MEASURES OF VARIATION

Text

Contents

Measures of Variation

1 Cumulative Frequency

Cumulative frequencies are useful if more detailed information is required about a set of data. In particular, they can be used to find the median and inter-quartile range.

The *inter-quartile range* contains the middle 50% of the sample and describes how spread out the data are. This is illustrated in Worked Example 2.

Worked Example 1

For the data given in the table, draw up a cumulative frequency table and then draw a cumulative frequency graph.

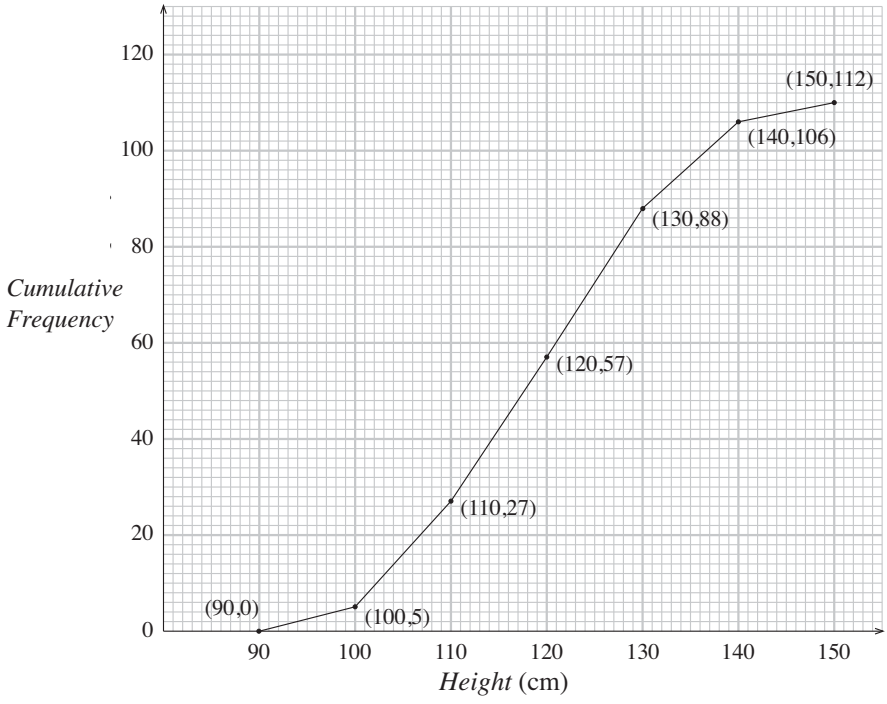
Height (cm)	*Frequency*
$90 < h \leq 100$	5
$100 < h \leq 110$	22
$110 < h \leq 120$	30
$120 < h \leq 130$	31
$130 < h \leq 140$	18
$140 < h \leq 150$	6

Solution

The table below shows how to calculate the cumulative frequencies.

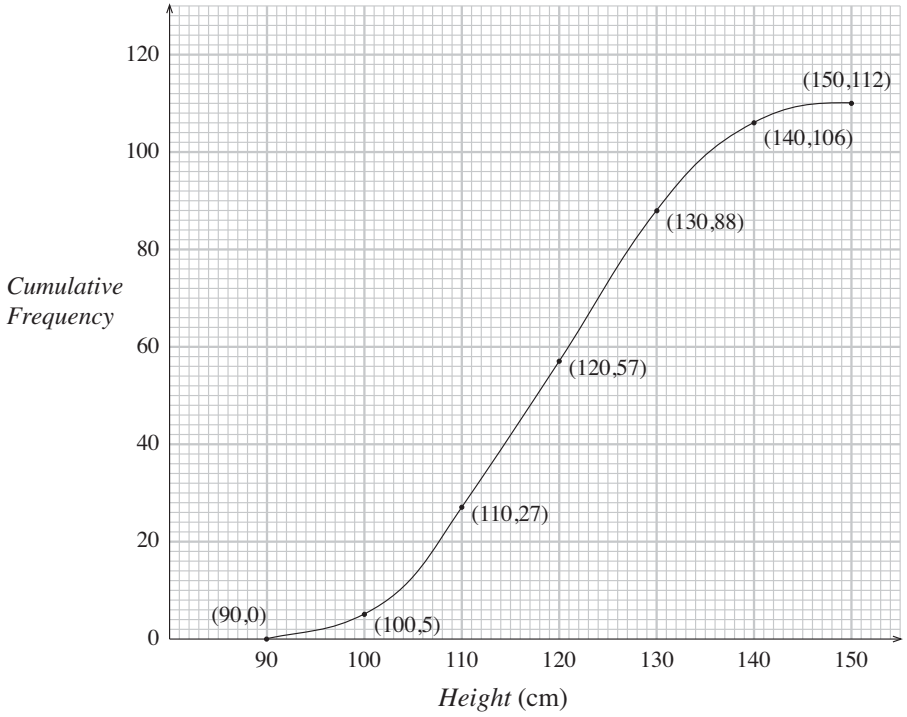
Height (cm)	*Frequency*	*Cumulative Frequency*
$90 < h \leq 100$	5	5
$100 < h \leq 110$	22	$5 + 22 = 27$
$110 < h \leq 120$	30	$27 + 30 = 57$
$120 < h \leq 130$	31	$57 + 31 = 88$
$130 < h \leq 140$	18	$88 + 18 = 106$
$140 < h \leq 150$	6	$106 + 6 = 112$

A graph can then be plotted using points as shown below.



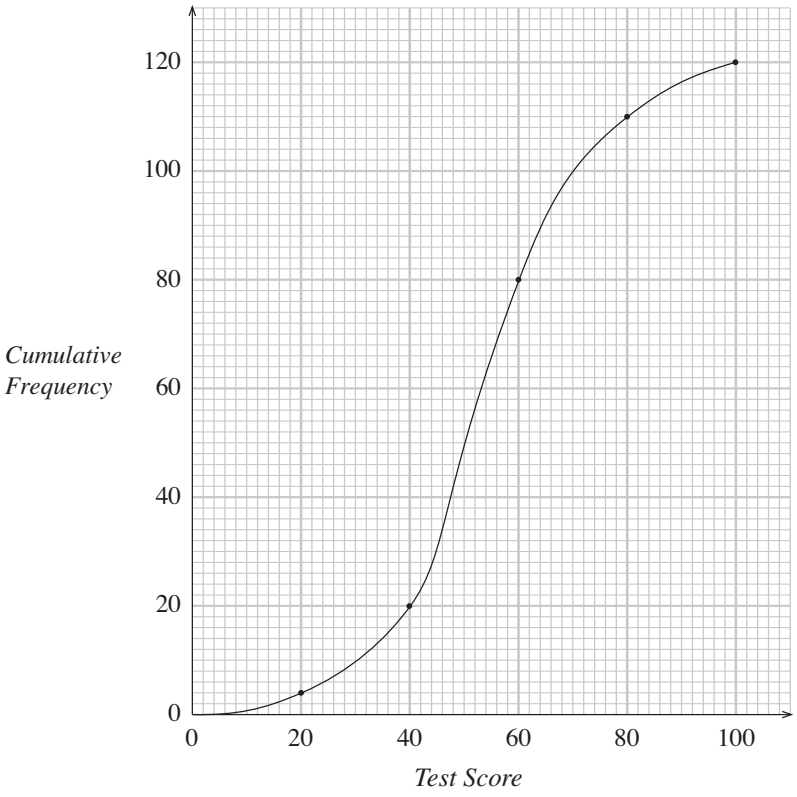
Note

A graph is found by drawing a smooth curve through the points, rather than using straight line segments.



Worked Example 2

The cumulative frequency graph below gives the results of 120 students on a test.



Use the graph to find:

(a) the median score, (b) the inter-quartile range,

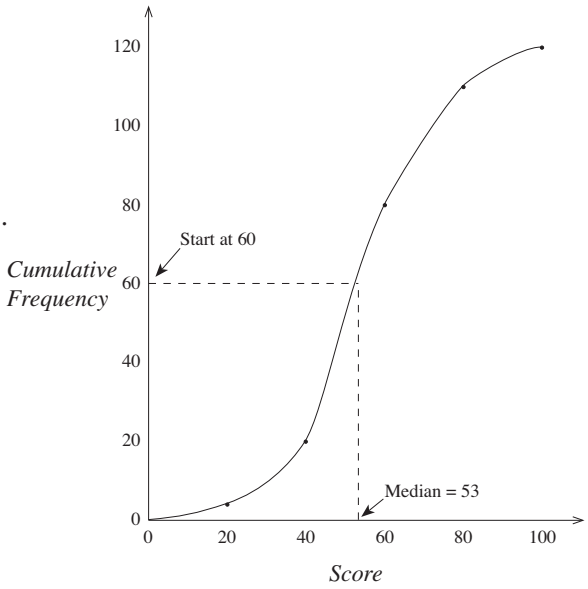
(c) the mark which was attained by only 10% of the students,

(d) the number of students who scored more than 75 on the test.

Solution

(a) Since $\frac{1}{2}$ of 120 is 60, the *median* can be found by starting at 60 on the vertical scale, moving horizontally to the graph line and then moving vertically down to meet the horizontal scale.

In this case the median is 53.



(b) To find out the *inter-quartile range*, we must consider the middle 50% of the students.

To find the *lower quartile*, start at $\frac{1}{4}$ of 120, which is 30.

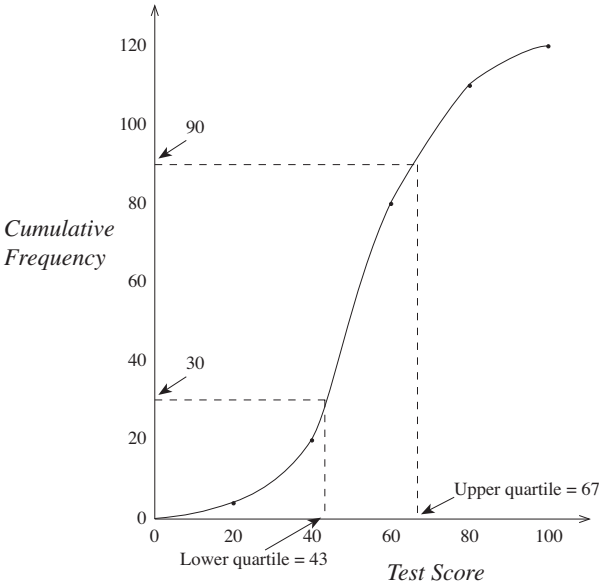
This gives

Lower Quartile = 43

To find the *upper quartile*, start at $\frac{3}{4}$ of 120, which is 90.

