Part G

Chapter 5

Descriptive statistics

Contents:

▶ **Mode**

- The mode's main advantage as a representative figure is that it is the most usual value within a set of data.
- The mode has an advantage over the mean in that it is not affected by extreme values contained in the data.
- The main disadvantage is that it does not take into account all values within the data and this makes its representativeness questionable.

G MEASURING THE SPREAD OF DATA

If, in addition to having measures of the middle of a data set, we also have an indication of the **spread** of the data, then a more accurate picture of the data set is possible.

For example, 1, 4, 5, 5, 6, 7, 8, 9, 9 has a mean value of 6 and so does
4, 4, 5, 6, 6, 7, 7, 7, 8. However, the first data set is more widely spread than the second one.

Two commonly used statistics that indicate the spread of a set of data are:

- the range
- the interquartile range.

THE RANGE

The **range** is the difference between the **maximum** (largest) data value and the **minimum** (smallest) data value.

range = maximum data value − minimum data value

Example 14

Find the range of the data set: 4, 7, 5, 3, 4, 3, 6, 5, 7, 5, 3, 8, 9, 3, 6, 5, 6

Searching through the data we find: minimum value = 3 maximum value = 9

$\therefore$ range $= 9 - 3 = 6$

THE UPPER AND LOWER QUARTILES AND THE INTERQUARTILE RANGE

The median divides the ordered data set into two halves and these halves are divided in half again by the **quartiles**.

The middle value of the lower half is called the **lower quartile** (Q_1). One-quarter, or 25%, of the data have a value less than or equal to the lower quartile. 75% of the data have values greater than or equal to the lower quartile.

The middle value of the upper half is called the **upper quartile** (Q_3). One-quarter, or 25%, of the data have a value greater than or equal to the upper quartile. 75% of the data have values less than or equal to the upper quartile.

interquartile range = upper quartile − lower quartile

The interquartile range is the range of the middle half (50%) of the data.

The data set has been divided into quarters by the lower quartile (Q_1), the median (Q_2) and the upper quartile (Q_3).

So, the **interquartile range**, is $\mathbf{IQR = Q_3 - Q_1}$.

Example 15

For the data set 6, 7, 3, 7, 9, 8, 5, 5, 4, 6, 6, 8, 7, 6, 6, 5, 4, 5, 6 find the:

a median b lower quartile
c upper quartile d interquartile range

The ordered data set is:

3 4 4 5 5 5 5 6 6 6 6 6 6 7 7 7 8 8 9 (19 of them)

a The median $= \left(\frac{19+1}{2}\right)$th score $=$ 10th score $= 6$

b/c As the median is a data value we ignore it and split the remaining data into two groups.

3 4 4 5 5 5 5 6 6 6 6 6 7 7 8 8 9

Q_1 = median of lower half $= 5$

Q_3 = median of upper half $= 7$

d $\text{IQR} = Q_3 - Q_1 = 2$

Example 16

For the data set 9, 8, 2, 3, 7, 6, 5, 4, 5, 4, 6, 8, 9, 5, 5, 5, 4, 6, 6, 8 find the:

a median b lower quartile
c upper quartile d interquartile range

The ordered data set is:

2 3 4 4 4 5 5 5 5 5 6 6 6 6 7 8 8 8 9 9 (20 of them)

a As $n = 20$, $\frac{n+1}{2} = \frac{21}{2} = 10.5$

$$\therefore \text{ median} = \frac{\text{10th value} + \text{11th value}}{2} = \frac{5+6}{2} = 5.5$$

b/c As the median is not a data value we split the original data into two equal groups of 10.

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

$\therefore Q_1 = 4.5$ $\therefore Q_3 = 7.5$

d $\text{IQR} = Q_3 - Q_1 = 3$

EXERCISE 5G

1 For each of the following sets of data find:

i the upper quartile ii the lower quartile
iii the interquartile range iv the range

a 2, 3, 4, 7, 8, 10, 11, 13, 14, 15, 15

b 35, 41, 43, 48, 48, 49, 50, 51, 52, 52, 52, 56

c

Stem	*Leaf*
1	3 5 7 7 9
2	0 1 3 4 6 7 8 9
3	0 1 2 7
4	2 6
5	1

Scale: 4 | 2 means 42

d

Score	0	1	2	3	4	5
Frequency	1	4	7	3	3	1

2 The time spent (in minutes) by 24 people in a queue at a bank, waiting to be attended by a teller, has been recorded as follows:

