OCR Statistics 01

Past paper questions on

- Representation of data
- Measures of location
- Measures of spread

Edited by K V Kumaran
Email: kvkumaran@gmail.com
Phone: 07961319548

Representation Of Data

- You must be happy constructing unordered, back-to-back and ordered stem and leaf diagrams. They show the overall distribution of the data and back-to-back diagrams allow you to compare two sets of data.

- Cumulative frequency graphs. The cumulative frequency is a "running total" of the frequencies as you go up the values. For example

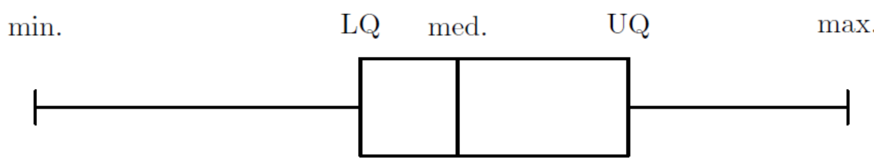
x	f
$0 \leqslant x < 5$	8
$5 \leqslant x < 10$	13
$10 \leqslant x < 15$	17
$15 \leqslant x < 20$	10

$\Rightarrow$ Create $\Rightarrow$
$\Rightarrow$ Cumulative $\Rightarrow$
$\Rightarrow$ Frequency $\Rightarrow$

x (upper limit of)	cum. freq
5	8
10	21
15	38
20	48

Plot the second of these tables and join it with a smooth curve to form the *cumulative frequency curve*. From this the median and the two quartiles can be found.

- Once these values are found we can draw a *box and whisker diagram*. The box and whisker diagram uses five values: the minimum, the maximum, the lower quartile, the upper quartile and the median. It is good for showing spread and comparing two quantities.



- Histograms are usually drawn for continuous data in classes. If the classes have equal widths, then you merely plot amount against frequency.

- If the classes do *not* have equal widths then we need to create a new column for *frequency density*. Frequency density is defined by f.d. $= \frac{\text{frequency}}{\text{class width}}$. The *area* of the bars are what represents the frequency, *not* the height.

- Frequency polygons are made by joining together the mid-points of the bars of a histogram with a ruler.

Measures Of Location

- The *mean* (arithmetic mean) of a set of data $\{x_1, x_2, x_3 \dots x_n\}$ is given by

$$\overline{x} = \frac{\text{sum of all values}}{\text{the number of values}} = \frac{\sum x}{n}.$$

When finding the mean[1] of a frequency distribution the mean is given by

$$\frac{\sum(xf)}{\sum f} = \frac{\sum(xf)}{n}.$$

- If a set of numbers is arranged in ascending (or descending) order the *median* is the number which lies half way along the series. It is the number that lies at the $\left(\frac{n+1}{2}\right)^{\text{th}}$ position. Thus the median of $\{13, 14, 15, 15\}$ lies at the $2\frac{1}{2}$ position $\Rightarrow$ average of 14 and 15 $\Rightarrow$ median $= 14.5$.

- The *mode* of a set of numbers is the number which occurs the most frequently. Sometimes no mode exists; for example with the set $\{2, 4, 7, 8, 9, 11\}$. The set $\{2, 3, 3, 3, 4, 5, 6, 6, 6, 7\}$ has two modes 3 and 6 because each occurs three times. One mode $\Rightarrow$ "unimodal". Two modes $\Rightarrow$ "bimodal". More than two modes $\Rightarrow$ "multimodal".

	ADVANTAGES	DISADVANTAGES
MEAN	⋆ The best known average. ⋆ Can be calculated exactly. ⋆ Makes use of all the data. ⋆ Can be used in further statistical work.	⋆ Greatly affected by extreme values. ⋆ Can't be obtained graphically. ⋆ When the data are discrete can give an impossible figure (2.34 children).
MEDIAN	⋆ Can represent an actual value in the data. ⋆ Can be obtained even if some of the values in a distribution are unknown. ⋆ Unaffected by irregular class widths and unaffected by open-ended classes. ⋆ Not influenced by extreme values.	⋆ For grouped distributions its value can only be estimated from an ogive. ⋆ When only a few items available or when distribution is irregular the median may not be characteristic of the group. ⋆ Can't be used in further statistical calculations.
MODE	⋆ Unaffected by extreme values. ⋆ Easy to calculate. ⋆ Easy to obtain from a histogram.	⋆ May exist more than one mode. ⋆ Can't be used for further statistical work. ⋆ When the data are grouped its value cannot be determined exactly.

Measures Of Spread

- The simplest measure of spread is the *range*. Range $= x_{\text{max}} - x_{\text{min}}$.

- The interquartile range is simply the upper quartile take away the lower quartile. Both of these values are usually found from a cumulative frequency graph (above).

- The *sum of squares from the mean* is called the *sum of squares* and is denoted

$$S_{xx} = \sum(x - \overline{x})^2 = \sum x^2 - n\overline{x}^2.$$

For example given the data set $\{3, 6, 7, 8\}$ the mean is 6; $\sum x^2 = 9 + 36 + 49 + 64 = 158$; so $S_{xx} = \sum x^2 - n\overline{x}^2 = 158 - 4 \times 6^2 = 14$.[2]

- The *standard deviation* (σ) is defined: $\sigma = \sqrt{\text{variance}} = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2}$.