This gives

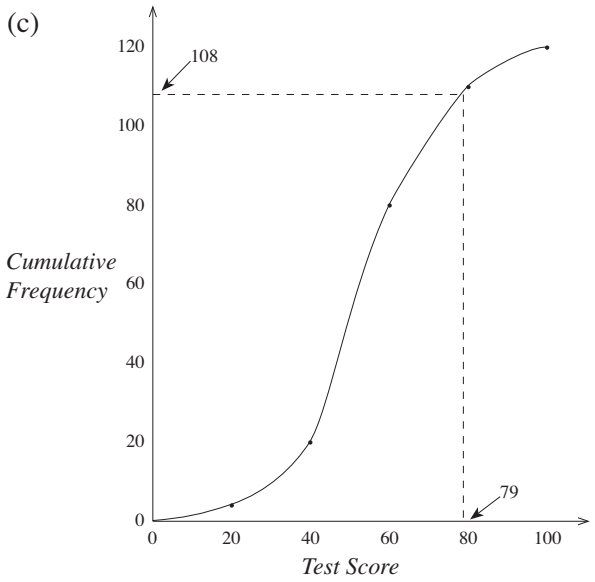
Upper Quartile = 67



The *inter-quartile range* is then

$$\begin{aligned}\text{Inter-quartile Range} &= \text{Upper Quartile} - \text{Lower Quartile}\\ &= 67 - 43\\ &= 24\end{aligned}$$

(c)



Here the mark which was attained by the top 10% is required.

$$10\% \text{ of } 120 \;=\; 12$$

so start at 108 on the cumulative frequency scale.

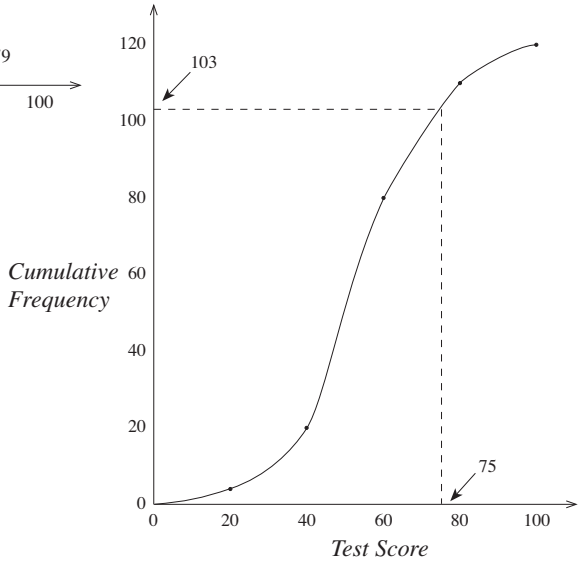
This gives a mark of 79.

(d) To find the number of students who scored more than 75, start at 75 on the horizontal axis.

This gives a cumulative frequency of 103.

So the number of students with a score greater than 75 is

$$120 - 103 \;=\; 17$$



Information

*A **quartile** is one of 3 values (lower quartile, median and upper quartile) which divides data into 4 equal groups.*

*A **percentile** is one of 99 values which divides data into 100 equal groups.*

*The **lower quartile** corresponds to the 25th percentile. The **median** corresponds to the 50th percentile. The **upper quartile** corresponds to the 75th percentile.*

We denote the lower quartile by Q_1, *the median by* Q_2 *and the upper quartile by* Q_3.

Worked Example 3

Mark	Frequency	Cumulative Frequency
1 - 10	2	2
11 - 20	5	7
21 - 30	9	16
31 - 40	14	
41 - 50	16	
51 - 60	12	
61 - 70	8	
71 - 80	4	70

The table below shows the distribution of marks on a test for a group of 70 students.

(a) Copy and complete the table to show the cumulative frequency for the distribution.

(b) (i) Using a scale of 1 cm to represent 5 marks on the horizontal axis and 1 cm to represent 5 students on the vertical axis, draw the cumulative frequency curve for the scores.

(ii) What assumption have you made in drawing your curve through the point $(0, 0)$?

(c) The pass mark for the test is 47. Use your graph to determine the number of students who passed the test.

(d) What is the probability that a student chose at random had a mark less than or equal to 30 ?

Solution

(a)

Mark	Frequency	Cumulative Frequency
1 - 10	2	2
11 - 20	5	7
21 - 30	9	16
31 - 40	14	30
41 - 50	16	46
51 - 60	12	58
61 - 70	8	66
71 - 80	4	70

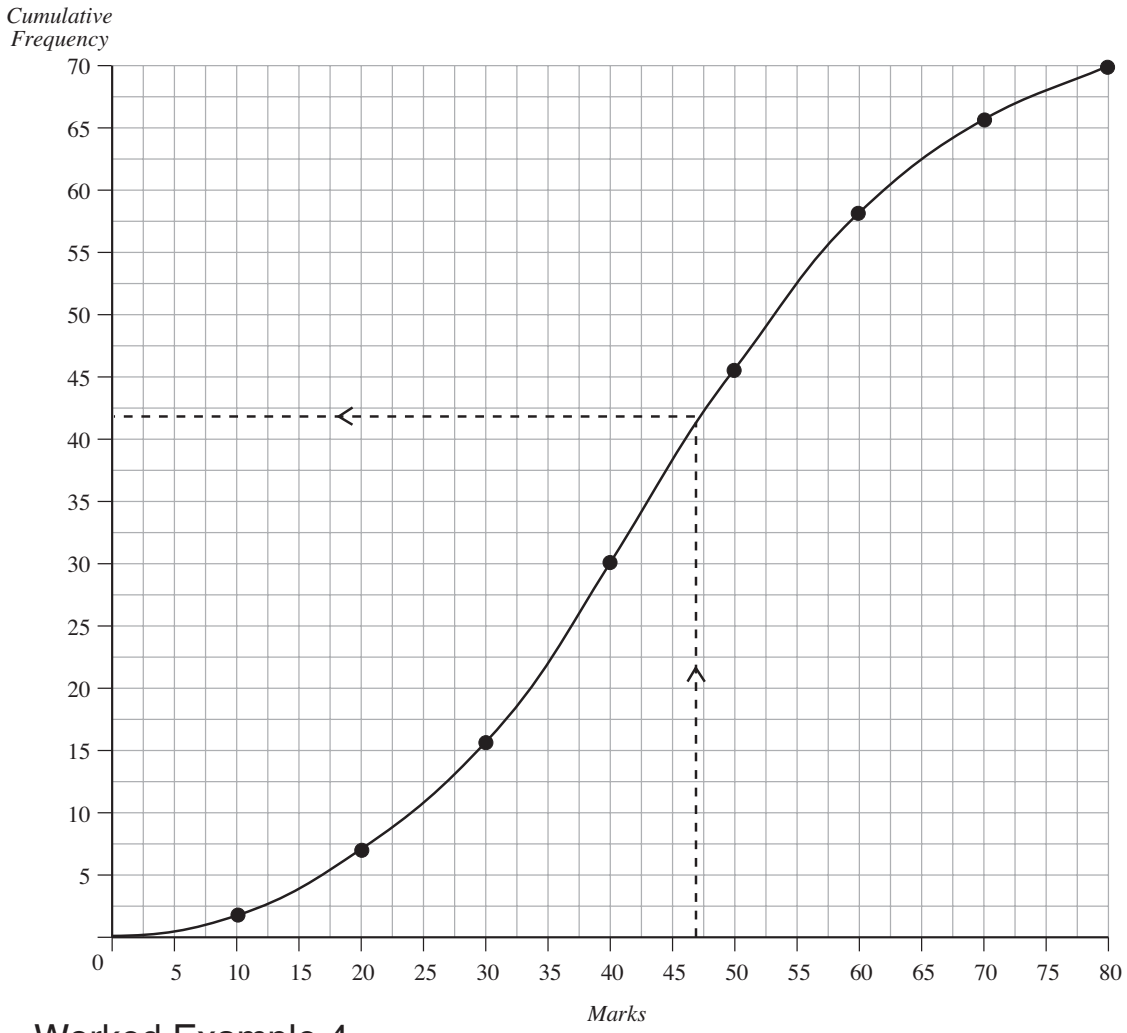
(b) (i) Graph on following page.

(ii) Assumption that no student scored zero.

(c) $70 - 42 = 28$ students passed the test.

(d) Probability $= \frac{16}{70} = \frac{8}{35}$.

(b) (i)



Worked Example 4

The heights, in metres, of a random sample of 40 soldiers from a regiment were measured. The heights are summarised in the following table.

Height in metres (x)	*Frequency*	*Cumulative frequency*
$1.75 \leq x < 1.80$	1	1
$1.80 \leq x < 1.85$	1	2
$1.85 \leq x < 1.90$	4	
$1.90 \leq x < 1.95$	13	
$1.95 \leq x < 2.00$	14	
$2.00 \leq x < 2.05$	3	
$2.05 \leq x < 2.10$	3	
$2.10 \leq x < 2.15$	1	40

(a) Copy and complete the cumulative frequency column in the table.

(b) Construct a cumulative frequency curve for the data.

(c) Estimate from this cumulative frequency curve:

(i) the median (Q_2) (ii) the upper quartile (Q_3) (iii) the lower quartile (Q_1).

(d) (i) Compare your values of $(Q_3 - Q_2)$ and $(Q_2 - Q_1)$.

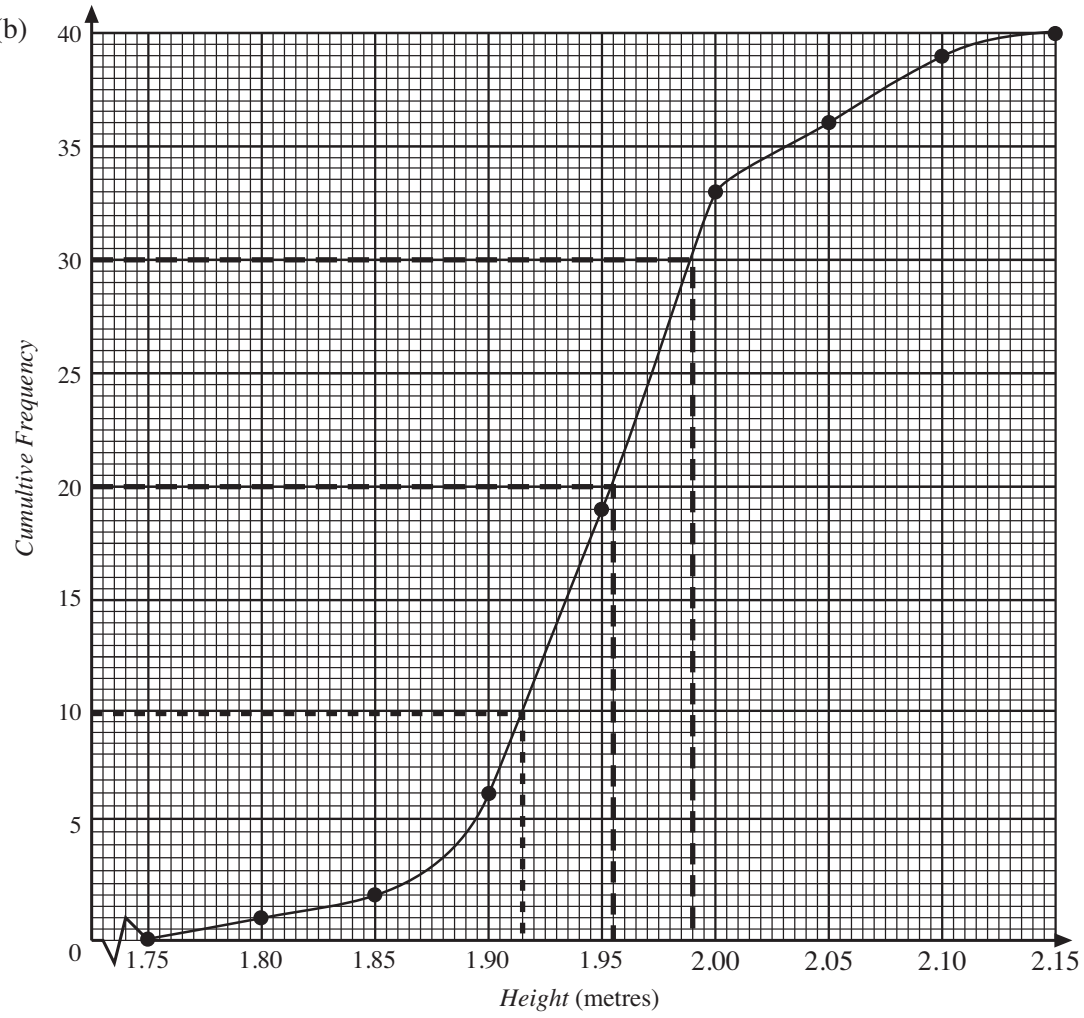
(ii) What does this indicate about the shape of the distribution?