0	3.2	0	2.4	3.2	0	1.3	0
1.6	2.8	1.4	2.9	0	3.2	4.8	1.7
3.0	0.9	3.7	5.6	1.4	2.6	3.1	1.6

a Find the median waiting time and the upper and lower quartiles.

b Find the range and interquartile range of the waiting time.

c Copy and complete the following statements:

i "50% of the waiting times were greater than ..... minutes."

ii "75% of the waiting times were less than ..... minutes."

iii "The minimum waiting time was ..... minutes and the maximum waiting time was ..... minutes. The waiting times were spread over ..... minutes."

3

Stem	*Leaf*
6	0 3 8
7	0 1 5 6 7 7
8	1 1 2 4 4 8 9 9
9	0 4 7 9
10	1

Scale: 7 | 5 means 7.5

For the data set given, find:

a the minimum value b the maximum value
c the median d the lower quartile
e the upper quartile f the range
g the interquartile range

INTERQUARTILE RANGE FROM CUMULATIVE FREQUENCY GRAPHS

The IQR of a distribution of grouped scores can also be obtained using a cumulative frequency graph.

Remember that: $\text{IQR} = Q_3 - Q_1$

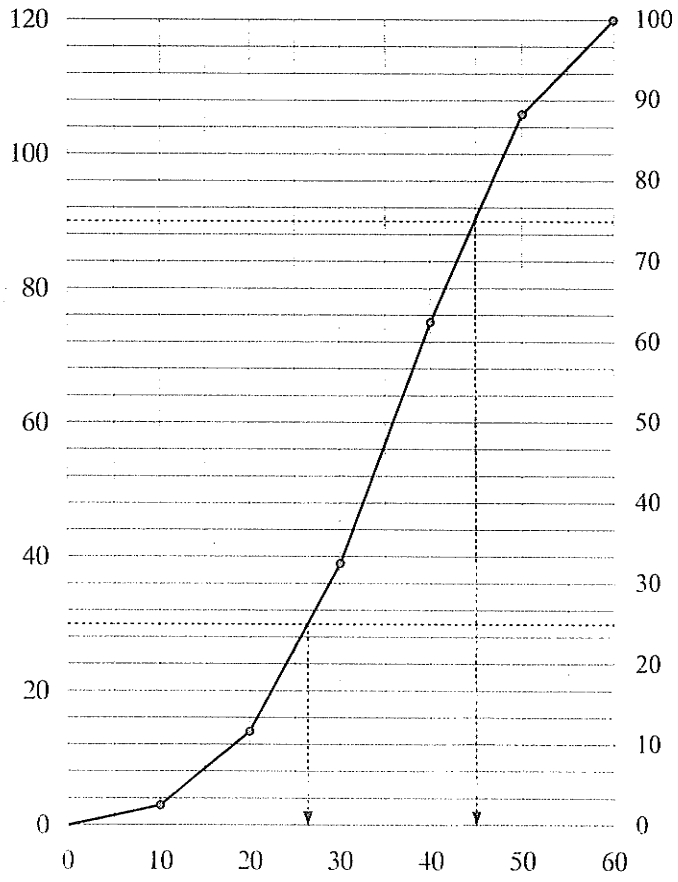
$= 75\text{th percentile} - 25\text{th percentile}$

Example 17

Draw a cumulative frequency graph for the following distribution and hence determine the interquartile range.

Scores	*Frequencies*
1 - 9.99	3
10 - 19.99	11
20 - 29.99	25
30 - 39.99	36
40 - 49.99	31
50 - 59.99	14

Scores	*Upper End point*	*Frequency*	*Cumulative Frequency*
0 - 9.99	9.995	3	3
10 - 19.99	19.995	11	14
20 - 29.99	29.995	25	39
30 - 39.99	39.995	36	75
40 - 49.99	49.995	31	106
50 - 59.99	59.995	14	120
Total		120	



75% of 120 = 90

$\therefore$ 75th percentile is $\doteqdot$ 90th score.

