than 4 significant figures. Allow probabilities given to 5 sig fig. PLEASE HIGHLIGHT ANY OVER-SPECIFICATION Please note that there are no G or E marks in scoris, so use B instead
			Total	18		

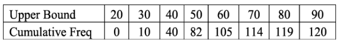
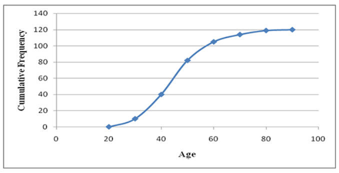
<table>
<tr><th colspan="3">Question</th><th>Answer/Indicative content</th><th>Marks</th><th colspan="2">Part marks and guidance</th></tr>
<tr><td>4</td><td></td><td>i</td><td>
<table>
<tr><th>Weight</th><th>Frequency</th><th>Group Width</th><th>Frequency density</th></tr>
<tr><td>$30 \le w < 50$</td><td>11</td><td>20</td><td>0.55</td></tr>
<tr><td>$50 \le w < 60$</td><td>10</td><td>10</td><td>1</td></tr>
<tr><td>$60 \le w < 70$</td><td>18</td><td>10</td><td>1.8</td></tr>
<tr><td>$70 \le w < 80$</td><td>14</td><td>10</td><td>1.4</td></tr>
<tr><td>$80 \le w < 90$</td><td>7</td><td>10</td><td>0.7</td></tr>
</table></td><td>M1</td><td>For fd's – at least 3 correct Accept any suitable unit for fd such as eg freq per 10g.</td><td>M1 can be also be gained from freq per 10 – 5.5, 10, 18, 14, 7 (at least 3 correct) or similar.
If fd not explicitly given, M1
A1 can be gained from all heights correct (within half a square) on histogram (and M1A0 if at least 3 correct)</td></tr>
<tr><td></td><td></td><td>i</td><td></td><td>A1</td><td></td><td>Linear scale and label on vertical axis IN RELATION to first M1 mark ie fd or frequency density or if relevant freq/10, etc (NOT eg fd/10).</td></tr>
<tr><td></td><td></td><td>i</td><td>
</td><td>G1</td><td>linear scales on both axes and labels
Vertical scale starting from zero (not broken – but can get final mark for heights if broken)</td><td>However allow scale given as fd × 10, or similar.
Accept f/w or f/cw (freq/width or freq/class width)
Ignore horizontal label
Can also be gained from an accurate key
G0 if correct label but not fd's.</td></tr>
<tr><td></td><td></td><td>i</td><td></td><td>G1</td><td>width of bars</td><td>Must be drawn at 30, 50 etc NOT 29.5 or 30.5 etc NO GAPS
ALLOWED
Must have linear scale.
No inequality labels on their own such as 30 ≤W<50, 50 ≤W<60 etc but allow if 30, 50, 60 etc occur at the correct boundary position.
See additional notes.
Allow this mark even if not using fd's</td></tr>
</table>

Question			Answer/Indicative content	Marks	Part marks and guidance	
		i		G1	height of bars **Examiner's Comments** Most candidates found the frequency densities correctly. They usually then went on to draw the axes correctly although a few failed to start the frequency density scale at zero or to label the axes. A few candidates used inequalities on the horizontal axis, which attracted a penalty of one mark. The choice of scales on the vertical axis was not always ideal, and this left some candidates vulnerable to drawing the heights at incorrect positions. In particular the height of the first bar was frequently incorrectly plotted at 0.5 rather than 0.55.	Height of bars – must be linear vertical scale. FT of heights dep on at least 3 heights correct and all must agree with their fds If fds not given and at least 3 heights correct then max M1A0G1G1G0 Allow restart with correct heights if given fd wrong (for last three marks only)
		ii	Mean= $\frac{(40\times11)+(55\times10)+(65\times18)+(75\times14)+(85\times7)}{60}=\frac{3805}{60}$	M1	For midpoints Products are 440, 550, 1170, 1050, 595	For midpoints (at least 3 correct) No marks for mean or sd unless using midpoints

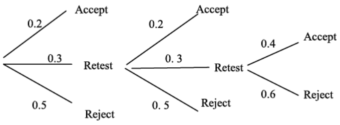
Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii	= 63.4 (or 63.42)	A1	CAO (exact answer 63.41666…) **Examiner's Comments** The calculation of the mean of the grouped data was in most cases accurately performed using correct mid-points. The calculation of the standard deviation was less well executed. Whilst there were many correct solutions seen, some forgot to factor in the frequencies and worked with Σ x^2 rather than Σ fx^2. Over specification of either or both of the answers caused some candidates to lose one mark.	Answer must NOT be left as improper fraction as this is an estimate Accept correct answers for mean and sd from calculator even if eg wrong Sxx given
		ii	$\sum x^2 f = (40^2 \times 11) + (55^2 \times 10) + (65^2 \times 18) + (75^2 \times 11) + (85^2 \times 7)$ = 253225			
		ii	$S_{xx} = 253225 - \dfrac{3805^2}{60} = 11924.6$	M1	For attempt at S_{xx} Should include sum of at least 3 correct multiples $fx^2 - \sum x^2/n$	Allow M1 for anything which rounds to 11900
		ii	$s = \sqrt{\dfrac{11924.6}{59}} = \sqrt{202.11} = 14.2$		At least 1dp required Use of mean 63.4 leading to answer of 14.29199.. with S_{xx} = 12051.4 gets full credit.	Allow SC1 for RMSD 14.1 (14.0976…) from calculator.
		ii		A1		Only penalise once in part (ii) for over specification, even if mean and standard deviation both over specified.
		ii			63.42 leads to 14.2014… Do not FT their incorrect mean (exact answer14.2166…)	If using $(x - \bar{x})^2$ method, B2 if 14.2 or better (14.3 if use of 63.4), otherwise B0

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	$\bar{x}$– 2*s* = 63.4 – (2 × 14.2) = 35	M1	For either No marks in (iii) unless usin $\bar{x}$ + 2*s* or $\bar{x}$ – 2*s*	FT their positive mean and their positive sd / rmsd for M1A1. Only follow through numerical values, not variables such as *s*, so if a candidate does not find *s* but then writes here 'limit is 63.4 + 2 × standard deviation', do NOT award M1
		iii	$\bar{x}$+ 2*s* = 63.4 + (2 × 14.2) = 91.8	A1	For both (FT)	Do not penalise for over-specification
		iii	So there are probably some outliers at the lower end, but none at the upper end	E1	Must include an element of doubt and must mention both ends **Examiner's Comments** Most candidates scored at least the first two marks. However many omitted the fact that there were definitely no outliers at the top end of the data and/or stated that there were definitely some outliers present at the bottom end, thus missing the final mark.	Must have correct limits to get this mark
		iv	$\text{Mean} = \dfrac{3624.5}{50} = 72.5\text{g}$ (or exact answer 72.49g)	B1	CAO Ignore units	
		iv	$S_{xx} = 265416 - \dfrac{3624.5^2}{50} = 2676$	M1	For S_{xx}	M1 for 265416 – 50 × their mean^2 BUT NOTE M0 if their S_{xx} < 0