Solution

(a) The completed cumulative frequency table is given below.

Height in metres (x)	*Frequency*	*Cumulative frequency*
$1.75 \le x < 1.80$	1	1
$1.80 \le x < 1.85$	1	2
$1.85 \le x < 1.90$	4	6
$1.90 \le x < 1.95$	13	19
$1.95 \le x < 2.00$	14	33
$2.00 \le x < 2.05$	3	36
$2.05 \le x < 2.10$	3	39
$2.10 \le x < 2.15$	1	40

(b)



(c) (i) 1.955 (ii) 1.99 (iii) 1.915

(d) (i) $Q_3 - Q_2 = 0.035$, $Q_2 - Q_1 = 0.04$

(ii) $Q_2 - Q_1 > Q_3 - Q_2$, hence (positive) *skew* to the data; that is, it is *not* symmetric.

Exercises

1. Make a cumulative frequency table for each of the three sets of data given below. Then, for each set of data, draw a cumulative frequency graph and use it to find the median and inter-quartile range.

(a) John weighed each apple in a large box. His results are given in this table.

Weight of apple (g)	$60 < w \leq 80$	$80 < w \leq 100$	$100 < w \leq 120$	$120 < w \leq 140$	$140 < w \leq 160$
Frequency	4	28	33	27	8

(b) Paul asked the students in his class how far they travelled to school each day. His results are given below.

Distance (km)	$0 < d \leq 1$	$1 < d \leq 2$	$2 < d \leq 3$	$3 < d \leq 4$	$4 < d \leq 5$	$5 < d \leq 6$
Frequency	5	12	5	6	5	3

(c) An athletics coach recorded the distances students could reach in the long jump event. His records are summarised in the table below.

Length of jump (m)	$1 < d \leq 2$	$2 < d \leq 3$	$3 < d \leq 4$	$4 < d \leq 5$	$5 < d \leq 6$
Frequency	5	12	5	6	5

2. A farmer grows a type of wheat in two different fields. He takes a sample of 50 heads of corn from each field at random and weighs the grains he obtains.

Mass of grain (g)	$0 < m \leq 5$	$5 < m \leq 10$	$10 < m \leq 15$	$15 < m \leq 20$	$20 < m \leq 25$	$25 < m \leq 30$
Frequency Field A	3	8	22	10	4	3
Frequency Field B	0	11	34	4	1	0

(a) Draw cumulative frequency graphs for each field.

(b) Find the median and inter-quartile range for each field.

(c) Comment on your results.

3. A consumer group tests two types of batteries using a DVD player.

Lifetime (hours)	$2 < l \le 3$	$3 < l \le 4$	$4 < l \le 5$	$5 < l \le 6$	$6 < l \le 7$	$7 < l \le 8$
Frequency Type A	1	3	10	22	8	4
Frequency Type B	0	2	2	38	6	0

(a) Use cumulative frequency graphs to find the median and inter-quartile range for each type of battery.

(b) Which type of battery would you recommend and why?

4. The table below shows how the height of children of a certain age vary. The data was gathered using a large-scale survey.

Height (cm)	$50 < h \le 55$	$55 < h \le 60$	$60 < h \le 65$	$65 < h \le 70$	$70 < h \le 75$	$75 < h \le 80$	$80 < h \le 85$
Frequency	100	300	2400	1300	700	150	50

A doctor wishes to be able to classify children as:

Category	*Percentage of Population*
Very Tall	5%
Tall	15%
Normal	60%
Short	15%
Very short	5%

Use a cumulative frequency graph to find the heights of children in each category.

5. The manager of a glazing company employs 30 salesmen. Each year he awards bonuses to his salesmen.

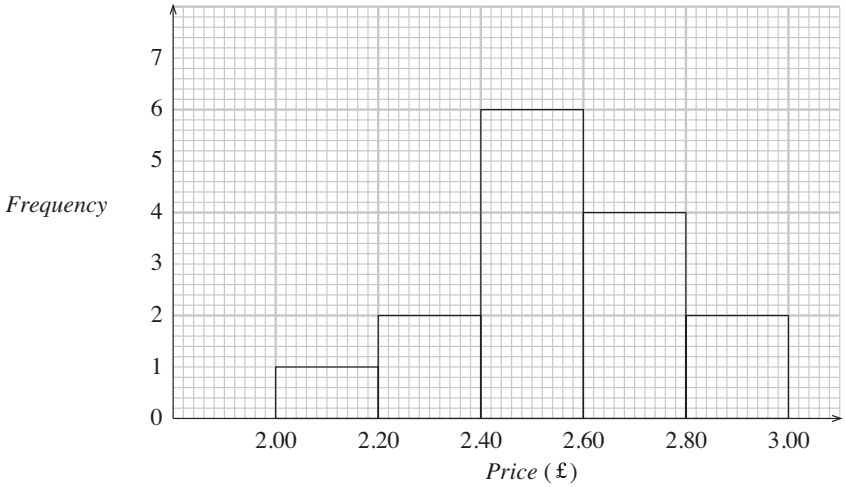
Bonus	*Awarded to*
£500	Best 10% of salesmen
£250	Middle 70% of salesmen
£50	Bottom 20% of salesmen

The sales made during 2005 and 2006 are shown in the table below.

Value of sales (£1000)	$0 < V \le 100$	$100 < V \le 200$	$200 < V \le 300$	$300 < V \le 400$	$400 < V \le 500$
Frequency 2006	0	2	15	10	3
Frequency 2005	2	8	18	2	0

Use cumulative frequency graphs to find the values of sales needed to obtain each bonus in the years 2005 and 2006.

6. The histogram shows the cost of buying a certain toy in a number of different shops.



(a) Draw a cumulative frequency graph and use it to answer the following questions.

(i) How many shops charged more than £2.65?

(ii) What is the median price?

(iii) How many shops charged less than £2.30?

(iv) How many shops charged between £2.20 and £2.60?

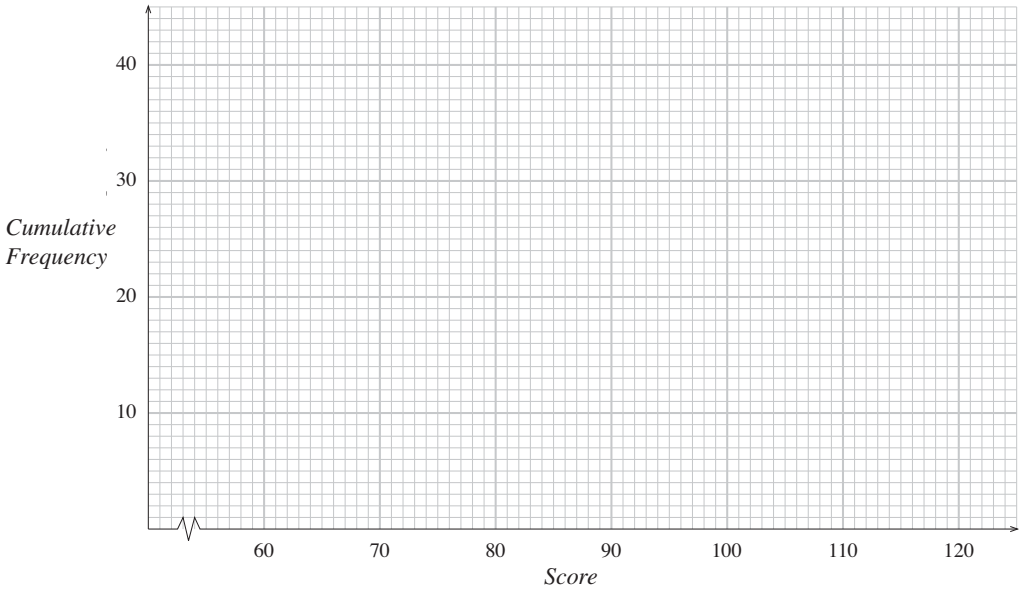
(v) How many shops charged between £2.00 and £2.50?

(b) Comment on which of your answers are exact and which are estimates.

7. Darrita and Jenine played 40 games of golf together. The table below shows Darrita's scores.

Scores (x)	$70 < x \leq 80$	$80 < x \leq 90$	$90 < x \leq 100$	$100 < x \leq 110$	$110 < x \leq 120$
Frequency	1	4	15	17	3

(a) On a grid similar to the one below, draw a cumulative frequency diagram to show Darrita's scores.



(b) Making your method clear, use your graph to find

(i) Darrita's median score,

(ii) the inter-quartile range of her scores.

(c) Jenine's median score was 103. The inter-quartile range of her scores was 6.

(i) Who was the more consistent player?
Give a reason for your choice.

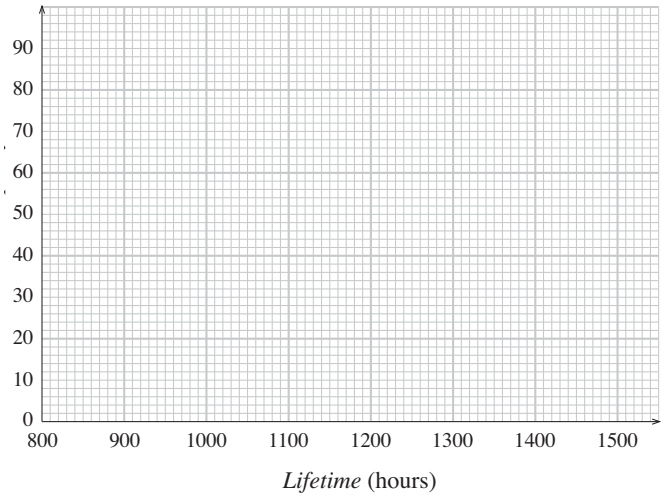
(ii) The winner of a game of golf is the one with the lowest score.
Who won most of these 40 games? Give a reason for your choice.