- *Example*: Given the set of data $\{5, 7, 8, 9, 10, 10, 14\}$ calculate the standard deviation. Firstly we note that $\overline{x} = 9$.

$$\sigma = \sqrt{\frac{\sum x^2}{n} - \overline{x}^2} = \sqrt{\frac{(5^2 + \cdots + 14^2)}{7} - 9^2}$$
$$= \sqrt{\frac{615}{7} - 81} = 2.6186\ldots$$

- When dealing with frequency distributions such as

x	1	2	3	4	5
f	4	5	7	5	4

, we *could* calculate σ by writing out the data[3] and carrying out the calculations as above, but this is clearly slow and inefficient. To our rescue comes a formula for σ that allows direct calculation from the table. This is

$$\sigma = \sqrt{\frac{\sum(x^2 f)}{n} - \overline{x}^2}.$$

- *Example*: Calculate mean and sd for the above frequency distribution. For easy calculation we need to add certain columns to the usual x and f columns thus;

x	f	xf	$x^2 f$
1	4	4	4
2	5	10	20
3	7	21	63
4	5	20	80
5	4	20	100
	$n = \sum f = 25$	$\sum(xf) = 75$	$\sum(x^2 f) = 267.$

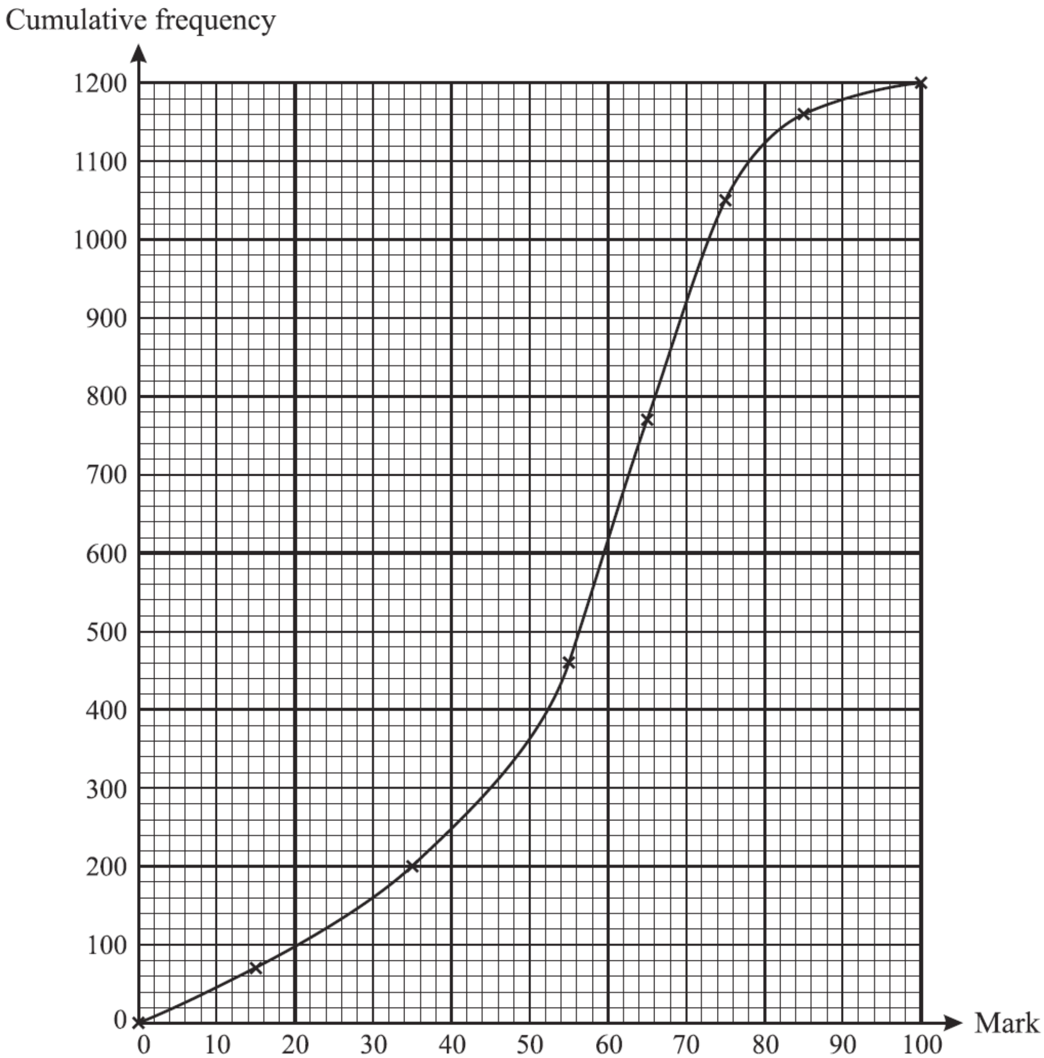
So $\overline{x} = \frac{\sum(xf)}{n} = \frac{75}{25} = 3$ and $\sigma = \sqrt{\frac{\sum(x^2 f)}{n} - \overline{x}^2} = \sqrt{\frac{267}{25} - 3^2} = 1.2961\ldots$

- *Linear Coding*. Given the set of data $\{2, 3, 4, 5, 6\}$ we can see that $\overline{x} = 4$ and it can be calculated that $\sigma = 1.414$ (3dp). If we add 20 to all the data points we can see that the mean becomes 24 and the standard deviation will be unchanged. If the data set is multiplied by 3 we can see that the mean becomes 12 and the standard deviation would become three times as large (4.743 (3dp)).

- If, instead of being given $\sum x$ and $\sum x^2$, you were given $\sum(x - a)$ and $\sum(x - a)^2$ for some constant a, you just use the substitution $u = x - a$ and use $\sum u$ and $\sum u^2$ to work out the mean of u and the standard deviation of u. Then, using the above paragraph, we know $\bar{x} = \bar{u} + a$ and $\sigma_x = \sigma_u$.

1.

The examination marks obtained by 1200 candidates are illustrated on the cumulative frequency graph, where the data points are joined by a smooth curve.



Use the curve to estimate

(i) the interquartile range of the marks, [3]

(ii) x, if 40% of the candidates scored more than x marks, [3]

(iii) the number of candidates who scored more than 68 marks. [2]

Five of the candidates are selected at random, with replacement.

(iv) Estimate the probability that all five scored more than 68 marks. [3]

It is subsequently discovered that the candidates' marks in the range 35 to 55 were evenly distributed — that is, roughly equal numbers of candidates scored 35, 36, 37, …, 55.