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iv	$s = \sqrt{\frac{2676}{49}} = \sqrt{54.61} = 7.39\text{g}$	A1	CAO ignore units Allow 7.4 but NOT 7.3 (unless RMSD with working) **Examiner's Comments** This was generally very well answered.	For s^2 of 54.6 (or better) allow M1A0 with or without working. For RMSD of 7.3 (or better) allow M1A0 provided working seen For RMSD^2 of 53.5 (or better) allow M1A0 provided working seen
		v	Variety A have lower average than Variety B oe	E1	FT their means Do not condone lower central tendency or lower mean	Allow 'on the whole' or similar in place of 'average'.
		v	Variety A have higher variation than Variety B oe	E1	FT their sd **Examiner's Comments** For this type of question candidates should be taught to discuss 'average' and 'variation'. Simply stating for example that the mean of A is lower than the mean of B does not attract any credit.	Allow 'more spread' or similar but not 'higher range' or 'higher variance' Condone less consistent.
			Total	17		

Question			Answer/Indicative content	Marks	Part marks and guidance	
5		i	 Upper Bound | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 Cumulative Freq | 0 | 10 | 40 | 82 | 105 | 114 | 119 | 120	B1	Cumulative frequencies All correct	May be implied from graph. Condone omission of 0 at this stage.
		i		G1	For plotted points (Provided plotted at correct UCB positions)	Plotted as (UCB, their cf). Ignore (20,0) at this stage. No midpoint or LCB plots. Plotted within ½ small square If cf not given then allow G1 for good attempt at cf. e.g. if they have 0,10,40,72,95,104,109,110
		i		G1	For joining points (within ½ a square)	For joining all of 'their points' (line or smooth curve) AND now including (20,0) Not for midpoint or LCB plots.
		i		G1	For scales	Linear horizontal scale. Allow if start at 30 (no inequality scales - Not even <20, <30, <40 …) Linear vertical scale Allow full credit if axes reversed correctly

Question			Answer/Indicative content	Marks	Part marks and guidance	
		i		G1	For labels All marks dep on good attempt at cumulative frequency, but not cumulative fx's or other spurious values. **Examiner's Comments** Many candidates gained full credit. A common error which resulted in the loss of 2 marks was to plot the correct height but at mid-points. Only a few used the lower class boundaries. Some candidates drew cumulative frequency bars and a small number just plotted frequency against midpoints. Some candidates forgot to label their axes or more often omitted the word "cumulative" on their vertical axis.	Age or *x* and Cumulative frequency or just CF or similar but not just frequency or fd nor cumulative fd Mid-point or LCB plots may score first and last two marks Can get up to 3/5 for cum freq bars Lines of best fit could attract max 4 out of 5.
		ii	Median = 45	B1	Allow answers between 44 and 46 without checking curve. Otherwise check curve. No marks if not using diagram.	Based on 60th value ft their curve (not LCB's) Allow 40 for m.p. plot without checking graph B0 for interpolation If max value wrong (eg 110) FT their max value for all 3 marks
		ii	Q1 = 37 Q3 = 53	B1	For Q3 or Q1 Allow Q1 between 37 and 38 without checking Allow Q3 between 52 and 54 without checking	Based on 30th and 90th values ft their curve (not LCB's) Allow Q1 = 32; Q3 = 48 without checking graph

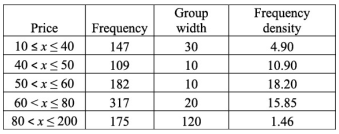
Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii	Inter-quartile range = 53 – 37 = 16	B1	For IQR providing both Q1 and Q3 are correct	B0 for interpolation B2 for correct IQR from graph if quartiles not stated but indicated on graph Allow from mid-point plot Must be good attempt at cumulative frequency in part (i) to score any marks here Lines of best fit: B0 B0 B0 here. Also cumulative frequency bars: B0 B0 B0 here
		ii	 0.2 Accept 0.3 Retest 0.5 Reject 0.2 Accept 0. 3 Retest 0. 5 Reject 0.4 Accept 0.6 Reject		Alternative version of tree diagram for Q2(i) **Examiner's Comments** This part was very well answered with many candidates picking up the follow through marks for correctly identifying the median and quartiles from their mid-point plotted graph.	
			Total	8		

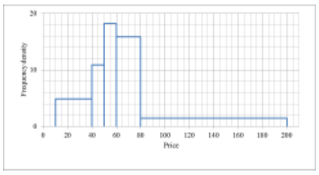
Question			Answer/Indicative content	Marks	Part marks and guidance	
6		i	0 | 6 1 | 5 8 2 | 1 5 8 3 | 1 1 3 5 8 9 Key 1 | 8 represents 18 people	G1	Stem (in either order) and leaves	Do not allow leaves 21 ,25, 28 etc Ignore commas between leaves Allow stem 0, 10, 20, 30
		i		G1	Sorted and aligned	Allow errors in leaves if sorted and aligned. Use paper test if unsure about alignment – hold a piece of paper vertically and the columns of leaves should all be separate. Alternatively place a pencil vertically over each column. If any figures protrude then deem this as non-alignment. Highlight this error
		i		G1	Key **Examiner's Comments** Most candidates scored all three marks, although some did not accurately align the leaves or did not provide a suitable key and thus scored only 2 marks. Very few candidates scored less than 2 out of 3.	
		ii	Negative	B1	**Examiner's Comments** This was very well-answered with only a few thinking that the skew was positive.	Allow -ve but NOT skewed to the left Do not allow 'negative correlation'
		iii	Median = 29.5	B1	CAO	
		iii	Mean = 26.7 (26.6666) or $26^{2}/_{3}$ or $^{80}/_{3}$ or 26.6	B1	CAO	Do not allow 27 but condone 26.6 www
		iii	Mode = 31	B1	CAO	