8. A sample of 80 electric light bulbs was taken. The lifetime of each light bulb was recorded. The results are shown below.

Lifetime (hours)	800–	900–	1000–	1100–	1200–	1300–	1400–
Frequency	4	13	17	22	20	4	0
Cumulative Frequency	4	17					

(a) Copy and complete the table of values for the cumulative frequency.

(b) Draw the cumulative frequency curve, using a grid as shown below.



(c) Use your graph to estimate the number of light bulbs which lasted more than 1030 hours.

(d) Use your graph to estimate the inter-quartile range of the lifetimes of the light bulbs.

(e) A second sample of 80 light bulbs has the same median lifetime as the first sample. Its inter-quartile range is 90 hours. What does this tell you about the difference between the two samples?

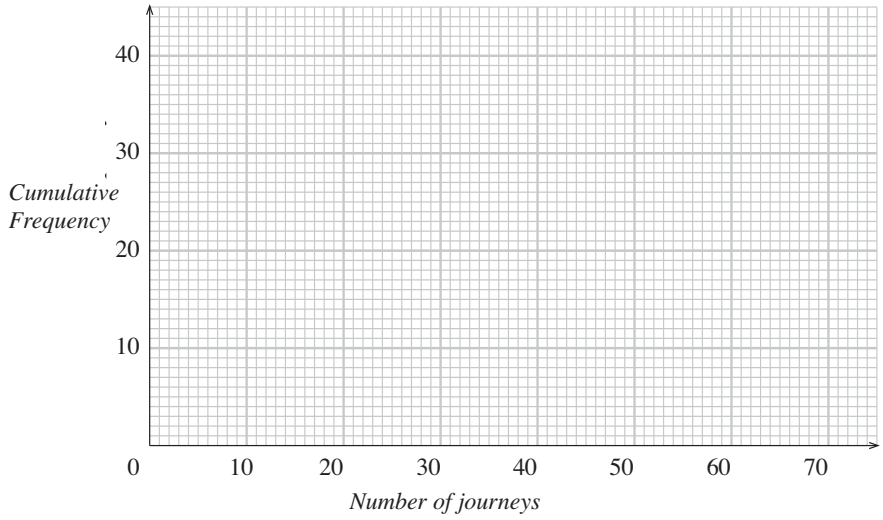
9. The number of journeys made by a group of people using taxis in one month is summarised in the table.

Number of journeys	0–10	11–20	21–30	31–40	41–50	51–60	61–70
Number of people	4	7	8	6	3	4	0

(a) Copy and complete the cumulative frequency table below.

Number of journeys	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 60	≤ 70
Cumulative frequency							

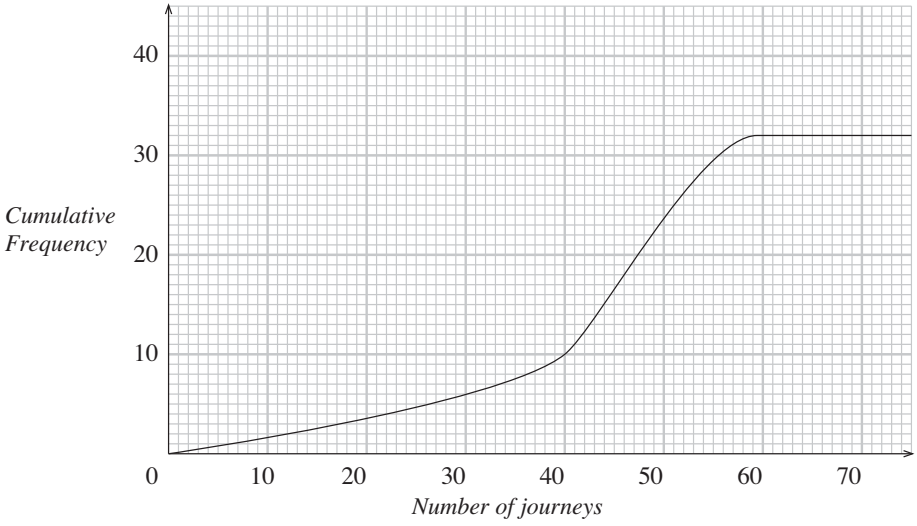
(b) (i) Draw the cumulative frequency graph, using a grid as below.



(ii) Use your graph to estimate the median number of journeys.

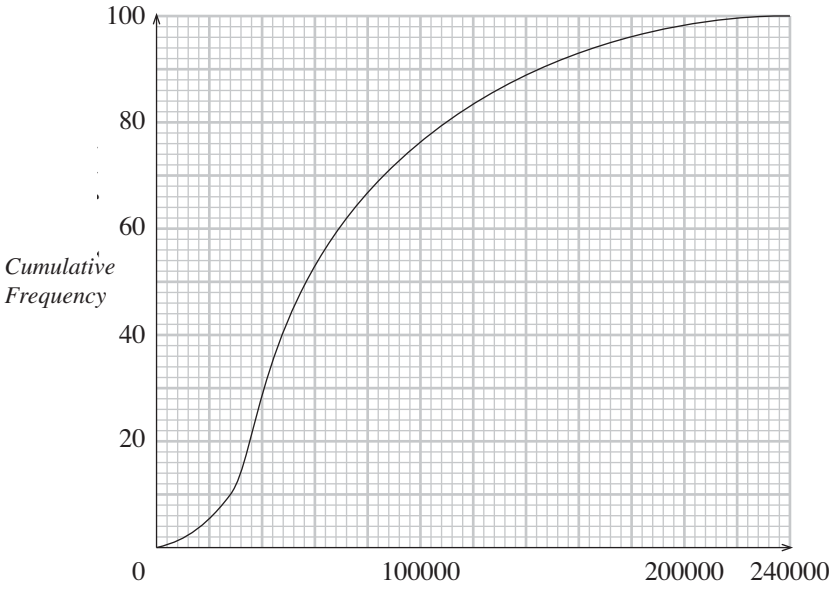
(iii) Use your graph to estimate the number of people who made more than 44 journeys in the month.

(c) The number of journeys made using taxis in one month, by another group of people, is shown in the graph.



Make **one** comparison between the numbers of journeys made by these two groups.

10. The cumulative frequency graph below gives information on holiday villa rental prices in 2002. The cumulative frequency is given as a percentage of all rental villas in the Montego Bay area of Jamaica.



Holiday villa rental prices in 2002 (J$), per week

This grouped frequency table gives the percentage distribution of villa rental prices (p) in the Montego Bay area in 2003.

Villa rental prices (p) in J$, 2003	*Percentage of villas in this class interval*
$0 \leq p < 40\,000$	26
$40\,000 \leq p < 52\,000$	19
$52\,000 \leq p < 68\,000$	22
$68\,000 \leq p < 88\,000$	15
$88\,000 \leq p < 120\,000$	9
$120\,000 \leq p < 160\,000$	5
$160\,000 \leq p < 220\,000$	4

(a) Use the data above to complete the cumulative frequency table below.

Villa rental prices (p) in J$, 2003	*Cumulative Frequency (%)*
$0 \leq p < 40\,000$	
$0 \leq p < 52\,000$	
$0 \leq p < 68\,000$	
$0 \leq p < 88\,000$	
$0 \leq p < 120\,000$	
$0 \leq p < 160\,000$	
$0 \leq p < 220\,000$	

(b) Trace or photocopy the grid for 2002, and on it construct a cumulative frequency graph for your table for 2003.

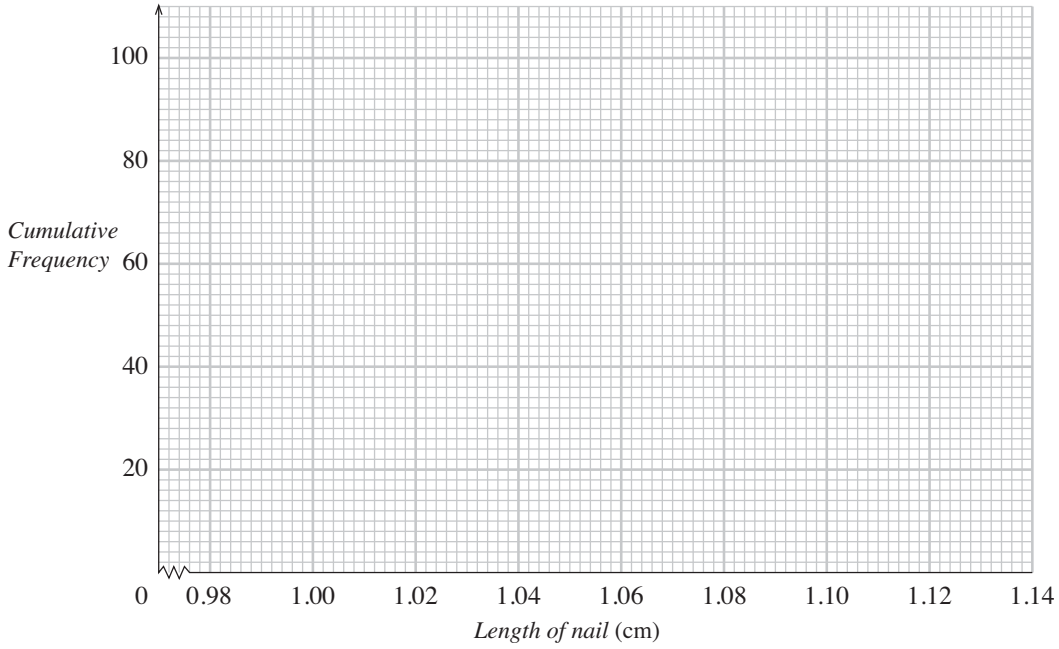
(c) In 2002 the rental for a holiday villa was J$100 000. Use both cumulative frequency graphs to estimate the rental for this house in 2003. Make your method clear.

11. The lengths of a number of carpenters' nails were measured to the nearest 0.01 cm and the following frequency distribution was obtained.

Length of nail (x cm)	*Number of nails*	*Cumulative Frequency*
$0.98 \le x < 1.00$	2	
$1.00 \le x < 1.02$	4	
$1.02 \le x < 1.04$	10	
$1.04 \le x < 1.06$	24	
$1.06 \le x < 1.08$	32	
$1.08 \le x < 1.10$	17	
$1.10 \le x < 1.12$	7	
$1.12 \le x < 1.14$	4	

(a) Complete the cumulative frequency column.