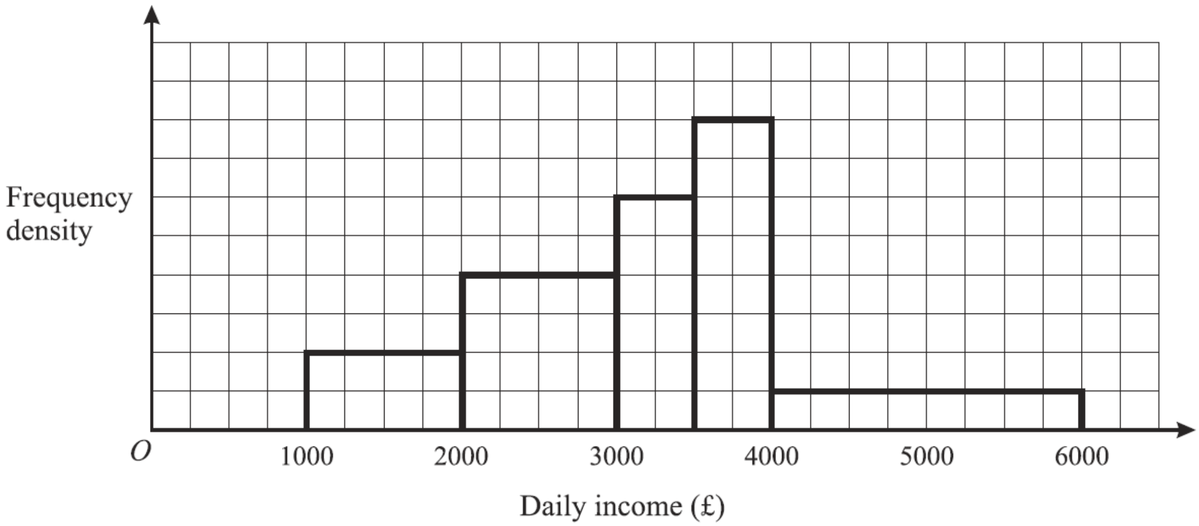
(v) What does this information suggest about the estimate of the interquartile range found in part **(i)**? [2]

Q5 June 2005

2.

Each day, the Research Department of a retail firm records the firm's daily income, to be used for statistical analysis. The results are summarised by recording the number of days on which the daily income is within certain ranges.

(i)



The histogram shows the results for 300 days. By considering the total area of the histogram,

(a) find the number of days on which the daily income was between £4000 and £6000, [4]

(b) calculate an estimate of the number of days on which the daily income was between £2700 and £3200. [3]

(ii) The Research Department offers to provide any of the following statistical diagrams: histogram, frequency polygon, box-and-whisker plot, cumulative frequency graph, stem-and-leaf diagram and pie chart.

Which one of these statistical diagrams would most easily enable managers to

(a) read off the median and quartile values of the daily income, [1]

(b) find the range of the top 10% of values of the daily income? [1]

Q4 Jan 2006

3.

In a UK government survey in 2000, smokers were asked to estimate the time between their waking and their having the first cigarette of the day. For heavy smokers, the results were as follows.

Time between waking and first cigarette	1 to 4 minutes	5 to 14 minutes	15 to 29 minutes	30 to 59 minutes	At least 60 minutes
Percentage of smokers	31	27	19	14	9

Times are given correct to the nearest minute.

(i) Assuming that 'At least 60 minutes' means 'At least 60 minutes but less than 240 minutes', calculate estimates for the mean and standard deviation of the time between waking and first cigarette for these smokers. [6]

(ii) Find an estimate for the interquartile range of the time between waking and first cigarette for these smokers. Give your answer correct to the nearest minute. [4]

(iii) The meaning of 'At least 60 minutes' is now changed to 'At least 60 minutes but less than 480 minutes'. Without further calculation, state whether this would cause an increase, a decrease or no change in the estimated value of

(a) the mean, [1]

(b) the standard deviation, [1]

(c) the interquartile range. [1]

Q7 June 2006

4.

In the 2001 census, the household size (the number of people living in each household) was recorded. The percentages of households of different sizes were then calculated. The table shows the percentages for two wards, Withington and Old Moat, in Manchester.

	Household size						
	1	2	3	4	5	6	7 or more
Withington	34.1	26.1	12.7	12.8	8.2	4.0	2.1
Old Moat	35.1	27.1	14.7	11.4	7.6	2.8	1.3

(i) Calculate the median and interquartile range of the household size for Withington. [3]

(ii) Making an appropriate assumption for the last class, which should be stated, calculate the mean and standard deviation of the household size for Withington. Give your answers to an appropriate degree of accuracy. [6]

The corresponding results for Old Moat are as follows.

Median	Interquartile range	Mean	Standard deviation
2	2	2.4	1.5

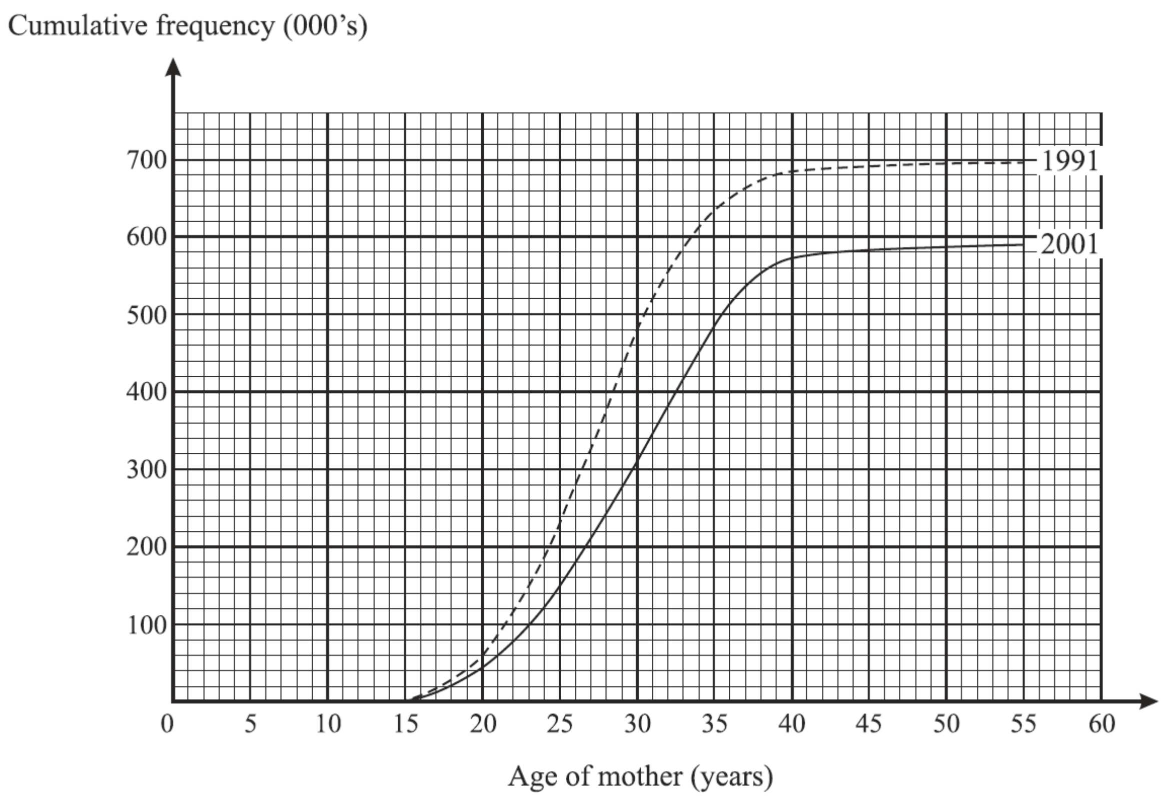
(iii) State one advantage of using the median rather than the mean as a measure of the average household size. [1]