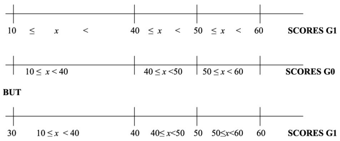
Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	The mode is not at all useful as it is just by chance that it is 31. Mark awarded for stating not useful and -not representative of data -does not represent Central Tendency -happened by chance (or similar) -comment about not appearing significantly more (only one repetition/only twice/etc) No mark for stating it would be useful OR NOT USEFUL because of -spread/range -sample size -negatively skewed -unaffected by outliers -isn't close to mean and median	E1	Allow any reasonable comment **Examiner's Comments** The mean, median and mode were usually given correctly although one or two candidates lost a mark due to over-specification of the mean or rounding of the median. However the final mark for the comment was awarded to only under a quarter of candidates. Many candidates gave general descriptions of the usefulness of the mode rather than commenting on this particular case. Too many candidates stated incorrectly that the mode was useful. Those who correctly stated that it was not useful, often followed this with an incorrect reason such as being unaffected by outliers; data being negatively skewed; or not being close to the mean and/or median.	
			Total	8		

Question			Answer/Indicative content	Marks	Part marks and guidance	
7		i	Mean = $\frac{(25\times147)+(45\times109)+(55\times182)+(70\times317)+(140\times175)}{930}$	M1	For midpoints (at least 3 correct) (allow 25.005, 45.005 etc leading to answer 70.20)	M0A0M0A0 unless using midpoints Answer must NOT be left as improper fraction as this is an estimate
		i	$\frac{750\times7+1250\times22+1750\times26+2500\times18+4000\times7}{80}$ $= \frac{151250}{80}$= (£)70.19 or (£)70.2 $\Sigma x^2 f$ = $(25^2 \times 147) + (45^2 \times 109) + (55^2 \times 182) + (70^2 \times 317) + (140^2 \times 175)$ = 91875 + 220725 + 550550 + 1553300 + 3430000 = 5846450	A1	CAO (exact answer 70.19355…) Correct answers obtained from use of calculator statistical functions gain full marks Condone answer of (£)70.20 For attempt at S_{xx} Should include sum of at least 3 correct multiples fx^2 $- \Sigma x^2/n$	Accept correct answers for mean and sd from calculator even if eg wrong *Sxx* given For use of midpoints 25.5, 45.5, 55.5, 70.5, 140.5 allow SC1 for £70.69 and SC1 for 36.89
		i	$S_{xx} = 5846450 - \frac{65280^2}{930}$ = 1264215.161 or 5846450 -930×70.19^2	M1	Do not FT their incorrect mean for A1	If using $(x-\bar{x})^2$ method, B2 if 36.9 or better, otherwise B0

Question			Answer/Indicative content	Marks	Part marks and guidance	
		i	$s =$ $\sqrt{\frac{1264215}{929}} = \sqrt{1360.83}$ = 36.89 or (£)36.9 Allow any answer between 36.87 and 36.90 without checking working	A1	(exact answer 36.88949…) Condone answer of (£)36.90 If both mean and sd overspecified, just deduct one mark **Examiner's Comments** This part was fairly well answered with over half of candidates gaining full credit. A few had no idea how to proceed, but most used correct midpoints, although some made slips with them or occasionally used figures such as 25.5, 45.5, etc. The standard deviation proved more difficult for a number of candidates with a variety of wrong methods seen. Very few used the statistical functions on their calculator to do this question, despite this being the recommended method. A few candidates over-specified either or both of their final answers and so lost a mark.	Allow use of 70.2 in calculation of S_{xx} = 1263372.8 leading to 36.87719… Condone RMSD of 36.87 (36.86985…) since *n* is so large

<table>
<tr><th colspan="3">Question</th><th>Answer/Indicative content</th><th>Marks</th><th colspan="2">Part marks and guidance</th></tr>
<tr><td></td><td></td><td>ii</td><td>100/120 × 175 = 145.83</td><td>M1*</td><td>For 175/120
<b>Examiner's Comments</b>
Candidates found this part rather more challenging, although almost half scored full marks. Trying to establish the proportion they were after was the biggest stumbling block. However, some were then unsure what to do with the figure of 145.83 once they had found it. Some rounded down to 145 (probably the most common mistake of those who understood what they needed to do) and others failed to finish by finding the percentage, just giving the final answer as a decimal 0.157.</td><td>Or 20/120 × 175 = 29.166 oe</td></tr>
<tr><td></td><td></td><td>ii</td><td>145.83/930 = 0.1568</td><td>*M1dep</td><td></td><td>(175 – 29.166)/930</td></tr>
<tr><td></td><td></td><td>ii</td><td>So 15.7%</td><td>A1</td><td></td><td>Accept 16% with working</td></tr>
<tr><td></td><td></td><td>iii</td><td>
<table>
<tr><th>Price</th><th>Frequency</th><th>Group width</th><th>Frequency density</th></tr>
<tr><td>$10 \le x \le 40$</td><td>147</td><td>30</td><td>4.90</td></tr>
<tr><td>$40 < x \le 50$</td><td>109</td><td>10</td><td>10.90</td></tr>
<tr><td>$50 < x \le 60$</td><td>182</td><td>10</td><td>18.20</td></tr>
<tr><td>$60 < x \le 80$</td><td>317</td><td>20</td><td>15.85</td></tr>
<tr><td>$80 < x \le 200$</td><td>175</td><td>120</td><td>1.46</td></tr>
</table></td><td>M1</td><td>For fds - at least 3 correct Accept any suitable unit for fd such as eg freq per cm.</td><td>M1 can be also be gained from freq per 10 – 4.9, 10.9, 18.2, 15.35, 0.146 (at least 3 correct) or similar.</td></tr>
</table>

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii		A1	Allow 15.9 and 1.5 and condone 1.45 **Examiner's Comments** This part was again well answered with around 80% of candidates gaining at least 4 marks out of 5. Various errors were seen, but none very commonly. The most frequently seen were: using frequency rather than frequency density, using a non-linear scale on one of the axes (usually the horizontal axis), stopping the horizontal axis at 120, and labelling the horizontal axis ‘Class width’.	If fd not explicitly given, M1 A1 can be gained from all heights correct (within ≤ one square) on histogram (and M1A0 if at least 3 correct)
		iii		B1	linear scales on both axes and label on both axes (Allow horizontal axis labelled *x*) Vertical scale starting from zero (not broken - but can get final mark for heights if broken)	Linear scale and label on vertical axis IN RELATION to first M1 mark ie fd or frequency density or if relevant freq/10, etc (NOT eg fd / 10). However allow scale given as fd × 10, or similar Accept f / w or f / cw (freq / width or freq / class width) Can also be gained from an accurate key G0 if correct label but not fd's.
		iii		B1	width of bars (within half a square) (NO GAPS ALLOWED)	Must have linear scale. Condone starting at 10 rather than 0. For inequality labels see additional notes below.
		iii	NB If not using fd's only mark available is B1 for width of bars			