(b) Draw a cumulative frequency diagram on a grid similar to the one below.



Use your graph to estimate

(i) the median length of the nails (ii) the inter-quartile range.

12. A wedding was attended by 120 guests. The distance, d km, that each guest travelled was recorded in the frequency table below.

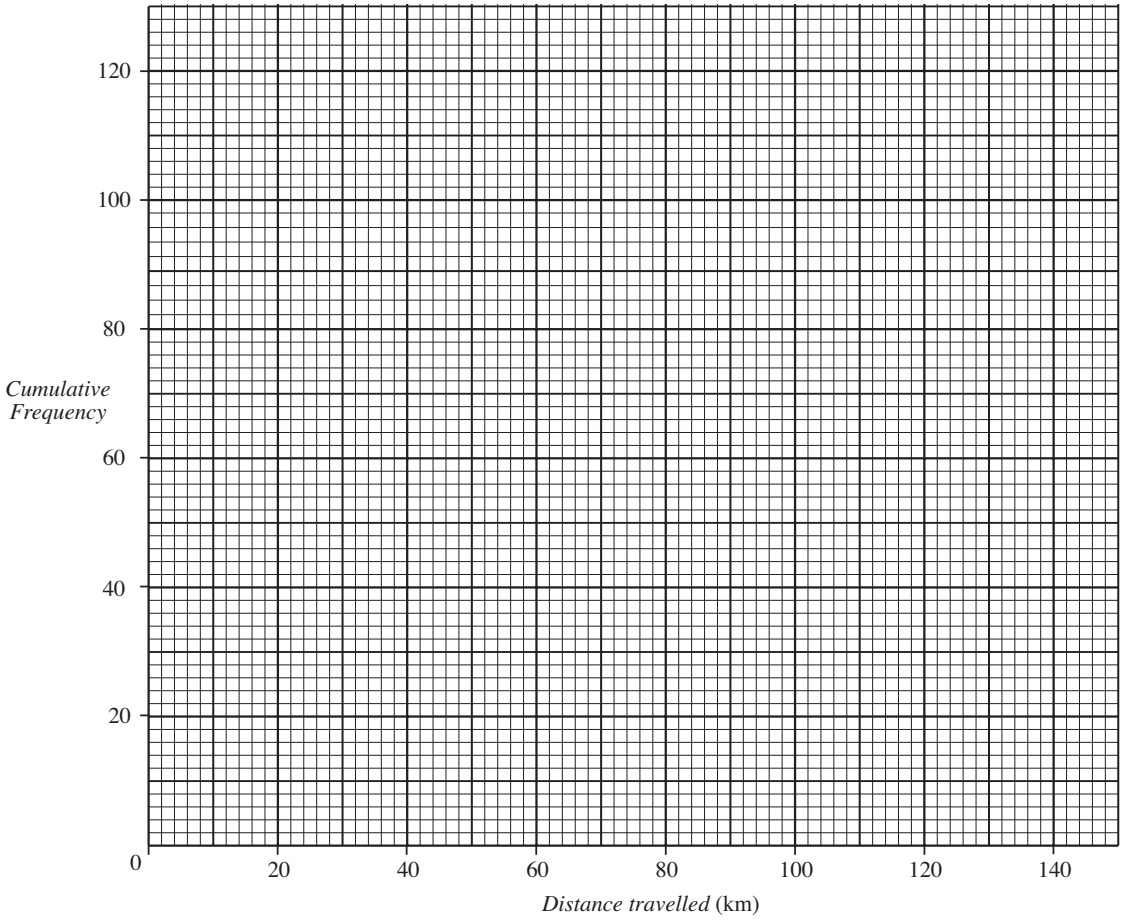
Distance (d km)	$0 < d \leq 10$	$10 < d \leq 20$	$20 < d \leq 30$	$30 < d \leq 50$	$50 < d \leq 100$	$100 < d \leq 140$
Number of guests	26	38	20	20	12	4

(a) Using the mid-interval values, calculate an estimate of the mean distance travelled.

(b) (i) Copy and complete the cumulative frequency table below.

Distance (d km)	$d \leq 10$	$d \leq 20$	$d \leq 30$	$d \leq 50$	$d \leq 100$	$d \leq 140$
Number of guests						120

(ii) On a grid similar to that shown below, draw a cumulative frequency curve to represent the information in the table.



(c) (i) Use the cumulative frequency curve to estimate the median distance travelled by the guests.

(ii) Give a reason for the large difference between the mean distance and the median distance.

13. Mr Isaacs is checking the time taken for documents to be word-processed. The first 12 documents which he samples take the following times, in minutes, to process.

5	12	13	14	16	18	19	23	34	35	40	43

(a) Calculate

(i) the median time,

(ii) the interquartile range.

His completed survey gives him the following information.

Time (minutes)	0 -	10 -	20 -	30 -	40 -	50 -	60 -	70 - 80
Number of documents	3	5	12	14	16	7	2	1

(b) Copy and complete the cumulative frequency table for this information.

Time less than (minutes)								
Number of documents	3	8	20					

(c) Use your cumulative frequency table to draw a cumulative frequency graph. Use your graph to find

(i) the median time,

(ii) the interquartile range.

You must mark your graph clearly showing how you found your answers.

(d) Compare the median times and the interquartile ranges found in the sample and the completed survey. Suggest a reason for any differences.

2 Box and Whisker Plots

The *box and whisker plot* is an important and useful way to illustrate and compare the measures of both location and variation through the median and quartiles.

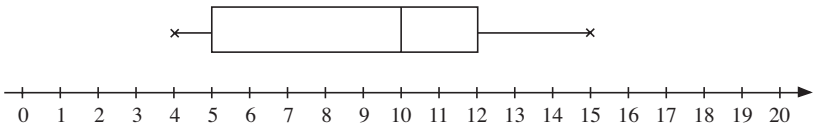
For example, for the data set below, we can easily find the median and quartiles.

4, 5, 10, 10, 11, 12, 15

↑ *lower quartile* ↑ *median* ↑ *upper quartile*

The *box i*s formed by the two quartiles, with the median marked by a line, whilst the *whiskers* are fixed by the two extreme values, 4 and 15.

The plot is shown below, relative to a scale.



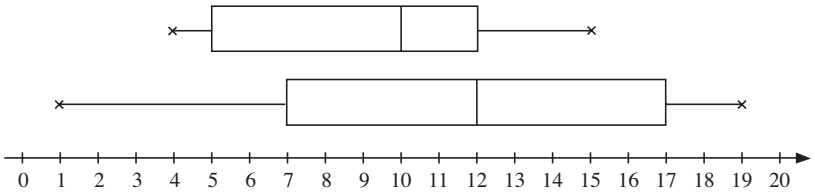
Box and whisker plots are particularly useful when comparing quickly two sets of data.

For example, if you wish to compare the data set above with the following data set,

1, 7, 9, 12, 14, 17, 19

lower quartile *median* *upper quartile*

then you can illustrate the two plots together. This is shown below – you can immediately see that the data in the second set are much more spread out than that in the first set.

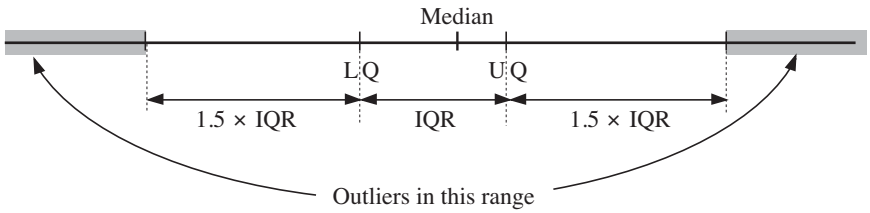


With the value of the lower quartile (Q_1), median (Q_2) an upper quartile (Q_3) together with the minimum and maximum values we have what is called the *5-number summary* of the data,

$$\left\{ \text{Min},\ Q_1, Q_2, Q_3, Max \right\}$$

Outliers

Outliers are extreme values in data sets and are often ignored as they can distort the data analysis. We make the concept precise by defining an outlier as 'any value which is either 1.5 times the interquartile range (IQR) more than the upper quartile (UQ) or 1.5 times the IQR less than the lower quartile (LQ). This is illustrated below.



Any outlier should be marked on the box and whisker diagram but the whisker should extend only to the lowest and highest values which are not outliers.

Sometimes outliers and spurious data, i.e.errors, are not relevant but in some contexts, they are important and cannot be neglected!

Worked Example 1

The number of goals scored by the 11 members of a football team in 2007 were as follows:

6 0 8 12 2 1 2 9 1 0 11

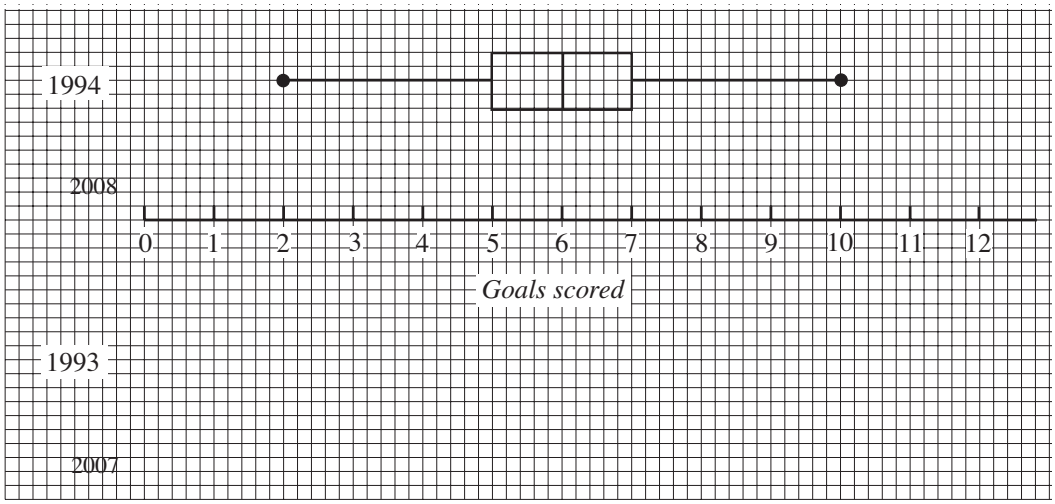
(a) Find the median.

(b) Find the upper and lower quartiles.

(c) Find the interquartile range.

(d) Explain why, for this set of data, the interquartile range is a more appropriate measure of spread than the range.

(e) The goals scored by the 11 members of the hockey team in 2008 are summarised in the box and whisker plot below.



(i) On a copy of the diagram, summarise the results for 2007 in the same way.

(ii) Do you think the team scored more goals in 2008? Explain your reasoning.