(iv) By comparing the values for Withington with those for Old Moat, explain briefly why the interquartile range may be less suitable than the standard deviation as a measure of the variation in household size. [1]

(v) For one of the above wards, the value of Spearman's rank correlation coefficient between household size and percentage is –1. Without any calculation, state which ward this is. Explain your answer. [2]

Q8 Jan 2007

The numbers of births, in thousands, to mothers of different ages in England and Wales, in 1991 and 2001 are illustrated by the cumulative frequency curves.



(i) In which of these two years were there more births? How many more births were there in this year? [2]

(ii) The following quantities were estimated from the diagram.

Year	Median age (years)	Interquartile range (years)	Proportion of mothers giving birth aged below 25	Proportion of mothers giving birth aged 35 or above
1991	27.5	7.3	33%	9%
2001				18%

(a) Find the values missing from the table. [5]

(b) Did the women who gave birth in 2001 tend to be younger or older or about the same age as the women who gave birth in 1991? Using the table and your values from part **(a)**, give two reasons for your answer. [3]

Q5 June 2007

6.

The stem-and-leaf diagram shows the age in completed years of the members of a sports club.

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

Key: 1 | 4 | 0 represents a male aged 41 and a female aged 40.

(i) Find the median and interquartile range for the males. [3]

(ii) The median and interquartile range for the females are 27 and 15 respectively. Make two comparisons between the ages of the males and the ages of the females. [2]

(iii) The mean age of the males is 30.7 and the mean age of the females is 27.5, each correct to 1 decimal place. Give one advantage of using the median rather than the mean to compare the ages of the males with the ages of the females. [1]

A record was kept of the number of hours, X, spent by each member at the club in a year. The results were summarised by

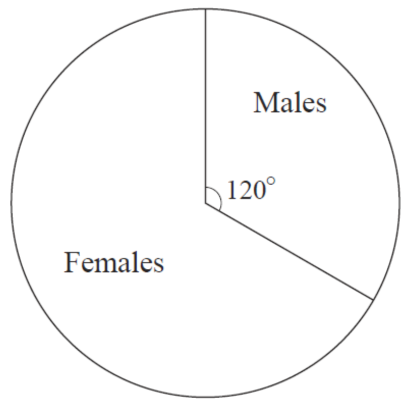
$$n = 49, \quad \Sigma(x - 200) = 245, \quad \Sigma(x - 200)^2 = 9849.$$

(iv) Calculate the mean and standard deviation of X. [6]

Q8 Jan 2008

7.

(i) The numbers of males and females in Year 12 at a school are illustrated in the pie chart. The number of males in Year 12 is 128.

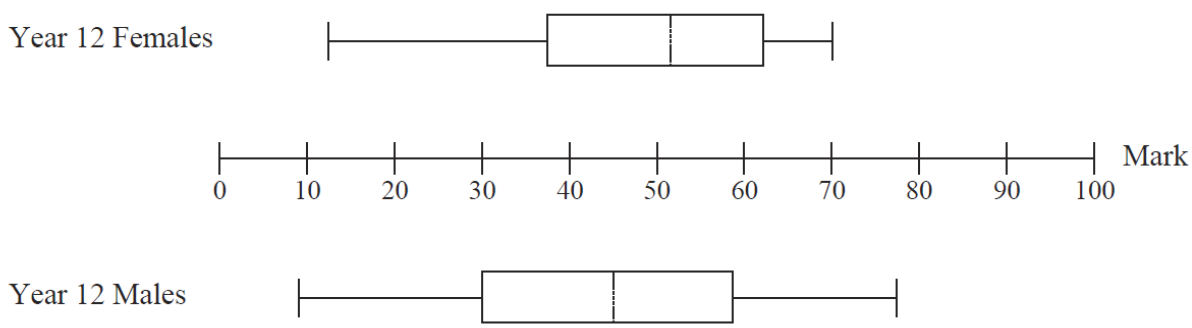


Year 12

(a) Find the number of females in Year 12. [1]

(b) On a corresponding pie chart for Year 13, the angle of the sector representing males is 150°. Explain why this does not necessarily mean that the number of males in Year 13 is more than 128. [1]

(ii) All the Year 12 students took a General Studies examination. The results are illustrated in the box-and-whisker plots.



(a) One student said "The Year 12 pie chart shows that there are more females than males, but the box-and-whisker plots show that there are more males than females."

Comment on this statement. [1]

(b) Give two comparisons between the overall performance of the females and the males in the General Studies examination. [2]

(c) Give one advantage and one disadvantage of using box-and-whisker plots rather than histograms to display the results. [2]

(iii) The mean mark for 102 of the male students was 51. The mean mark for the remaining 26 male students was 59. Calculate the mean mark for all 128 male students. [3]

Q6 June 2008

8.

