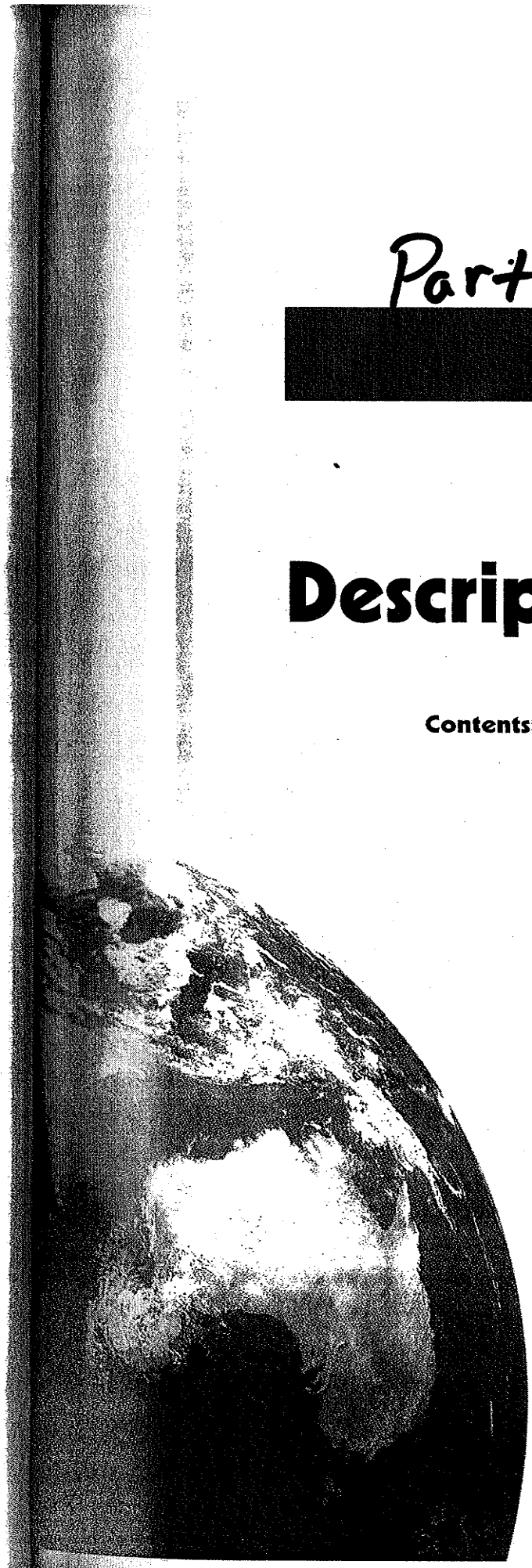


Part H

Chapter 5

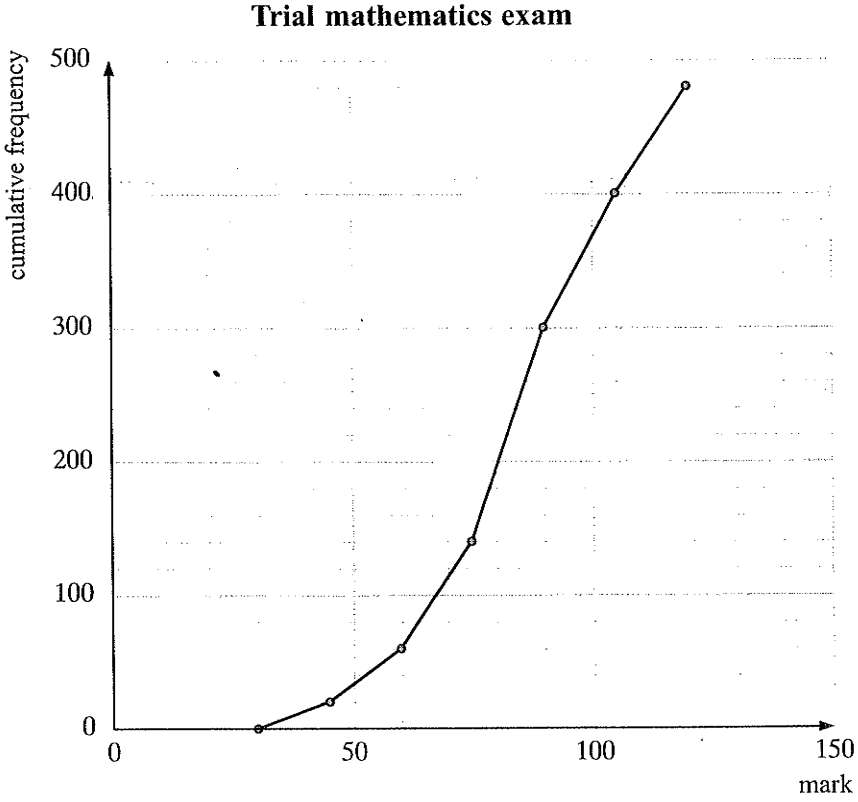
Descriptive statistics

Contents:

A Describing data
Investigation 1: Statistics from the internet
B Presenting and interpreting data
C Grouped discrete data
Investigation 2: Taxi Sir?
D Continuous data
Investigation 3: Choosing class intervals
E Frequency distribution tables
F Summarising the data
Investigation 4: Effects of outliers
G Measuring the spread of data
H Box-and-whisker plots
I The standard deviation
Investigation 5: Heart stopper
J Statistics using technology
K Parallel boxplots
Investigation 6: How do you like your eggs?

Review set 5A
Review set 5B

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

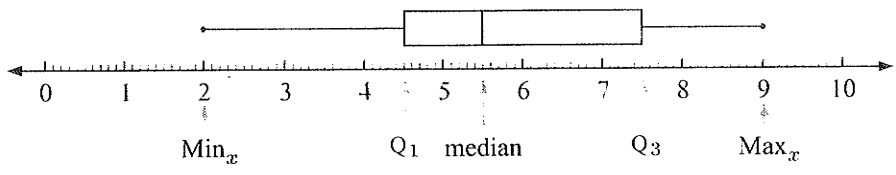
A **box-and-whisker plot** (or simply a **boxplot**) is a visual display of some of the descriptive statistics of a data set. It shows:

- 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.

Example 18

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

a construct the five-number summary b draw a boxplot

c find the i range ii interquartile range

d find the percentage of data values below 7