Solution

We first put the number of goals in increasing order, i.e:

0 0 1 1 2 2 6 8 9 11 12

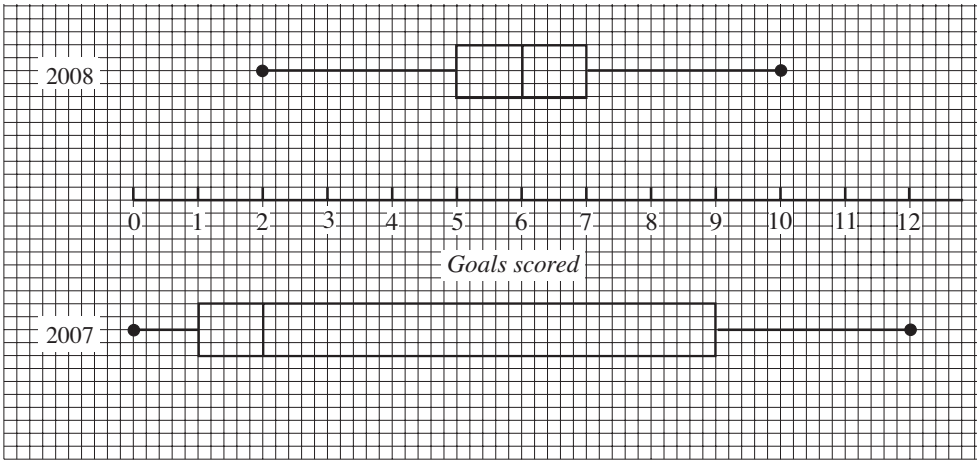
(a) There are 11 data points, so the median is the $\left(\frac{11+1}{2}\right)$ th value, i.e. the 6th value, which is 2.

(b) The lower quartile is the $\left(\frac{11+1}{4}\right)$ th value, i.e. the 3rd value, which is 1; the upper quartile is the 9th value, i.e. 9.

(c) Interquartile range $= 9 - 1 = 8$.

(d) The interquartile range is a better measure to represent the 'average' spread, rather than the range, as it excludes the outlying values.

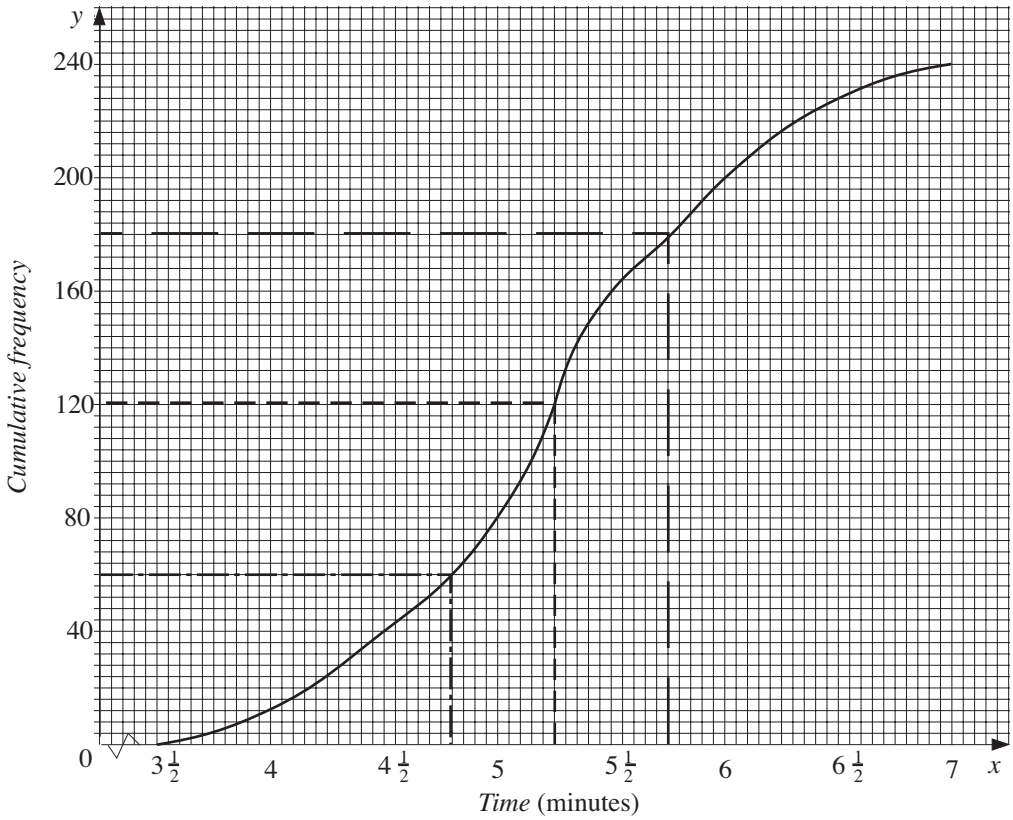
(e)



The team scored more goals in 2008; the median is much lower.

Worked Example 2

The cumulative frequency curve represents the times taken to run 1500 metres by each of the 240 members of the athletics club, Weston Harriers.



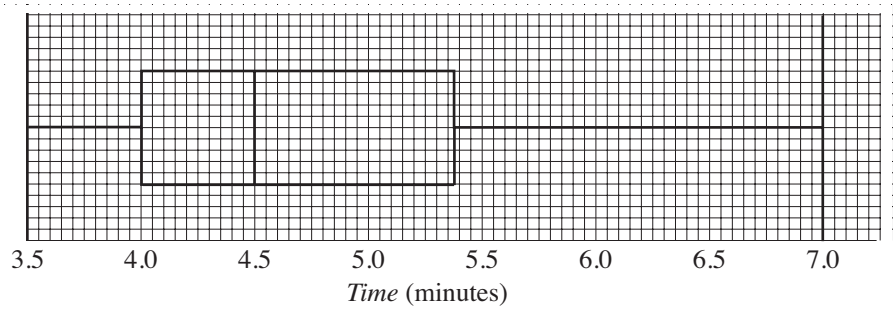
(a) From the graph, find:

(i) the median time; (ii) the upper quartile and the lower quartile.

(b) Draw a box and whisker plot to illustrate the data.

(c) Use your box and whisker plot to make *on*e comment about the shape of a histogram for these data.

A rival athletics club, Eastham Runners, also has 240 members. The time taken by each member to run 1500 metres is recorded and these data are shown in the following box and whisker plot.



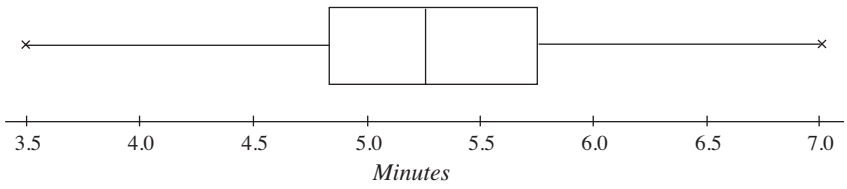
(d) Use this diagram to make *one* comment about the data for Eastham Runners compared with the data for Weston Harriers.

Solution

(a) (i) From the dashed line (120 on the vertical axis), the median is $5\frac{1}{4}$ minutes or 5 minutes 15 seconds.

(ii) Similarly:
the upper quartile is $5\frac{3}{4}$ minutes or 5 minutes 45 seconds,
the lower quartile is 4 min 48 sec (width of each small square is 3 seconds)

(b)



(c) The data are almost symmetric about the median.

(d) The data for Eastham Runners are skewed to the left, with a lower median time. Hence Eastham Runners' data are on average better than Western Harriers' data, because relatively more athletes have faster times.

Worked Example 3

The ages (in years) of a group of people visiting a club are given below.

23	31	14	27	32	34	28	29	40	29	37	27	28	20
40	76	26	31	42	34	25	26	30	40	27	52	36	

(a) Identify any outlier.

(b) Illustrate the data using a box and whisker plot.

Solution

(a) First identify the median and upper and lower quartiles of the 27 data values.

Putting the data in increasing order gives:

14	20	23	25	26	26	27	27	27	28	28	29	29
30	31	31	32	34	34	36	37	40	40	40	42	
52	76											

The *median* is the $\left(\frac{27+1}{2}\right)$ th value, i.e. the 14th value $\Rightarrow$ 30

The *lower quartile* is the $\left(\frac{27+1}{4}\right)$ th value, i.e. the 7th value $\Rightarrow$ LQ = 27

The *upper quartile* is the $\left(\frac{3(27+1)}{4}\right)$ th value, i.e. the 21st value $\Rightarrow$ UQ = 37

The *interquartile range* is 37 – 27 = 10.

Now check for outliers, remembering that the outliers must be less than

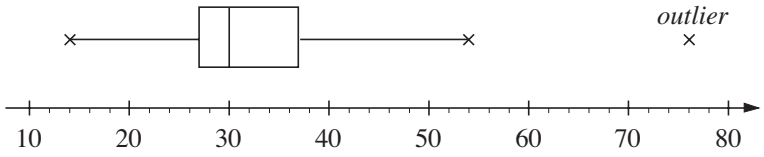
$$\text{LQ} - 1.5\times \text{IQR} = 27 - 1.5\times 10 = 27 - 15 = 12$$

or more than

$$\text{UQ} + 1.5 \times \text{IQR} = 37 + 1.5 \times 10 = 37 + 15 = 52$$

The only outlier is 76.

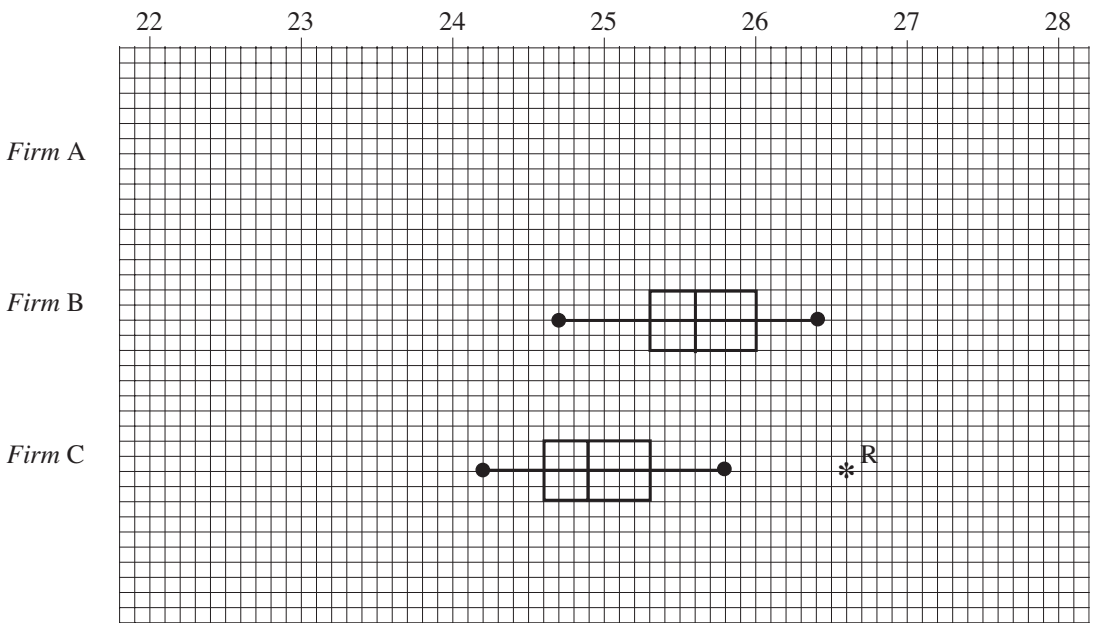
(b) This box and whisker plot illustrates the data.