25% of 120 = 30

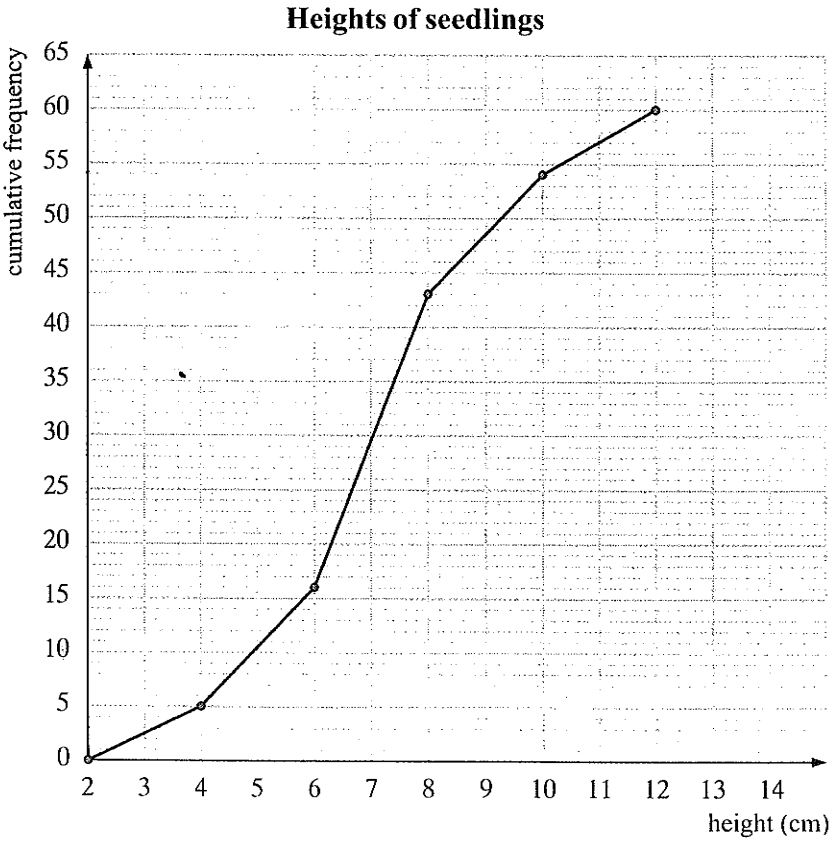
$\therefore$ 25th percentile is $\doteqdot$ 30th score.

IQR = 75th percentile − 25th percentile

$\doteqdot 45 - 26$

$\doteqdot 19$

4 A botanist has measured the heights of 60 seedlings and has presented her findings on the cumulative frequency graph below.



a How many seedlings have heights of 5 cm or less?

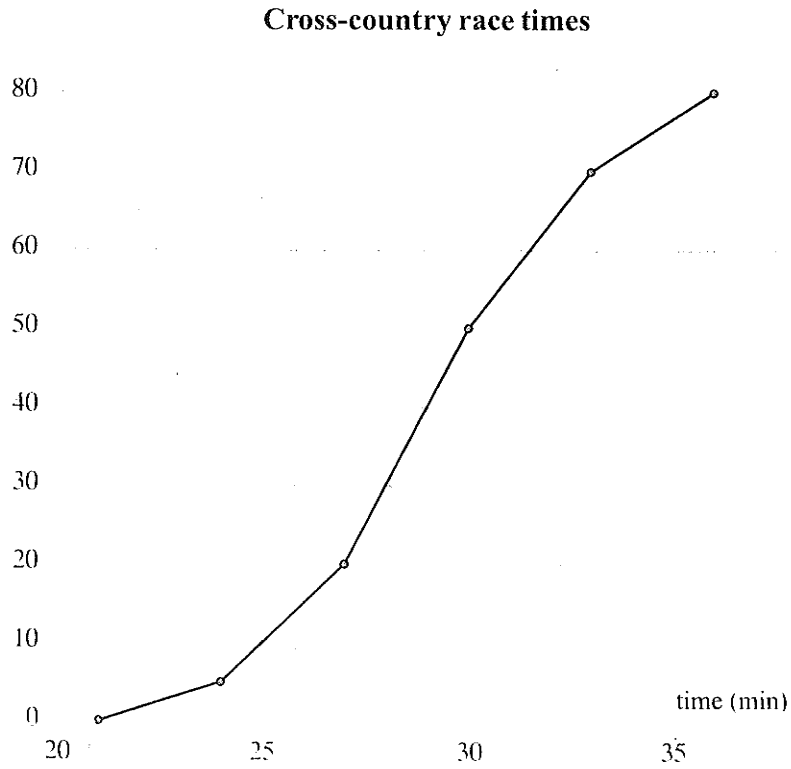
b What percentage of seedlings are taller than 8 cm?

c What is the median height?

d What is the interquartile range for the heights?

e Find the 90th percentile for the data and explain what your answer means.