The stem-and-leaf diagram shows the masses, in grams, of 23 plums, measured correct to the nearest gram.

5	5 6 7 8 8 9	
6	1 2 3 5 6 8 9	
7	0 0 2 4 5 6 7 8	Key : 6 | 2 means 62
8	0	
9	7	

(i) Find the median and interquartile range of these masses. **[3]**

(ii) State one advantage of using the interquartile range rather than the standard deviation as a measure of the variation in these masses. **[1]**

(iii) State one advantage and one disadvantage of using a stem-and-leaf diagram rather than a box-and-whisker plot to represent data. **[2]**

(iv) James wished to calculate the mean and standard deviation of the given data. He first subtracted 5 from each of the digits to the left of the line in the stem-and-leaf diagram, giving the following.

0	5 6 7 8 8 9	
1	1 2 3 5 6 8 9	
2	0 0 2 4 5 6 7 8	Key : 1 | 2 means 12
3	0	
4	7	

The mean and standard deviation of the data in this diagram are 18.1 and 9.7 respectively, correct to 1 decimal place. Write down the mean and standard deviation of the data in the original diagram. **[2]**

Q5 Jan 2009

9.

The diameters of 100 pebbles were measured. The measurements rounded to the nearest millimetre, x, are summarised in the table.

x	$10 \leqslant x \leqslant 19$	$20 \leqslant x \leqslant 24$	$25 \leqslant x \leqslant 29$	$30 \leqslant x \leqslant 49$
Number of stones	25	22	29	24

These data are to be presented on a statistical diagram.

(i) For a histogram, find the frequency density of the $10 \leqslant x \leqslant 19$ class. **[2]**

(ii) For a cumulative frequency graph, state the coordinates of the first two points that should be plotted. **[2]**

(iii) Why is it not possible to draw an exact box-and-whisker plot to illustrate the data? **[1]**

Q5 June 2009

10.

Last year Eleanor played 11 rounds of golf. Her scores were as follows:

79, 71, 80, 67, 67, 74, 66, 65, 71, 66, 64.

(i) Calculate the mean of these scores and show that the standard deviation is 5.31, correct to 3 significant figures. **[4]**

(ii) Find the median and interquartile range of the scores. **[4]**

This year, Eleanor also played 11 rounds of golf. The standard deviation of her scores was 4.23, correct to 3 significant figures, and the interquartile range was the same as last year.

(iii) Give a possible reason why the standard deviation of her scores was lower than last year although her interquartile range was unchanged. **[1]**

In golf, smaller scores mean a better standard of play than larger scores. Ken suggests that since the standard deviation was smaller this year, Eleanor's overall standard has improved.

(iv) Explain why Ken is wrong. **[1]**

(v) State what the smaller standard deviation does show about Eleanor's play. **[1]**

Q6 June 2009

11.

40 people were asked to guess the length of a certain road. Each person gave their guess, l km, correct to the nearest kilometre. The results are summarised below.

l	10–12	13–15	16–20	21–30
Frequency	1	13	20	6

(i) **(a)** Use appropriate formulae to calculate estimates of the mean and standard deviation of l. **[6]**

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

(ii) A histogram is to be drawn to illustrate the data. Calculate the frequency density of the block for the 16–20 class. **[2]**

(iii) Explain which class contains the median value of l. **[2]**

(iv) Later, the person whose guess was between 10 km and 12 km changed his guess to between 13 km and 15 km. Without calculation state whether the following will increase, decrease or remain the same:

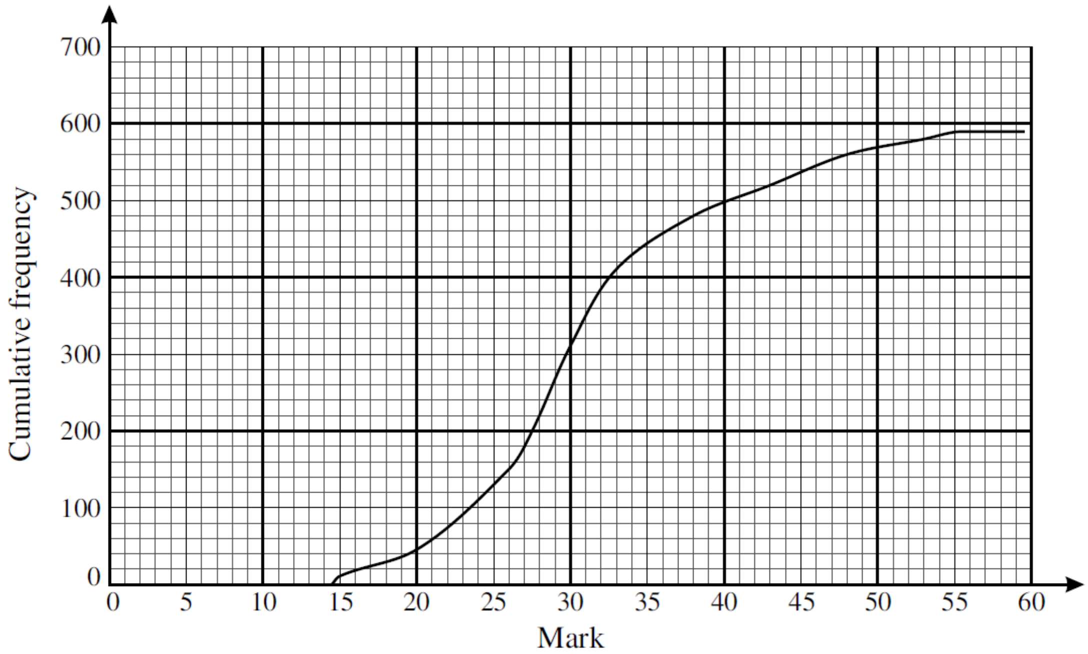
(a) the mean of l, **[1]**

(b) the standard deviation of l. **[1]**

Q2 Jan 2010

12.