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	Heights must be within ≤ 1 square of overlay (only for scales 2cm = 4 units (blue) or 5 units (red)) – otherwise check heights. Note that you must make sure that the overlay is aligned correctly with the vertical axis.	B1	height of bars  10 ≤ x < 40 ≤ x < 50 ≤ x < 60 SCORES G1 10 ≤ x < 40 40 ≤ x <50 50 ≤ x < 60 SCORES G0 BUT 30 10 ≤ x < 40 40 40≤ x<50 50 50≤x<60 60 SCORES G1 **Examiner's Comments** This part was again well answered with around 80% of candidates gaining at least 4 marks out of 5. Various errors were seen, but none very commonly. The most frequently seen were: using frequency rather than frequency density, using a non-linear scale on one of the axes (usually the horizontal axis), stopping the horizontal axis at 120, and labelling the horizontal axis 'Class width'.	Height of bars – must be linear vertical scale. FT of heights dep on at least 3 heights correct and all must agree with their fds If fds not given and 3 or 4 heights correct then max M1A0G1G1G0
		iv	Positive skewness	B1	Allow +ve **Examiner's Comments** Over 90% of candidates scored the one mark available here.	
		v	Area for men from 100 to 200 = 100 × 2 = 200 200/990 = 0.202	M1		Or $^{100}/_{120}$ × 240
		v	So 20.2%	A1		20% with working

Question			Answer/Indicative content	Marks	Part marks and guidance	
		v	Cannot be certain as both figures are estimates	E1	Independent **Examiner's Comments** A good number of candidates achieved full marks, and the question was answered better than question 6 part (ii) which is a similar calculation. Of those who got the calculation incorrect most started with 240/990 or 20/990, rather than 200/990. The explanation over certainty was well answered with most candidates achieving this mark, whether or not they got the first 2 marks.	Allow comments such as ‘grouped data so cannot be certain’ or ‘Values are not exact so cannot be certain’ oe or ‘midpoints have been used so cannot be certain’ oe
		vi	Men's running shoes have a lower average price than women's (as their mean is only £68.83 compared to £70.19). Or equivalent for women	E1	FT their mean Do NOT condone lower central tendency or lower mean	Allow ‘on the whole’ or similar in place of ‘average’.

Question			Answer/Indicative content	Marks	Part marks and guidance	
		vi	Men's running shoes have a more variation in price than women's (as their sd is £42.93 compared to £36.89). Or equivalent for women	E1	FT their SD **Examiner's Comments** Although this is essentially a simple question, almost a third of candidates scored zero. Candidates struggled to provide acceptable comparisons, with many relying on terms such as “central tendency” when comparing the means, and relatively few discussing averages. A more encouraging proportion of candidates were able to provide a good interpretation for the differences in the standard deviations. Some thought that central tendency was something to do with variation. A number of candidates were unable to construct a proper, legible, grammatically correct sentence.	Allow ‘more spread’ or similar but not ‘higher range’ or ‘higher variance’ or ‘less distributed’ Condone less consistent
			Total	18		

Question			Answer/Indicative content	Marks	Part marks and guidance	
8		i	Median = 29.0	B1		Condone wrong method
		i	IQR = 31.8 – 27.8	M1	For either quartile – allow	Allow 27.75 and 31.9 leading to 4.15
		i	= 4.0	A1	alternative definitions of quartiles **Examiner's Comments** The vast majority of candidates found the median correctly. A small minority misread/ignored the key to the stem and leaf diagram and gave an incorrect answer of 290. However under half of candidates found the quartiles correctly, with many using 5th and 15th values, which was penalised.	Do not allow 27.7, 27.9, 31.6,32.0
		ii	Lower limit = 27.8 – 1.5 × 4.0 = 21.8	M1	Method for either	For use of mean (29.44) and SD
		ii	27.75, 31.9 lead to 21.525 and 38.125 27.7, 31.6 lead to 21.85 and 37.45	A1	FT sensible quartiles and IQR	(2.516765…) 29.44 ± 2 × 2.516765 M1 Lower Limit = 24.4 A1 Upper limit = 34.5 A1 So no outliers B1
		ii	Upper limit = 31.8 + 1.5 × 4.0 = 37.8	A1	FT sensible quartiles and IQR	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii	So there are no outliers (at either end of the distribution)	B1	Dep on at least one A1 Use of median scores 0/4 **Examiner's Comments** Most candidates gained full marks, often on follow through from quartiles which were slightly out. The most common error was to use the median in calculations. A few candidates started from scratch and calculated mean and standard deviation. Some managed this successfully, but others made errors in their calculations, or incorrectly used a combination of both methods such as mean ± 1.5 × interquartile range.	
			Total	7		

9			Increases a value by 6
			New value is closer to 62 than the old value is to 61.4
			51 changes to 57
			or 57 changes to 63
			or 58 changes to 64
				M1(AO3.1b)
				M1(AO2.2a)
				A1(AO2.2a)
				[3]
					Implied by correct answer or pair of values differing by 6
					Implied by correct answer or new value closer to 62 than old value	

			Total	3		

Question			Answer/Indicative content	Marks	Part marks and guidance	

10		a	Allocate numbers 001 to 200 to the trees
			Choose 10 (3 digit) random numbers
				B1(AO1.2)
				B1(AO2.4)
				[2]
					e.g. use calculator to get 10 different random numbers	

		b	Mean = 27.61kg SD = 4.04 kg (3sf)	B1(AO1.1) B1(AO1.1) [2]	BC BC	

		c	Upper limit = 27.61 + 2 × 4.04 = 35.69
			So the value of 38.1 is an outlier
			This value should be investigated to check if it is genuine. If so, it should not be removed from the data
				M1(AO1.1)
				A1(AO1.1)
				B1(AO2.2b)
				[3]
					For mean + 2 × sd OR UQ + 1.5 IQR = 28.3 + 1.5 × 3.2 = 33.1
					OR e.g. If the value is not representative of the other 199 trees because e.g. this tree is a different type it should be ignored	