Exercises

1. A manufacturing company needs to place a regular order for components. The manager investigates components produced by three different firms and measures the diameters of a sample of 25 components from each firm.

   The results of the measurements for the samples of components from Firm B and Firm C are illustrated in the two box plots shown below.



(a) (i) Find the range of the sample of measurement for Firm B.

(ii) Find the interquartile range of the sample of measurement for Firm C.

(iii) Explain why the result labelled R is shown as an outlier on the box plot for Firm C.

(b) The results of the measurements for the sample from Firm A are summarised as follows.

Median = 25.0 mm, lower quartile = 23.4 mm, upper quartile = 26.5 mm, lowest value = 22.5 mm, highest value = 27.3 mm.

Draw, on a copy of the grid above, a box plot to illustrate the sample results for Firm A.

(c) The manager studies the three box plots to decide which firm's components he should use. The components he requires should have a diameter of 25 mm, but some variation above and below this measurement is bound to happen and is acceptable. Any components with diameters below 24 mm or above 26 mm will have to be thrown away. State which firm's components you think the manager should choose. Explain carefully why you think he should choose this firm rather than the other two.

2. A random sample of 51 people were asked to record the number of km they travelled by car in a given week. The distances, to the nearest km, are shown below.

67	76	85	42	92	48	93	46
52	72	77	53	41	48	86	78
56	80	70	70	66	62	54	85
60	58	43	58	74	44	52	74
52	82	78	47	66	50	67	87
78	86	94	63	72	63	44	47
57	68	81					

(a) Find the median and the quartiles of this distribution.

(b) Draw a box plot to represent these data.

(c) Give one advantage of using the box plot to illustrate data such as that given above.

3. The weights, to the nearest kilogram, of 19 pigs were:

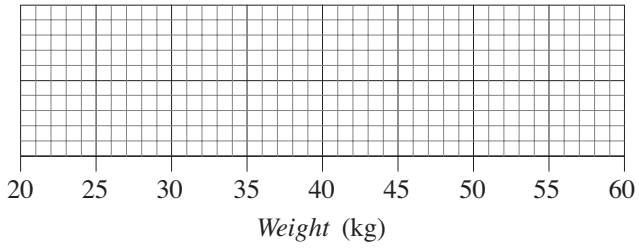
36 38 30 31 38 43 55 38 37 30 48 41 33 25 34 43 37 40 36

(a) Find the inter-quartile range of the weights.

(b) Find any weights that are outliers.

The median of the data is 37 kg.

(c) Draw a box plot for the data.



(d) Name a distribution that could be used to model the weight of these pigs. Give a reason for your choice.

The weight of a full-grown pig is about 45 kg.

(e) What does this suggest about the 19 pigs?

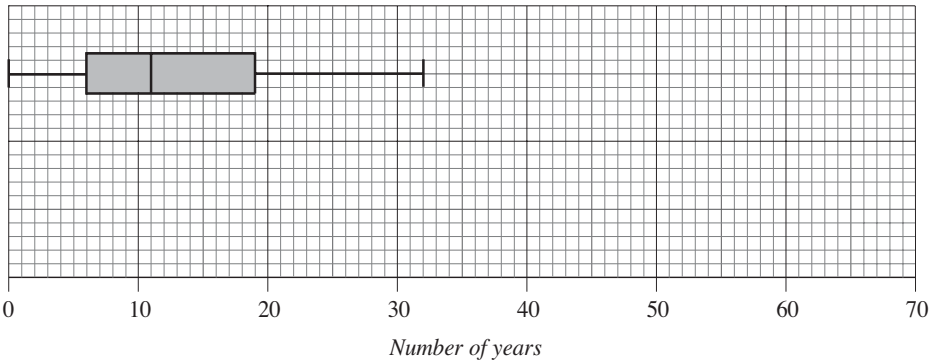
4. The length of reign of each of the last 19 English monarchs is given in the table.

George VI	16 years	George IV	10 years	James II	3 years
Edward VIII	0 years	George III	60 years	Charles II	25 years
George V	26 years	George II	33 years	Charles I	24 years
Edward VII	9 years	George I	13 years	James I	22 years
Victoria	64 years	Anne	12 years	Elizabeth I	45 years
William IV	7 years	William III	14 years	Mary	5 years
				Edward VI	6 years

(a) Find the median and quartiles of the length of reign of these 19 monarchs.

(b) Write down the name of any monarch whose length of reign is an outlier. You **must** show calculations to support your answer.

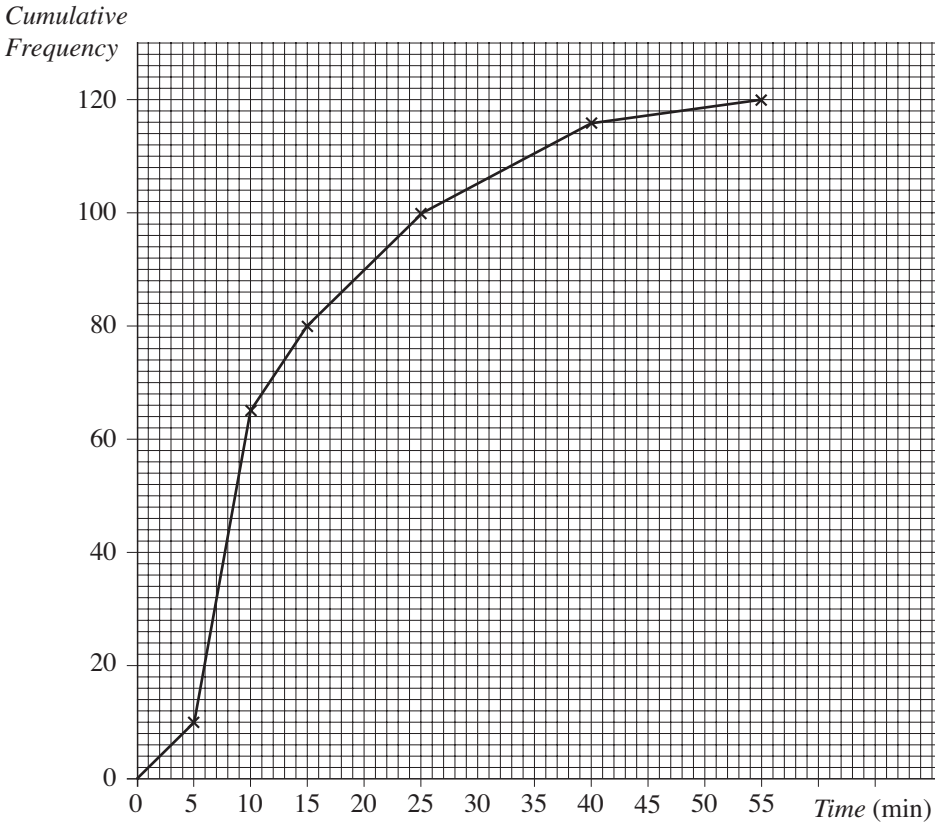
(c) The box and whisker plot shows the length of reign of the last 19 Popes.



Draw a box and whisker plot for the length of reign of the last 19 monarchs on a copy of the diagram.

(d) Compare the length of reign of monarchs and Popes.

5. The cumulative frequency polygon below shows the times taken to travel to a city centre school by a group of students.



(a) Estimate from the graph:

(i) the median;

(ii) the interquartile range;

(iii) the percentage of students taking more than 35 minutes to reach school.

(b) A school of equivalent size in a rural area showed the following distribution of times taken to travel to school.

Time taken (min)	*No. of students*	*Cumulative frequency table*	
		Time	*No. of pupils*
0 and under 5	8	<5	8
5 and under 10	44	<10	
10 and under 15	15	<15	
15 and under 25	9	<25	
25 and under 40	7	<40	
40 and under 55	37	<55	

(i) Complete a copy of the cumulative frequency table for the data.

(ii) Draw on the same axes the cumulative frequency polygon for this school, labelling the polygon clearly.

(iii) Estimate from this polygon the median and the interquartile range.

(c) Construct box and whisker plots for each set of data and comment on the main differences that are apparent between the two distributions.

6. In a village a record was kept of the ages of those people who died in 2008. The data are shown on the stem and leaf diagram.

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

Key: 2 \| 3 denotes 23 years

(a) How many people died in this village in 2008?

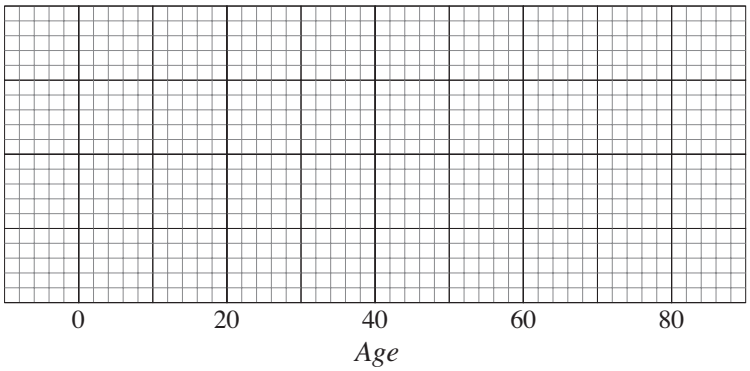
At the start of the year there were 750 people in the village.

(b) Calculate the death rate for this village.

(c) Use the stem and leaf diagram to obtain values for:

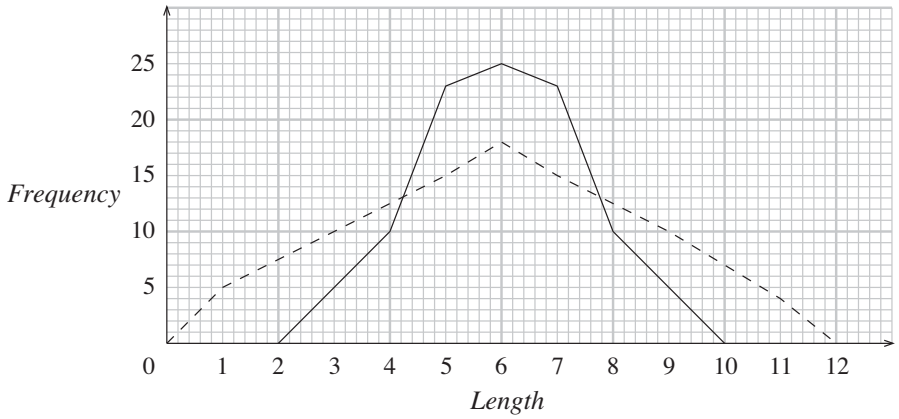
(i) the median, (ii) the lower and upper quartiles.