The marks of some students in a French examination were summarised in a grouped frequency distribution and a cumulative frequency diagram was drawn, as shown below.



(i) Estimate how many students took the examination. **[1]**

(ii) How can you tell that no student scored more than 55 marks? **[1]**

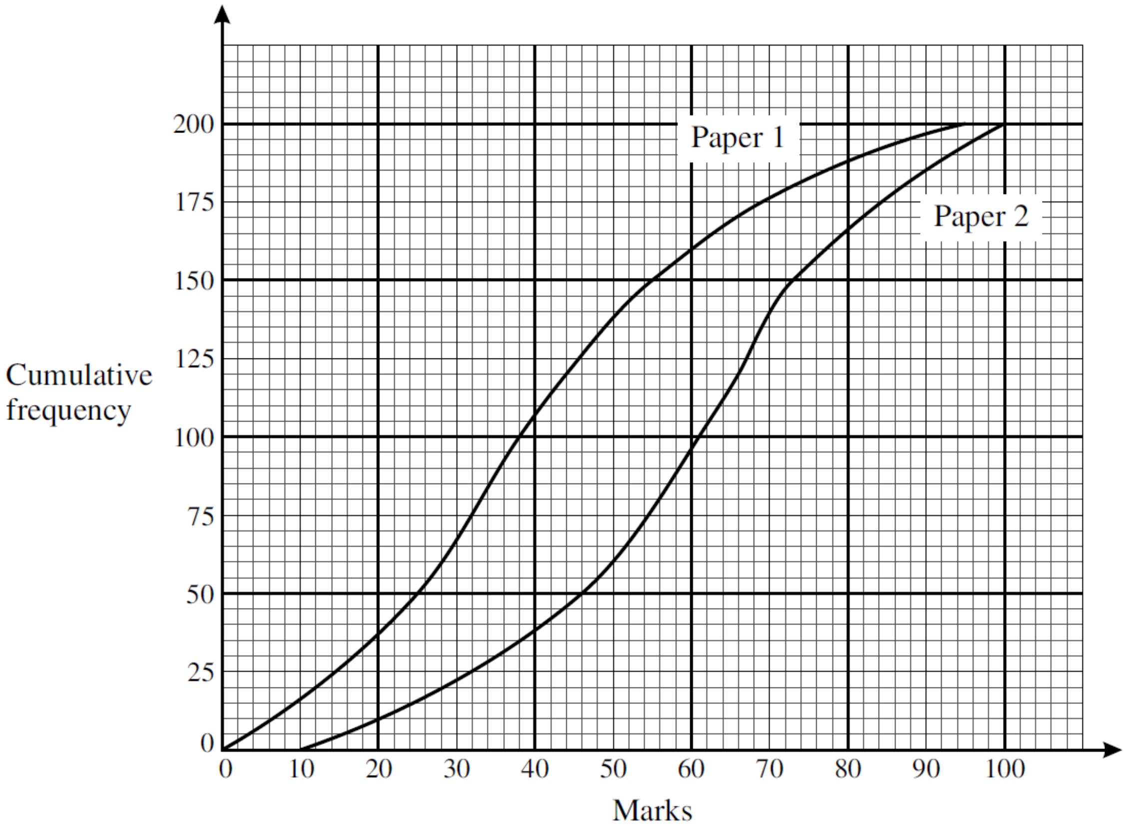
(iii) Find the greatest possible range of the marks. **[1]**

(iv) The minimum mark for Grade C was 27. The number of students who gained exactly Grade C was the same as the number of students who gained a grade lower than C. Estimate the maximum mark for Grade C. **[3]**

(v) In a German examination the marks of the same students had an interquartile range of 16 marks. What does this result indicate about the performance of the students in the German examination as compared with the French examination? **[3]**

Q1 June 2010

200 candidates took each of two examination papers. The diagram shows the cumulative frequency graphs for their marks.



(i) Estimate the median mark for each of the papers. **[2]**

(ii) State, with a reason, which of the two papers was the easier one. **[2]**

(iii) It is suggested that the marks on Paper 2 were less varied than those on Paper 1. Use interquartile ranges to comment on this suggestion. **[4]**

(iv) The minimum mark for grade A, the top grade, on Paper 1 was 10 marks lower than the minimum mark for grade A on Paper 2. Given that 25 candidates gained grade A in Paper 1, find the number of candidates who gained grade A in Paper 2. **[2]**

(v) The mean and standard deviation of the marks on Paper 1 were 36.5 and 28.2 respectively. Later, a marking error was discovered and it was decided to add 1 mark to each of the 200 marks on Paper 1. State the mean and standard deviation of the new marks on Paper 1. **[2]**

Q1 Jan 2011

14.

The table shows information about the time, t minutes correct to the nearest minute, taken by 50 people to complete a race.

Time (minutes)	$t \leqslant 27$	$28 \leqslant t \leqslant 30$	$31 \leqslant t \leqslant 35$	$36 \leqslant t \leqslant 45$	$46 \leqslant t \leqslant 60$	$t \geqslant 61$
Number of people	0	4	28	14	4	0

(i) In a histogram illustrating the data, the height of the block for the $31 \leqslant t \leqslant 35$ class is 5.6 cm. Find the height of the block for the $28 \leqslant t \leqslant 30$ class. (There is no need to draw the histogram.) **[3]**

(ii) The data in the table are used to estimate the median time. State, with a reason, whether the estimated median time is more than 33 minutes, less than 33 minutes or equal to 33 minutes. **[3]**

(iii) Calculate estimates of the mean and standard deviation of the data. **[6]**

(iv) It was found that the winner's time had been incorrectly recorded and that it was actually less than 27 minutes 30 seconds. State whether each of the following will increase, decrease or remain the same:

(a) the mean, **[1]**

(b) the standard deviation, **[1]**

(c) the median, **[1]**

(d) the interquartile range. **[1]**

Q4 June 2011

15.