			Total	7		

Question			Answer/Indicative content	Marks	Part marks and guidance	
11		a	E.g. There is a greater spread of birth rates for countries in sub-Saharan African than for countries in the Caribbean E.g. The range for countries in Africa is greater than for countries in East and South East Asia but this could be caused by outliers as the IQRs are similar E.g. sub-Saharan Africa has a mixture of economically rich and poor countries resulting in a large IQR E.g. Countries in East and South East Asia tend to have higher life expectancy than countries in sub-Saharan Africa so their populations are older, on average, and have lower birth rates	B1(AO2.2b) B1(AO2.2b) B1(AO2.2b) [3]	B1 Correct relevant comment that can be inferred from the source material B1 Distinct correct relevant comment that can be inferred from the source material B1 Third distinct relevant comment that can be inferred from the source material (this mark is only available if the candidate's comments include reference to both features of the LDS and fig 9.1)	
		b	(*A*) E.g. The calculation doesn't use the populations as weights E.g. Does not take the populations into account	E1(AO2.3) [1]		
		b	(*B*) E.g. Lower because Australia has the highest population but the lowest birth rate oe E.g. answer given is too high as too much weight is given to Papua New Guinea	E1(AO2.2a) [1]		

Question			Answer/Indicative content	Marks	Part marks and guidance	
		c	[weak] negative	B1(AO1.2) [1]		
		d	E.g. Correlation / association does not imply causality E.g. Some countries with low birth rates have quite low physician density E.g. Some countries with low physician density have quite low birth rates E.g. Data do not show what happens after an increase in physicians Therefore it is not possible to be certain	E1(AO2.3) [1]		
			Total	7		

Question			Answer/Indicative content	Marks	Part marks and guidance	
12		a	Symmetrical with one possible outlier	B1(AO1.2) [1]	or negative skew	

		b	24th value – 8th value
			27 – 16 = 11
				M1(AO1.1)
				A1(AO1.1)
				[2]		

		c	16 – 1.5 × 11 = –0.5
			3 > – 0.5 so it is not an outlier.
				M1(AO1.1)
				A1(AO2.2a)
				[2]
					Check for outliers using their Q_1– 1.5 × IQR	

			Total	5		

Question			Answer/Indicative content	Marks	Part marks and guidance	

13		a	Consistent use of midpoints in either calculation
			Mean £36.25
			Sample sd 8.313…
			£36 250 and £8313
				M1(AO1.1)
				A1(AO1.1a)
				A1(AO1.1)
				A1(AO1.1)
				[4]
					soi
					BC
					BC
					FT their calculator values	

		b	We are using grouped data not the original values	B1(AO2.4) [1]		
		c	Any valid reason which suggests that the sample is not necessarily representative	B1(AO3.2b) [1]		
		d	It would increase	B1(AO2.2a) [1]		
			Total	7		

Question			Answer/Indicative content	Marks	Part marks and guidance	
14		a	Population because these are all the countries of interest.	B1(AO1.1) [1]		

		b	Eg Consistent with the correlation for all countries oe
			And
			Eg You would expect countries with higher populations to tend to have higher numbers of both mobile phone subscribers and internet users.
			Or
			Eg people who use mobile phones will be more likely to use the internet
				E1(AO2.4)
				E1(AO2.2b)
				[2]		

		c	Ukraine
			Has high mobile phone usage with lower internet provision.
			Suggests people are used to and/or like technology so potential customers for the internet.
				B1(AO1.1)
				E1(AO3.2a)
				[2]		

Question			Answer/Indicative content	Marks	Part marks and guidance		

		d	Number the companies from 000 to 599
			or 001 to 600
			Generate 3-digit random numbers (from a calculator or spreadsheet). Match number with number given to companies. [Discard 600 to 999.]
			Do not use any numbers twice.Stop when you have selected 20 different companies.
				E1(AO1.2)
				E1(AO1.1)
				E1(AO2.4)
				[3]
					Or from a table of random numbers.
					or discard 000 and 601 to 999.
						OR Input list of companies into a spreadsheet
						NB Writing the names on paper. Putting these in a hat. Selecting 20. **E1 E1E0**. Not practical.	

			Total	8			
15		a	Vertical scale	B1(AO 2.5) [1]			
		b	The sales are lower in the final time period (assuming a linear vertical scale) so the director is not correct.	E1(AO 2.3) [1]			
			Total	2			

Question			Answer/Indicative content	Marks	Part marks and guidance	

16		a	(ai) Mean for world = 7 174 654 290 ÷ 239
			= 30 million (approx.)
			Mean for sample is 9.36 million so much smaller
				M1(AO 1.1)
				A1(AO 1.1)
				B1(AO 1.1)
				[3]
					Any degree of accuracy	

		b	(aii) There are a lot of small countries in the world.
			Therefore there is no reason to suppose the sample is not random,
				B1(AO 2.4)
				E1(AO 2.2b)
				[2]		

		c	(b) 0.091 × 55 400
			=[$] 5041.40
				M1(AO 3.3)
				A1(AO 1.1)
				[2]
					Condone missing units.
					Answer can be rounded to 5041 or 5040	

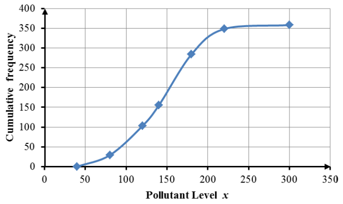
		d	(ci) Identify physicians per 1000 as highest R^2 with positive correlation
			and positive correlation
				B1(AO 2.2b)
				E1(AO 2.4)
				[2]
					NB graph C
					e.g. Reject D as it shows negative correlation	

		e	(cii) There is hardly any correlation.	E1(AO 3.5b) [1]		