5 The following cumulative frequency graph displays the performance of 80 competitors in a cross-country race.



Find:

a the lower quartile time

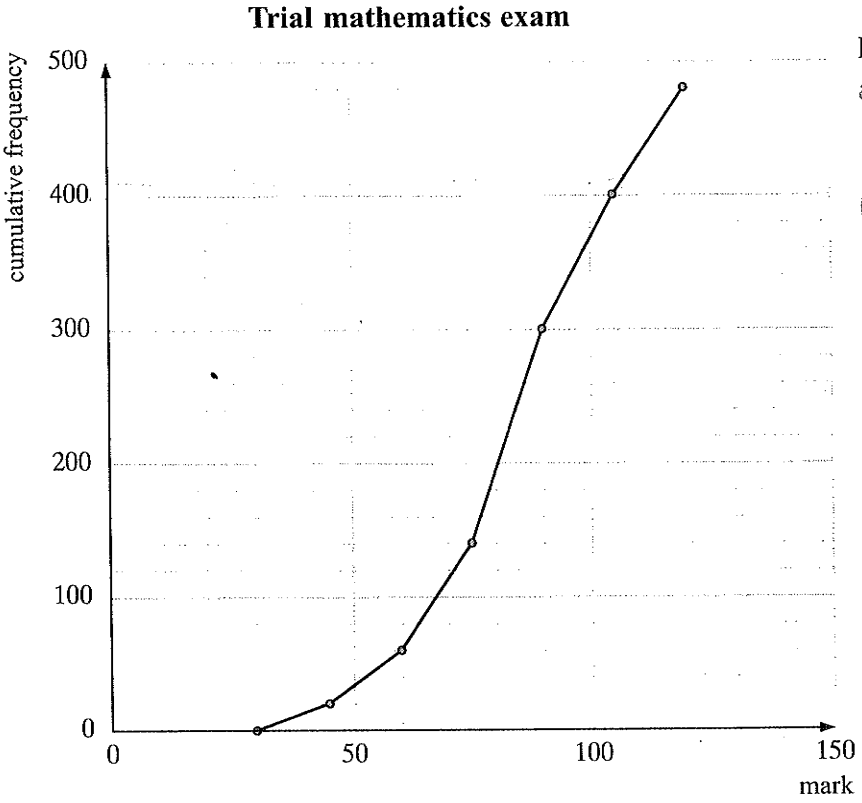
b the median

c the upper quartile

d the interquartile range

e an estimate of the 40th percentile.

6 The cumulative frequency graph below displays the marks scored by year 12 students from a cluster of schools in a common trial mathematics exam.



Find:

a how many students sat for the examination

b the probable maximum possible mark for the exam

c the median mark

d the interquartile range

e an estimate of the 85th percentile.

H BOX-AND-WHISKER PLOTS

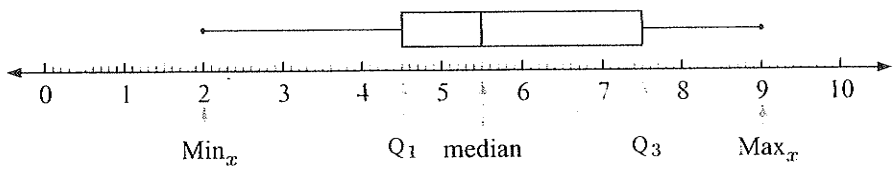
A **box-and-whisker plot** (or simply a **boxplot**) is a visual display of some of the descriptive statistics of a data set. It shows:

- the minimum value (Min_x)
- the lower quartile (Q_1)
- the median (Q_2)
- the upper quartile (Q_3)
- the maximum value (Max_x)

} These five numbers form what is known as the **five-number summary** of a data set.

In **Example 16** the five-number summary and the corresponding boxplot is:

minimum $= 2$
$Q_1 = 4.5$
median $= 5.5$
$Q_3 = 7.5$
maximum $= 9$



Note:
- The rectangular box represents the 'middle' half of the data set.
- The lower whisker represents the 25% of the data with smallest values.
- The upper whisker represents the 25% of the data with greatest values.