The masses, x kg, of 50 bags of flour were measured and the results were summarised as follows.

$$n = 50 \qquad \Sigma(x - 1.5) = 1.4 \qquad \Sigma(x - 1.5)^2 = 0.05$$

Calculate the mean and standard deviation of the masses of these bags of flour. **[6]**

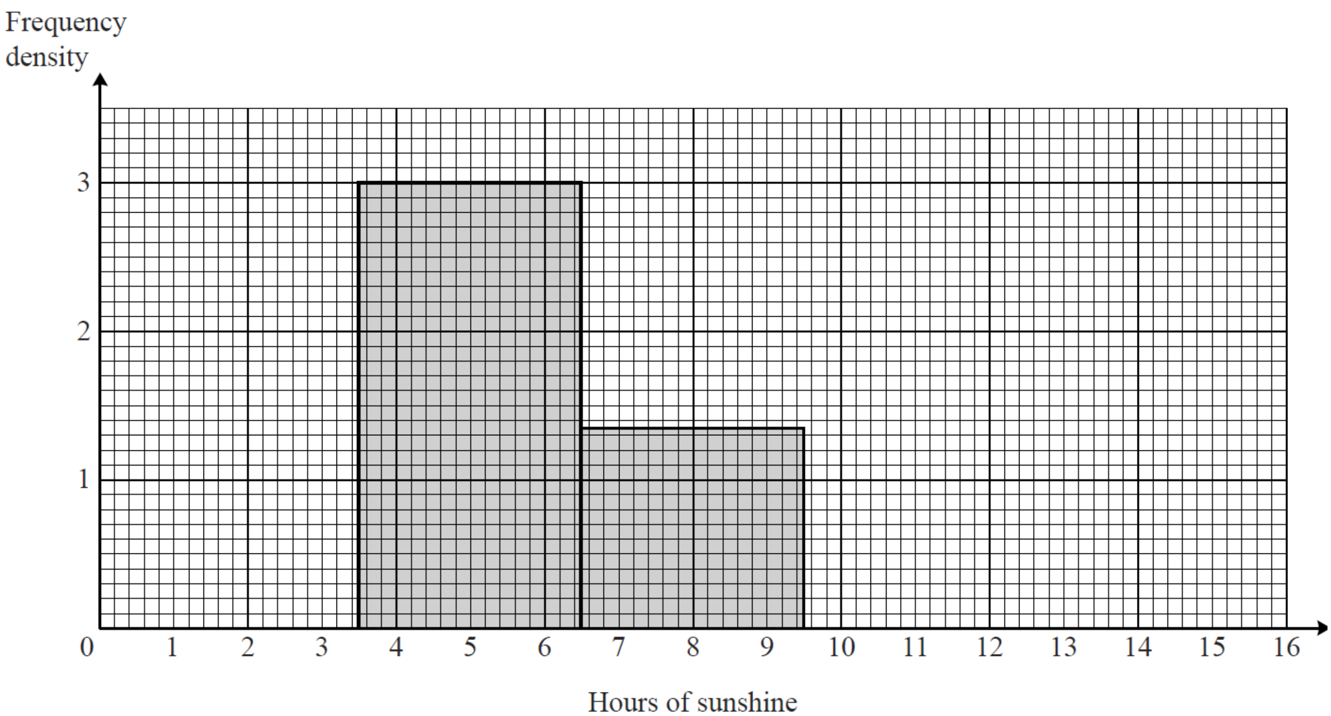
Q2 June 2012

16.

5 At a certain resort the number of hours of sunshine, measured to the nearest hour, was recorded on each of 21 days. The results are summarised in the table.

Hours of sunshine	0	1 – 3	4 – 6	7 – 9	10 – 15
Number of days	0	6	9	4	2

The diagram shows part of a histogram to illustrate the data. The scale on the frequency density axis is 2 cm to 1 unit.



(i) (a) Calculate the frequency density of the 1 – 3 class. **[1]**

(b) Fred wishes to draw the block for the 10 – 15 class on the same diagram. Calculate the height, in centimetres, of this block. **[2]**

(ii) A cumulative frequency graph is to be drawn. Write down the coordinates of the first two points that should be plotted. You are not asked to draw the graph. **[2]**

(iii) (a) Calculate estimates of the mean and standard deviation of the number of hours of sunshine. **[5]**

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

Q5 Jan 2012

17.

The test marks of 14 students are displayed in a stem-and-leaf diagram, as shown below.

0	
1	2 6
2	1 3 5
3	w x 4 8 y z
4	6 7 7

Key: 1 | 6 means 16 marks

(i) Find the lower quartile. **[1]**

(ii) Given that the median is 32, find the values of w and x. **[2]**

(iii) Find the possible values of the upper quartile. **[2]**

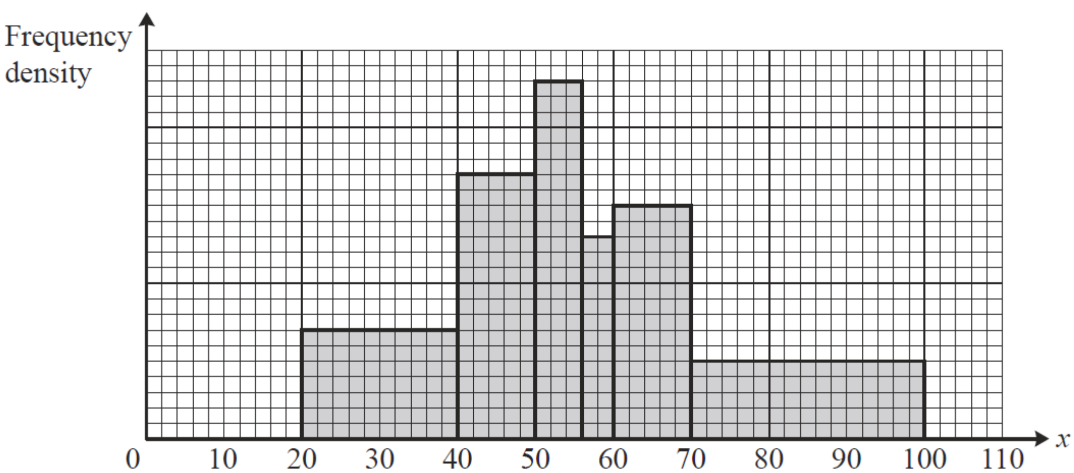
(iv) State one advantage of a stem-and-leaf diagram over a box-and-whisker plot. **[1]**

(v) State one advantage of a box-and-whisker plot over a stem-and-leaf diagram. **[1]**

Q3 June 2012

18.