Question			Answer/Indicative content	Marks	Part marks and guidance
			Total	10	

Question			Answer/Indicative content	Marks	Part marks and guidance	
17		i	Mean $= \dfrac{17100}{50} = 342$	B1	Ignore units CAO	
			$Sxx = 6115108 - \dfrac{17100^2}{50} = 266908$	M1	For S*xx* M1 for 6115108 – 50 x their mean^2 BUT NOTE M0 if their $S_{xx} < 0$	
			$s = \sqrt{\dfrac{266908}{49}} = \sqrt{5447.10} = 73.8 \ (73.8044\ldots)$	A1 [3]	CAO ignore units M1A0 for RMSD = 73.1 (73.062…) **Examiner's Comments** Nearly all candidates worked out the mean correctly. Many candidates also found the standard deviation but some over-specified the answer thus losing a mark. A minority of candidates made an error in the formula for standard deviation or worked out the RMSD. It was encouraging to see most candidates recalling and using correct formulae.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii	New mean = (0.108 × 342) + 7.2 = £44.14 New sd = 0.108 × 73.8 = £7.97 Using RMSD gives £7.89 Using variance gives 588.29	B1 M1 A1 [3]	FT their mean Allow £44.1 or better provided answer is positive FT their sd (unless negative)for M1 and A1 NB If candidate 'starts again' only award marks for CAO Do not penalise lack of units in mean or sd Deduct at most 1 mark overall in whole question for over-specification of either mean or SD or both **Examiner's Comments** Most candidates used the transformation to obtain the correct mean. Many candidates also obtained the correct standard deviation, with a pleasingly small number mistakenly adding 7.2 to their final answer. Some candidates lost marks for over-specification, although there was only a penalty of 1 mark in the whole question for this error. It was encouraging to see most candidates giving answers to an appropriate degree of accuracy.	
			Total	6		

<table>
<tr><th colspan="3">Question</th><th>Answer/Indicative content</th><th>Marks</th><th colspan="2">Part marks and guidance</th></tr>
<tr>
<td>18</td><td></td><td>i</td>
<td>
<table>
<tr><td>Upper bound</td><td>40</td><td>80</td><td>120</td><td>140</td><td>180</td><td>220</td><td>300</td></tr>
<tr><td>Cumulative frequency</td><td>0</td><td>29</td><td>103</td><td>155</td><td>284</td><td>348</td><td>358</td></tr>
</table>


NB If you receive a script where the graph is drawn on lined paper, rather than on the grid, please mark it and then refer it to your team leader <u>BEFORE</u> you submit it.
</td>
<td>B1

G1

G1

G1

G1

[5]</td>
<td>**Cumulative frequencies All correct.**
May be implied from graph. Condone omission of 0 at this stage.

For points Plotted as (UCB, their cf). Ignore (40,0) at this stage. No midpoint or LCB plots or non-linear scales
Plotted within ½ small square
If cf not given then allow B1G1 for all correct

For joining points (within ½ a square) For joining all of 'their points' (line or smooth curve) AND now including (40,0) Not for midpoint or LCB plots or non-linear scales

For scales
Linear horizontal scale. Allow if start at 40 (no inequality scales - Not even <40, <60, <80 …) Linear vertical scale Allow full credit if axes reversed correctly

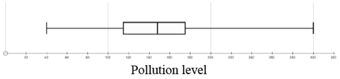
For labels
Pollutant level or *x* and Cumulative frequency or just CF or similar but not frequency or fd nor cumulative fd
All four dep on attempt at cumulative frequency.
Mid-point or LCB plots or cum freq bars may score first and last two marks
NOTE With one error in cfs last 4 marks still available (EG 0, 29, 103, 145, 274, 338, 348)</td>
<td></td>
</tr>
</table>

Question			Answer/Indicative content	Marks	Part marks and guidance	
					Examiner's Comments The majority of candidates made a good attempt at the graph. Very few failed to recognise that cumulative frequency was required, with only an occasional histogram or frequency graph seen. The values for the cumulative frequency were on the whole correctly calculated but a few tried to make them up to 365, failing to read the question correctly. The scales were usually linear but some chose difficult intervals, especially on the vertical scale, for example intervals of 24. Labelling was not as successful; missing labels or labelling the vertical scale as frequency was common. A number of candidates used mid-points rather than upper boundaries for plotting and a few used lower boundaries. Even if correct boundaries were used, the point (40,0) was often omitted with candidates either not joining their graph to the axis or joining it to (0,0). Just one third of candidates scored full marks in this question.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		ii	Estimate from curve is 327 Proportion = 327/358 = 0.913 or 91.3% 315/368 = 87.99% 316/358 = 88.3% NB Linear interpolation gives 284 + ½×64 = 316	M1 A1 [2]	**Allow 315 to 330 without checking graph (unless non-linear scales in which case allow 316 by LI)** **Otherwise FT their graph** within one square (allow a slight slip in scales – contact TL if unsure) **Max M1A0 if final answer given as a fraction** **Examiner's Comments** Candidates were fairly even spread between reading off the graph or using linear interpolation to find the cumulative frequency for x = 200. A sufficiently accurate value was usually obtained from the graph, but many responses stopped short of even writing it as a fraction over 358, let alone converting this value into a proportion (decimal or percentage).	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		iii	Median = 148 **Allow 145 to 152 without checking graph** Q_1 = 115 **Allow 110 to 115 without checking graph** Q_3 = 175 **Allow 175 to 180 without checking graph** IQR = 60 No marks if non-linear scales If quartiles not specified give B1B0 for 'IQR is 115 < x < 175' or similar If answer only for IQR, check if quartiles given in part (iv) or (v) – if not then check graph	**B1** **B1** **B1** **[3]**	For Q1 or Q3 For IQR FT their cf graph for all 3 marks within one square (on both scales) (allow a slight slip in scales - contact TL if unsure) **Examiner's Comments** This was a generally well done. The main problems were caused by unhelpful scales chosen in part (i), which candidates then interpreted wrongly in this part. Some candidates used cumulative frequencies of 100, 200 and 300, rather than the correct values to find the median and quartiles.	
		iv	Lower limit Q_1 – 1.5 × IQR '115 – (1.5 × 60)' (= 25) Upper limit Q_3 + 1.5 × IQR '175 + (1.5 × 60)' (= 265) There are definitely no outliers at the lower end as the lowest data value is 40 which is below the lower limit. It is uncertain whether there	**M1** **M1** **A1** **A1** **[4]**	FT their quartiles provided between 40 and 300 Allow 'No values below (their) 25' for first A1 Allow 'Lower limit = (their) 25 so no outliers' You must be convinced that comments about no outliers refer to **lower tail only**. Allow additional comment that since some data is lost there could be one or more outliers If their lower limit > 40 then A0 Do not allow 'There IS at least one outlier.' oe There must be an element	