a The ordered data set is:

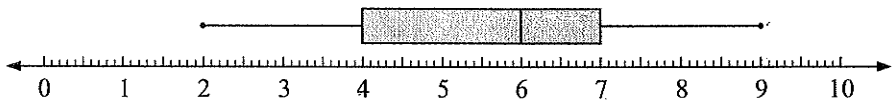
2 3 4 4 4 5 5 6 6 6 6 6 7 7 8 8 9 (17 of them)

$Q_1 = 4$ median $= 6$ $Q_3 = 7$

So the **5-number summary** is:

$$\begin{cases} \text{Min}_x = 2 \\ Q_1 = 4 \\ \text{median} = 6 \\ Q_3 = 7 \\ \text{Max}_x = 9 \end{cases}$$

b



c i range $= \text{Max}_x - \text{Min}_x$
$= 9 - 2$
$= 7$

ii IQR $= Q_3 - Q_1$
$= 7 - 4$
$= 3$

d 75% of the data values are less than or equal to 7.

BOXPLOTS AND OUTLIERS

Outliers are extraordinary data that are usually separated from the main body of the data. Outliers are either much larger or much smaller than most of the data.

There are several 'tests' that identify data that are outliers.

A commonly used test involves the calculation of 'boundaries':

- **The upper boundary = upper quartile + 1.5 × IQR.**
  Any data larger than the upper boundary is an outlier.
- **The lower boundary = lower quartile − 1.5 × IQR.**
  Any data smaller than the lower boundary is an outlier.

Outliers are marked with an asterisk on a boxplot and it is possible to have more than one outlier at either end.

The whiskers extend to the last value that is not an outlier.