The masses, x grams, of 800 apples are summarised in the histogram.



(i) On the frequency density axis, 1 cm represents a units. Find the value of a. **[3]**

(ii) Find an estimate of the median mass of the apples. **[4]**

Q6 Jan 2013

19.

The lengths, in centimetres, of 18 snakes are given below.

24 62 20 65 27 67 69 32 40 53 55 47 33 45 55 56 49 58

(i) Draw an ordered stem-and-leaf diagram for the data. **[3]**

(ii) Find the mean and median of the lengths of the snakes. **[2]**

(iii) It was found that one of the lengths had been measured incorrectly. After this length was corrected, the median increased by 1 cm. Give two possibilities for the incorrect length and give a corrected value in each case. **[2]**

Q1 June 2013

20.

At a stall in a fair, contestants have to estimate the mass of a cake. A group of 10 people made estimates, m kg, and for each person the value of $(m - 5)$ was recorded. The mean and standard deviation of $(m - 5)$ were found to be 0.74 and 0.13 respectively.

(i) Write down the mean and standard deviation of m. **[2]**

The mean and standard deviation of the estimates made by another group of 15 people were found to be 5.6 kg and 0.19 kg respectively.

(ii) Calculate the mean of all 25 estimates. **[2]**

(iii) Fiona claims that if a group's estimates are more consistent, they are likely to be more accurate. Given that the true mass of the cake is 5.65 kg, comment on this claim. **[2]**

Q4 June 2013

21.

The stem-and-leaf diagram shows the heights, in metres to the nearest 0.1 m, of a random sample of trees of species A.

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

Key: 6 | 4 means 6.4 m

(i) Find the median and interquartile range of the heights. **[3]**

(ii) The heights, in metres to the nearest 0.1 m, of a random sample of trees of species B are given below.

7.6 5.2 8.5 5.2 6.3 6.3 6.8 7.2 6.7 7.3 5.4 7.5 7.4 6.0 6.7

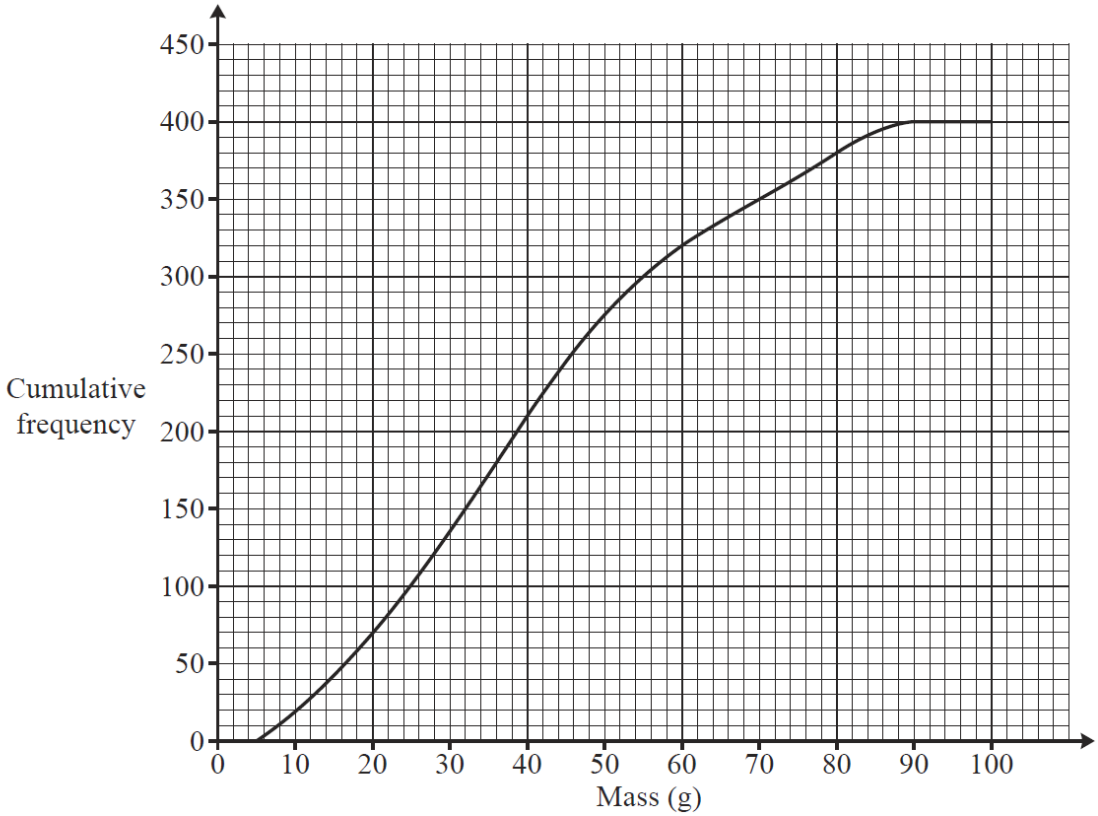
In the answer book, complete the back-to-back stem-and-leaf diagram. **[2]**

(iii) Make two comparisons between the heights of the two species of tree. **[2]**

Q1 June 2014

22.

The masses, in grams, of 400 plums were recorded. The masses were then collected into class intervals of width 5 g and a cumulative frequency graph was drawn, as shown below.



(i) Find the number of plums with masses in the interval 40 g to 45 g. **[1]**

(ii) Find the percentage of plums with masses greater than 70 g. **[2]**

(iii) Give estimates of the highest and lowest masses in the sample, explaining why their exact values cannot be read from the graph. **[2]**

(iv) On the graph paper in the answer book, draw a box-and-whisker plot to illustrate the masses of the plums in the sample. **[4]**

(v) Comment briefly on the shape of the distribution of masses. **[1]**

Q2 June 2015