Question			Answer/Indicative content	Marks	Part marks and guidance	
			are outliers at the upper end as the highest class includes the upper limit. Use of mean = 145.08 and sd = 45.09 gives 54.9 and 235.26 for M2 So could be some outliers at lower and could be some at upper end but not sure. E1E1		of doubt. However, condone 'There is probably at least one outlier.' You must be convinced that comments about some outliers refer to upper tail only. If their upper limit <220 or >300 then A0 **Examiner's Comments** This was again generally well done with most candidates correctly calculating the outlier limits. Most responses correctly stated that there were no outliers at the lower end but some stated that there were definitely outliers at the upper end rather than that there may be some. A number of candidates used the median instead of the lower and upper quartiles to find the limits and others used 2 × IQR, rather than 1.5 × IQR. A very few candidates found the mean and standard deviation and then using these, found the limits correctly.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		v		G1* G1*dp G1*dp [3]	FT their median and quartiles provided between 40 and 300 and Q_1 < median < Q_3 Can restart from graph For linear scale shown. Dep on attempt at box and whisker plot with at least a box and one whisker. Condone lack of label. For boxes (Q_1, median, Q_3) in correct positions, within half a square For whiskers at 40 and 300 within half a square Upper whisker could be partially dotted **Examiner's Comments** Although this part was generally answered well, a minority of candidates lost the final mark by not having the end of the whiskers plotted at 40 and/or 300, often plotting these at 29 and/or 358. Some candidates did not show a horizontal scale, making their response difficult to mark. Other candidates had trouble drawing the box and whisker diagram due the lack of a ruler.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		vi	The readings from Tower Hamlets show (stronger) positive skewness The readings from Marylebone Road show little evidence of skewness Accept 'No skewness' For 2 marks must suggest that TH has higher positive skew than MR	E1 E1 **[2]**	Allow '<u>slight</u> positive skewness' Do <u>not</u> FT their diagram but must have boxplot in part (v) to get second mark 'TH shows more evidence of positive skewness than MR' gets E2 **Examiner's Comments** Many candidates struggled to answer the question which was asked. Often zero marks were scored as the candidate wrote a short essay with no mention of skewness. Being precise and talking about both locations generally gained the marks. Some candidates still referred to it as left and right skew or mixed up positive and negative. The question did ask for a comparison, which was generally missed.	
			Total	19		

Question			Answer/Indicative content	Marks	Part marks and guidance	

19		a	Frequency density 4.6
			Number of runners 21
				B1(AO2.2a)
				B1(AO1.1)
				[2]
					Examiner's Comments
					This part was usually done correctly, though some candidates omitted one or both answers. Multiples of the correct answer were sometimes seen instead of 4.6.	

		b	Vertical axis should be labelled "number of runners per minute"	E1(AO2.3) [1]	Or frequency density **Examiner's Comments** The most common wrong answer was to say that the vertical axis should be labelled 'frequency'.	
			Total	3		

Question			Answer/Indicative content	Marks	Part marks and guidance	
20		a	9483	B1(AO1.1) [2]	BC **Examiner's Comments** This part was almost always done correctly.	

		b	7 × 10 112 +10259 – 10014 soi
			= 71029
			66381 – 9204 + *x* = "71029"
			Emma needs to make 13852 steps
				M1(AO3.1a)
				A1(AO1.1)
				M1(AO1.1)
				A1(AO3.2a)
				[4]
					$\frac{7 \times 10\,112 + 10259 - 10014}{7}$
					NB Rose's new mean is 10147 = 10147
					$= \frac{66381 - 9204 + x}{7}$
					Using 8 days can gain M1 only
					Examiner's Comments
					Most candidates found this part quite straightforward. One common error in this part was to use 10 112, the original 7-day mean for Rose, rather than work out the new 7-day mean for days 2 to 8 inclusive. Another common error was to find how many steps Emma should take on day 8 so that she has taken as many steps as Rose over the period of 8 days; this was often done by considering 8-day means.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		c	10341 + 2×948 soi = 12237	M1(AO3.1b)		
			Comparison of their 13852 with their 12237	M1(AO1.1)	Soi	
			13852 is an outlier, so Emma would need to make an unusually high number of steps on day 8	A1(AO3.2b) [3]	Conclusion; 'outlier' not essential. Dep M2 www	
					<u>Examiner's Comments</u> Candidates were expected to compare their answer to part (b) with the long-term mean number of steps for Emma given in the question, 10 341. In order to gain full credit candidates were expected to understand that the question was asking if the answer to part (b) was an outlier, and were expected to use the definition of outlier given in section D13 of the specification; candidates using other definitions of outlier were not given full credit. Candidates using the data in the table in the question to work out quartiles or mean and standard deviation were not given credit.	
			Total	8		

Question			Answer/Indicative content	Marks	Part marks and guidance	

21		a	Any two distinct reasons
			eg classes of different widths represented by bars of same width
			eg vertical axis should be frequency density
			eg final upper class boundary not given
			eg should have continuous horizontal scale / no gaps between bars
				E1(AO2.4)
				E1(AO1.1)
				[2]
					Examiner's Comments
					Many candidates gave valid answers. However, the question asked for two respects in which the presentation of the data is incorrect, and an answer like 'it should be a histogram' does not answer the question.	

		b	(i) Positive correlation
			(ii) 9.8395 or 9.8 or 9.84 or 9.840
				B1(AO2.2b)
				[1]
				B1(AO1.1)
				[1]
					oe
					Examiner's Comments
					Most candidates gave a correct answer to this part. However, stating that there is a correlation between birth and death rates was not given any credit, nor were statements like 'higher birth rates cause higher death rates'.	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		(iii)	extrapolation	B1(AO3.2b) [1]	Examiner's Comments This question was usually answered correctly. However, a small number of candidates appeared to have read an approximate answer from the graph (usually 10) despite the clear instruction to use the equation of the line of best fit to calculate this. oe	
		(iv)	Birth rates and death rates in the Caribbean, may be very different from those in Africa.	E1(AO2.2a) [1]	Examiner's Comments Many candidates correctly answered that this would require extrapolation; many others gave an acceptable equivalent answer by pointing out that there was no data in that part of the scatter diagram so using the line of best fit would be unreliable. oe | Advantage	
		(v)	eg other continents to select countries from	E1(AO2.2a) [1]	Examiner's Comments Most candidates were sufficiently familiar with the Large Data Set to answer this part. eg a random sample would | Advantage	

Question			Answer/Indicative content	Marks	Part marks and guidance	
					almost certainly not just include countries from Africa **Examiner's Comments** Very few candidates answered this part correctly. Many candidates gave answers that suggested that it would be difficult to take a random sample from all the population of Africa because of communication difficulties, size of the continent, unreliability of records, etc., whilst true, do not address the context of the question. The majority of candidates did not appreciate that the sample taken was of 55 countries, and that this sample was taken from all the countries of the world, as listed in the Large Data Set.	