(d) On a copy of the diagram below, draw a box and whisker diagram to illustrate the data.



3 Standard Deviation

The two *frequency polygons* drawn on the graph below show samples which have the same mean, but the data in one are much more spread out than in the other.



The *range* (highest value – lowest value) gives a simple measure of how much the data are spread out.

Standard deviation (s.d.) is a much more useful measure and is given by the formula:

$$\text{s.d.} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

where x_i represents each datapoint $(x_1, x_2, \ldots, x_n)$
$\bar{x}$ is the mean,
n is the number of values.

Then $(x_i - \bar{x})^2$ gives the square of the difference between each value and the mean (squaring exaggerates the effect of data points far from the mean and gets rid of negative values), and

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

sums up all these squared differences.

The expression

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

gives an average value to these differences. If all the data were the same, then each x_i would equal $\bar{x}$ and the expression would be zero.

Finally we take the square root of the expression so that the dimensions of the standard deviation are the same as those of the data.

So standard deviation is a measure of the spread of the data. The greater its value, the more spread out the data are. This is illustrated by the two frequency polygons shown above. Although both sets of data have the same mean, the data represented by the 'dotted' frequency polygon will have a greater standard deviation than the other.

Worked Example 1

Find the mean and standard deviation of the numbers,

$$6,\ 7,\ 8,\ 5,\ 9$$

Solution

The mean, $\bar{x}$, is given by,

$$\bar{x} = \frac{6 + 7 + 8 + 5 + 9}{5}$$

$$= \frac{35}{5}$$

$$= 7$$

Now the standard deviation can be calculated.

$$\text{s.d.} = \sqrt{\frac{(6-7)^2 + (7-7)^2 + (8-7)^2 + (5-7)^2 + (9-7)^2}{5}}$$

$$= \sqrt{\frac{1 + 0 + 1 + 4 + 4}{5}}$$

$$= \sqrt{\frac{10}{5}}$$

$$= \sqrt{2}$$

$$= 1.414 \quad \text{(to 3 decimal places)}$$

An alternative formula for standard deviation is

$$\text{s.d.} = \sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2}}{n} - \bar{x}^2}$$

This expression is much more convenient for calculations done without a calculator. The proof of the equivalence of this formula is given below although it is beyond the scope of the GCSE syllabus.

Proof

For completeness, we will show a proof of this result but the important point is to be able to use the result.

Note that
$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n}\left(x_i^{\,2} - 2x_i\,\bar{x} + \bar{x}^2\right)$$

$$= \sum_{i=1}^{n} x_i^{\,2} - \sum_{i=1}^{n}\left(2x_i\,\bar{x}\right) + \sum_{i=1}^{n}\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1$$

(since the expressions $2\bar{x}$ and $\bar{x}^2$ are common for each term in the summation).

But $\sum_{i=1}^{n} 1 = n$, since you are summing $\underbrace{1 + 1 + \ldots + 1}_{n \text{ terms}} = n$, and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$, by definition, thus

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\left\{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2 n\right\} \quad \left(\text{substituting } \sum_{i=1}^{n} 1 = n\right)$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\bar{x}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) + \bar{x}^2 \qquad \text{(dividing by } n\text{)}$$

$$= \frac{\sum_{i=1}^{n} x_i^{\,2}}{n} - 2\bar{x}^2 + \bar{x}^2 \qquad \left(\text{substituting } \bar{x} \text{ for } \frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$$

and the result follows.

Worked Example 2

Find the mean and standard deviation of each of the following sets of numbers.

(a) 10, 11, 12, 13, 14 (b) 5, 6, 12, 18, 19

Solution

(a) The mean, $\bar{x}$, is given by

$$\bar{x} = \frac{10 + 11 + 12 + 13 + 14}{5}$$

$$= \frac{60}{5}$$

$$= 12$$

The standard deviation can now be calculated using the alternative formula.

$$\text{s.d.} = \sqrt{\left(\frac{10^2 + 11^2 + 12^2 + 13^2 + 14^2}{5}\right) - 12^2}$$

$$= \sqrt{146 - 144}$$

$$= \sqrt{2}$$

$$= 1.414 \quad \text{(to 3 decimal places)}$$

(b) The mean, $\bar{x}$, is given by

$$\bar{x} = \frac{5 + 6 + 12 + 18 + 19}{5}$$

$$= 12 \quad \text{(as in part (a))}$$

The standard deviation is given by

$$\text{s.d.} = \sqrt{\left(\frac{5^2 + 6^2 + 12^2 + 18^2 + 19^2}{5}\right) - 12^2}$$

$$= \sqrt{178 - 144}$$

$$= 5.8 \quad \text{(to 1 decimal place)}$$

Note that both sets of numbers have the same mean value, but that set (b) has a much larger standard deviation. This is expected, as the spread in set (b) is clearly far more than in set (a).

Worked Example 3

The table below gives the number of road traffic accidents per day in a small town.

Accidents per day	0	1	2	3	4	5	6
Frequency	5	8	6	3	2	1	1

Find the mean and standard deviation of this data.

Solution

The necessary calculations for each datapoint, x_i, are set out below.

Accidents per day (x_i)	*Frequency* (f_i)	x_i^2	$x_i f_i$	$x_i^2 f_i$
0	5	0	0	0
1	8	1	8	8
2	6	4	12	24
3	3	9	9	27
4	2	16	8	32
5	0	25	0	0
6	1	36	6	36
TOTALS	**25**		**43**	**127**

From the totals,

$$n = 25, \quad \sum_{i=1}^{n} x_i f_i = 43, \quad \sum_{i=1}^{n} x_i^2 = 127$$

The mean, $\bar{x}$, is now given by

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$$

$$= \frac{43}{25}$$

$$= 1.72$$

The standard deviation is now given by

$$\text{s.d.} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 f_i}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{127}{25} - 1.72^2}$$

$$= 1.5 \quad \text{(to 1 decimal place)}$$

Most scientific calculators have statistical functions which will calculate the mean and standard deviation of a set of data.

Exercises

1. (a) Find the mean and standard deviation of each set of data given below.

A	51	56	51	49	53	62
B	71	76	71	69	73	82
C	102	112	102	98	106	124

(b) Describe the relationship between each set of numbers and also the relationship between their means and standard deviations.

2. Two machines, *A* and *B*, fill empty packets with soap powder. A sample of packets was taken from each machine and the weight of powder (in kg) was recorded.

A	2.27	2.31	2.18	2.2	2.26	2.24			
B	2.78	2.62	2.61	2.51	2.59	2.67	2.62	2.68	2.70

(a) Find the mean and standard deviation for each machine.

(b) Which machine is most consistent?

3. Two groups of students were trying to find the acceleration due to gravity. Each group conducted 5 experiments.

Group A	9.4	9.6	10.2	10.8	10.1
Group B	9.5	9.7	9.6	9.4	9.8

Find the mean and standard deviation for each group, and comment on their results.

4. The number of matches per box was counted for 100 boxes of matches.
The results are given in the table below.

Number of Matches	*Frequency*
44	28
45	31
46	14
47	15
48	8
49	2
50	2

Find the mean and standard deviation of this data.

5. When two dice were thrown 50 times the total scores shown below were obtained.

Score	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	0	4	8	12	9	7	5	3	1	0

Find the mean and standard deviation of these scores.

6. The length of telephone calls from an office was recorded. The results are given in the table below.

Length of call (mins)	$0 < t \leq 0.5$	$0.5 < t \leq 1.0$	$1.0 < t \leq 2.0$	$2.0 < t \leq 5.0$
Frequency	8	10	12	4

Estimate the mean and standard deviation using this table.

7. The charges (to the nearest £) made by a jeweller for repair work on jewellery in one week are given in the table below.

Charge (£)	20 – 29	30 – 49	50 – 99	100 – 149	150 – 199	200 – 300
Frequency	10	22	6	2	4	1

Use this table to estimate the mean and standard deviation.

8. Thirty families were selected at random in two different countries. They were asked how many children there were in each family.

Country A		
1	2	1
2	0	1
2	2	3
2	1	1
2	2	2
3	1	0
5	1	4
0	2	1
1	1	0
2	0	2

Country B		
2	2	1
4	1	1
2	3	5
3	5	1
9	4	2
6	5	1
4	2	2
4	4	5
3	2	2
4	7	0

Find the mean and standard deviation for each country and comment on the results.

9. (a) Calculate the standard deviation of the numbers

$$3,\ 4,\ 5,\ 6,\ 7.$$

(b) Show that the standard deviation of *every* set of five consecutive integers is the same as the answer to part (a).

10. Ten students sat a test in Mathematics, marked out of 50. The results are shown below for each student.

25, 27, 35, 4, 49, 10, 12, 45, 45, 48

(a) Calculate the mean and standard deviation of the data.

The same students also sat an English test, marked out of 50. The mean and standard deviation are given by

mean = 30, standard deviation = 3.6.

(b) Comment on and contrast the results in Mathematics and English.

11. Ten boys sat a test which was marked out of 50. Their marks were

28, 42, 35, 17, 49, 12, 48, 38, 24 and 27

(a) Calculate

(i) the mean of the marks,

(ii) the standard deviation of the marks.

Ten girls sat the same test. Their marks had a mean of 30 and a standard deviation of 6.5.

(b) Compare the performances of the boys and girls.

12. There are twenty students in class *A* and twenty students in class *B*. All the students in class *A* were given an I.Q. test. Their scores on the test are given below.

100, 104, 106, 107, 109, 110, 113, 114, 116, 117,
118, 119, 119, 121, 124, 125, 127, 127, 130, 134.

(a) The mean of their scores is 117. Calculate the standard deviation.

(b) Class *B* takes the same I.Q. test. They obtain a mean of 110 and a standard deviation of 21. Compare the data for class *A* and class *B*.

(c) Class *C* has only 5 students. When they take the I.Q. test they all score 105. What is the value of the standard deviation for class *C*?

13. The following are the scores in a test for a set of 15 students.

5	4	8	7	3
6	5	9	6	10
7	8	6	4	2

(a) (i) Calculate the mean score.

(ii) Calculate the standard deviation of the scores.

A set of 10 different students took the same test. Their scores are listed below.

5	6	6	7	7
4	7	8	3	7

(b) After making any necessary calculations for the second set, compare the two sets of scores. Your answer should be understandable to someone who does not study Statistics.

14. In a survey on examination qualifications, 50 people were asked,

"*How many subjects are listed on your GCSE certificate?*"

The frequency distribution of their responses is recorded in the table below.

Number of subjects	1	2	3	4	5	6	7
Number of people	5	3	7	8	9	10	8

(a) Calculate the mean and standard deviation of the distribution.

(b) *A Normal Distribution has approximately 68% of its data values within one standard deviation of the mean.*

Use your answers to part (a) to check if the given distribution satisfies this property of a Normal Distribution. Show your working clearly.