Example 19

Draw a boxplot for the following data, testing for outliers and marking them, if they exist, with an asterisk on the boxplot:

3, 7, 8, 8, 5, 9, 10, 12, 14, 7, 1, 3, 8, 16, 8, 6, 9, 10, 13, 7

The ordered data set is:

1 3 3 5 6 7 7 7 8 8 | 8 8 9 9 10 10 12 13 14 16 $(n = 20)$

Min_x	Q_1	median	Q_3	Max_x
$= 1$	$= 6.5$	$= 8$	$= 10$	$= 16$

$\text{IQR} = Q_3 - Q_1 = 3.5$

Test for outliers:

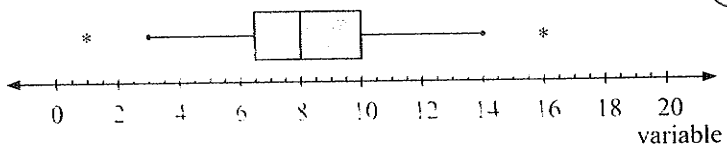
upper boundary
$= \text{upper quartile} + 1.5 \times \text{IQR}$
$= 10 + 1.5 \times 3.5$
$= 15.25$

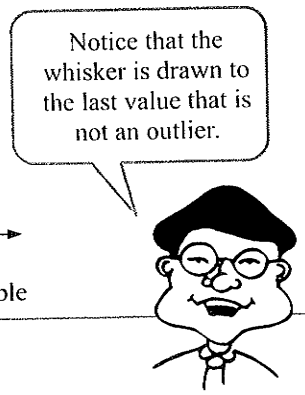
and

lower boundary
$= \text{lower quartile} - 1.5 \times \text{IQR}$
$= 6.5 - 1.5 \times 3.5$
$= 1.25$

As 16 is above the upper boundary it is an outlier.
As 1 is below the lower boundary it is an outlier.

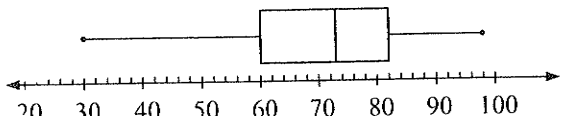
So, the boxplot is:





EXERCISE 5H

1 A boxplot has been drawn to show the distribution of marks (out of 100) in a test for a particular class.



a What was the i highest mark ii lowest mark scored?
b What was the median test score for this class?
c What was the range of marks scored for this test?
d What percentage of students scored 60 or more for the test?
e What was the interquartile range for this test?
f The top 25% of students scored a mark between ..... and .....
g If you scored 70 for this test, would you be in the top 50% of students in this class?
h Comment on the symmetry of the distribution of marks.

2 A set of data has a lower quartile of 31.5, median of 37 and an upper quartile of 43.5.

a Calculate the interquartile range for this data set.
b Calculate the boundaries that identify outliers.
c Which of the data 22, 13.2, 60, 65 would be outliers?

3 Julie examines a new variety of bean and does a count on the number of beans in 33 pods. Her results were:

5, 8, 10, 4, 2, 12, 6, 5, 7, 7, 5, 5, 5, 13, 9, 3, 4, 4, 7, 8, 9, 5, 5, 4, 3, 6, 6, 6, 6, 9, 8, 7, 6

a Find the median, lower quartile and upper quartile of the data set.

b Find the interquartile range of the data set.

c What are the lower and upper boundaries for outliers?

d Are there any outliers according to c? e Draw a boxplot of the data set.

4 Andrew counts the number of bolts in several boxes and tabulates the data as shown below:

Number of bolts	33	34	35	36	37	38	39	40
Frequency	1	5	7	13	12	8	0	1

a Find the five-number summary for this data set.

b Find the i range ii IQR for the data set.

c Are there any outliers? Test for them. d Construct a boxplot for the data set.

I THE STANDARD DEVIATION

The standard deviation is the most widely used measure of the spread of a sample.

The **standard deviation** measures the **deviation between scores and the mean**, i.e., is a measure of the **dispersal** of the data.

The differences between the scores and the mean are squared, and the average of these squares is then found. The standard deviation is the square root of this average.

The standard deviation gives insight into how the data is **dispersed**.

The larger the standard deviation, the more widely spread the data would be and vice versa. The standard deviation provides a better measure of the spread than either the range or the interquartile range because it considers all scores in the data.

UNGROUPED DATA

The standard deviation, s, can be determined using the formula: $s = \sqrt{\dfrac{\sum(x - \overline{x})^2}{n}}$

where x is any **score**

$\overline{x}$ is the **mean** of the distribution

n is the **total number of scores.**