Question			Answer/Indicative content	Marks	Part marks and guidance	

		c	Eg Generate a random number, *n*, between 1 and 4 and select the *n*th item in the data set.
			Eg Select every 4th item on the list thereafter (stopping when 14 have been selected)
				B1(AO1.2)
				B1(AO1.1)
				[2]
					Candidates may choose other valid starting points
					Candidates may choose other valid intervals
					Examiner's Comments
					There were very few completely correct answers to this part. A great many candidates described how to find a random sample by assigning random numbers to the countries. A great many others described stratified sampling. Those candidates who did describe taking, typically, every fourth country often omitted to point out the need to use a random starting point.	

			Total	9		

Question			Answer/Indicative content	Marks	Part marks and guidance	
22		a	2.031578947….rounded to two or more sf isw BC	**B1** **(AO 1.1)** **[1]**	NB 2.0, 2.03 or 2.032 **Examiner's Comments** Whilst it is acceptable to calculate the mean using a formal written method, there is an expectation in the new specification that candidates will use the statistical functions on their calculators for parts (a) and (b), hence the single mark allocation.	
		b	1.076367330…rounded to two or more sf isw BC	**B1** **(AO 1.1)** **[1]**	NB 1.1, 1.08 or 1.076 **Examiner's Comments** Candidates who did well in this question made efficient use of the appropriate calculator function. Candidates who did less well made laborious calculations and slipped up in the arithmetic. ? **Misconception** The choice of appropriate formula (using *n*, *n* –1 or other denominator corrections) is beyond the scope of this A Level Maths qualification, and the use of a single formula for all contexts is expected. The H640 OCR B (MEI) specification and formulae sheet makes clear that the divisor (*n* –1) should be used	

Question			Answer/Indicative content	Marks	Part marks and guidance
			Total	2	

Question			Answer/Indicative content	Marks	Part marks and guidance	
23		a	negative skew	**B1** **(AO 1.2)** [1]	Examiner's Comments Some confusion between positive and negative skew was seen.	
		b	(used) the mode	**B1** **(AO 1.1)** [1]	Examiner's Comments A simple statement was expected, and not a mini essay. Almost every candidate gained this mark.	
		c	(used) the median	**B1** **(AO 1.1)** [1]	Examiner's Comments A simple statement was expected, and not a mini essay. The majority of candidates gained the mark	

Question			Answer/Indicative content	Marks	Part marks and guidance	
		d	61 – 1.5 × (88 – 61) 20.5 < 35 [so 35 is not an outlier] so he does not move to set 2	M1 (AO 2.1) A1 (AO 2.2b) [2]	*Alternatively,* 73.61 – 2 × 17.03 39.6 > 35 [so 35 is an outlier] so he moves to set 2 allow eg only marks below 20.5 (or 39.6) would lead to a move down plus correct conclusion <u>Examiner's Comments</u> Candidates who did well in this question used the lower quartile and the interquartile range to determine whether Benson's mark is an outlier. Candidates who did less well used the median in conjunction with the lower quartile.	
			Total	5		
24		a	Positive skew	B1 (AO1.2) [1]		
		b	Layout correct scale on axis and range of boxplot correct their IQR and median shown IQR is 19 to 40 and median is 28	M1 (AO1.1a) A1 (AO1.1) M1 (AO1.1) A1 (AO1.1) [4]	allow 20.5 to 41 but not 19 to 41 or 20.5 to 40	
			Total	5		

Question			Answer/Indicative content	Marks	Part marks and guidance	
25		a	the pattern of data points suggest a straight line oe the value of *r* indicates that the fit is close oe	E1 (AO3.3) E1 (AO2.4) [2]		
		b	70	B1 (AO3.4) [1]	from 0.89×83 – 3.76 (= 70.11)	
		c	(10 + 3.76)÷ 0.89 13	M1 (AO3.4) A1 (AO1.1) [2]	(= 13.21 …)	
		d	The approximation for Tina’s mark is obtained by interpolation, whereas the approximation for Dave’s mark is from extrapolation	E1 (AO3.5b) [1]	allow eg should use the equation for *x* on y to estimate Dave’s mark	
			Total	6		
26		a	=B2*C2	E1 (AO2.4) [3]	must have “=”	
		b	A All populations are in the upper quartile B All total GDPs are in the upper quartile Both statements are consistent with the data.	E1 (AO2.4) E1 (AO2.4) B1 (AO2.2a) [3]		
		c	No because some countries do not take part in the Olympics OR having more Olympic athletes in a country may encourage others	E1 (AO2.2a) [1]		
			Total	5		

Question			Answer/Indicative content	Marks	Part marks and guidance		
27		a	Opportunity sampling	B1(AO 1.2) [1]			
		b	Each sample of this size does not have an equal probability of being selected since not all the customers will come at 11.00 am on a Monday	E1(AO 2.4) [1]	allow eg Some of Qasim's customers will be at work and never be able to visit café at 11am on Monday		
		c	Positive skew	B1(AO 1.2) [1]			

		d	Median is 10th value = 30
			43 – 22
			= 21
				B1(AO 1.1)
				M1(AO 1.1)
				A1(AO 1.1)
				[3]
					For their 15th value – their 5th value	NB other convention s for finding the quartiles are acceptable as long as the method is clear.	

		e	Any two reasonable statements eg Medians similar, so age of typical customer found to be the same IQR much smaller in larger sample, suggesting less variability range is larger, (but these values are both outliers and could be atypical)	B1(AO 2.4) B1(AO 2.4) [2]			
			Total	8			

Question			Answer/Indicative content	Marks	Part marks and guidance	
28		a	If a particular data value is not available, the code #N/A has been included in the Large Data Set – this is to prevent some software reading a blank as a zero.	E1 (AO 2.4) [1]		
		b	The populations of African countries vary considerably. The “mean of means” does not take weighting into account.	E1 (AO 2.4) [1]		
		c	The heights of the bars are proportional to the frequencies in the frequency diagram, the areas of the bars are proportional to frequency in a histogram (or the heights of the bars are proportional to frequency density)	E1 (AO 2.4) [1]		
		d	Frequency densities of 1.3, 2.4, 3.8 and 0.733 Horiziontal and vertical scales correct and correctly labelled Correct diagram with no gaps between bars	B1 (AO 1.1) B1 (AO 1.1) B1 (AO 1.1) [3]		
		e	eg there may some association between the variables, but it is not clear what sort, so using a value of one variable to predict a value of the other variable is unlikely to be reliable	B1 (AO 2.4) B1 (AO 2.2b) [2]	or eg in the subsection of scatter where median life expectancy is 17, there appears to be no correlation	
			Total	8		