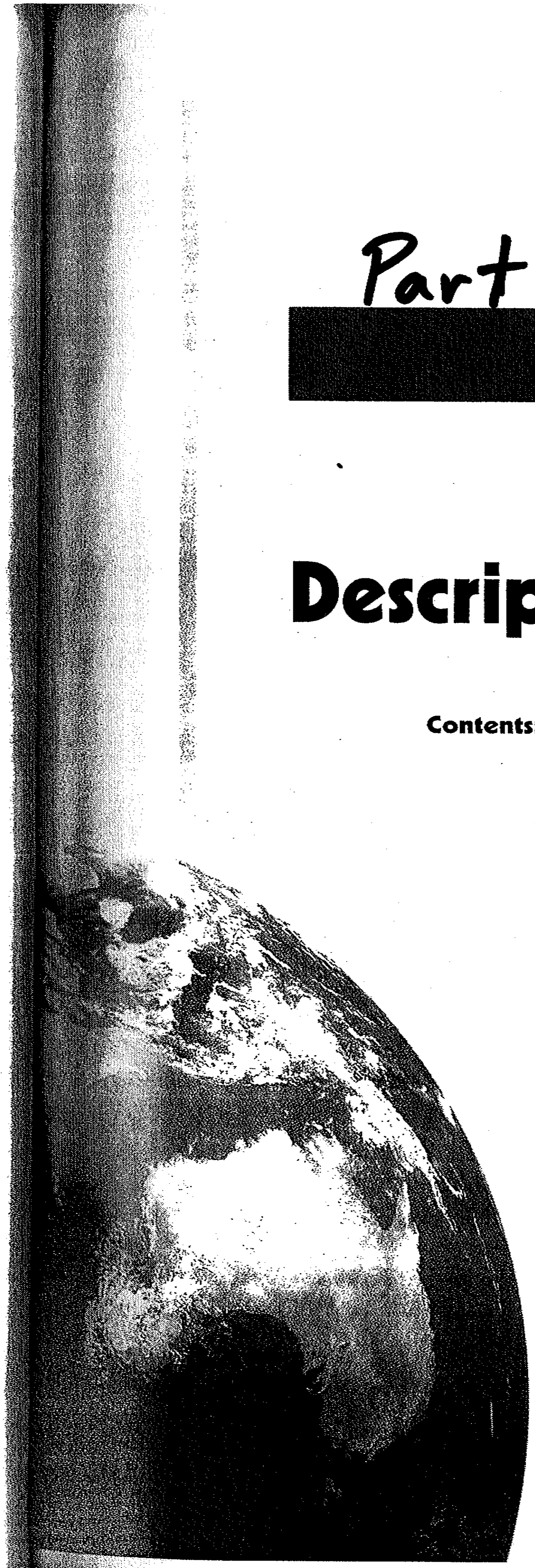


Part F

Chapter 5

Descriptive statistics

Contents:

4 A market research company conducts a survey on the number of times people eat out at restaurants each year. The results from the 100 people surveyed are as follows:

No. of times	*Freq.*
1 - 20	15
21 - 40	21
41 - 60	24
61 - 80	18
81 - 100	12
101 - 120	6
121 - 140	4

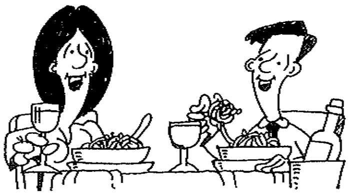
a Construct a frequency table which also shows cumulative frequencies.

b Construct a frequency polygon.

c Construct a cumulative frequency polygon and use it to answer the following:

i What percentage of people ate at restaurants less than 81 times a year?

ii If a person was in the lower 25% of the distribution, what is the maximum number of times they would have eaten at restaurants during the year?



5 The manager of a fast food restaurant is concerned that customers are waiting too long for their food. She decides to gather some statistics on customer waiting times and the following times (in minutes) are recorded:

1.25 2.5 8.5 1.8 4.6 10.5 3.4 7.0 6.25 4.10 5.15 5.95 7.35 5.8 2.9
0.9 3.4 6.0 8.8 2.7 10.2 4.5 5.2 4.1 2.5 7.7 3.8 2.1 5.5 6.25
4.3 1.8 8.0 9.9 3.7 4.4 6.2 3.3 7.2 8.6 3.45 6.55 2.85 9.4 4.25
5.6 11.9 6.4 4.8 5.8

a Group this data into classes of 0 - 1.99, 2 - 3.99, etc. and construct a frequency table which also shows cumulative frequencies.

b Construct a histogram of the data.

c Construct a cumulative frequency graph and answer the following questions:

i How many customers have to wait less than 4 minutes for their food?

ii What percentage of customers have to wait more than 5 minutes for their food?

iii If the restaurant's goal is for 90% of the customers to be given their food within 8 minutes, are they achieving this goal?

F SUMMARISING THE DATA

MEASURES OF THE MIDDLE OF A DISTRIBUTION

After collecting and presenting statistical data, you can now attempt to interpret the data. One way of doing this is to find the value of the **centre or middle** of the distribution.

There are three commonly used measures for the middle of a distribution; the **mean**, the **mode** and the **median**.

However, before proceeding further we will define some of the terms that will be used from now on:

- **Ungrouped data** comprises of single values which have not been put into groups or classes.

  For example, the heights of five children are 1.23 m, 1.56 m, 1.34 m, 1.09 m, 1.71 m.

- **Grouped data** has been grouped together according to the number of times each value occurs (i.e., the frequency). There are two types of grouped data:
  - ► **Grouped discrete data** which can be precisely determined and has been grouped together according to the number of times each value occurs.
  - ► **Grouped continuous data** which cannot be precisely determined and has been grouped together into classes.

MEAN, MODE AND MEDIAN

The **mean** of a set of scores is their arithmetic average obtained by adding all the scores and dividing by the total number of scores.

The **mode(s)** of a set of scores is the score(s) which occurs most frequently.

The **median** of a set of scores is the middle score after they have been placed in order of size from smallest to largest.

A set of scores is **bimodal** if it has two modes. If it has more than two modes we do not use them as a measure of the centre.

MEDIAN FINDER

UNGROUPED DATA

Example 8

Find the mean, mode and median of the following distributions:

a 3, 6, 5, 6, 4, 5, 5, 6, 7 **b** 13, 12, 15, 13, 18, 14, 16, 15, 15, 17.

a mean $= \dfrac{3+6+5+6+4+5+5+6+7}{9}$

$= \dfrac{47}{9}$

$\doteqdot 5.2$

modes are 5 and 6 i.e., is bimodal {both 5 and 6 occur with highest frequency}

median = 5 {In order of size: ~~3, 4, 5, 5~~, 5, ~~6, 6, 6, 7~~ }

↑ middle score

b mean $= \dfrac{13+12+15+13+18+14+16+15+15+17}{10}$

$= \dfrac{148}{10}$

$\doteqdot 14.8$

mode = 15 {occurs most frequently}

median = 15 {In order of size: ~~12, 13, 13, 14~~, 15, 15, ~~15, 16, 17, 18~~}

↑ middle scores

∴ take average, which is 15.

Note: For a sample containing:

- an **odd number** of scores, n say, the **median** score is the $\left(\frac{n+1}{2}\right)$th score
- an **even number** of scores, n say, the **median** score is the average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th scores.

Be sure that you distinguish between the position of a score and its value.

The table below shows the rank score from smallest to largest and the position of each score, for data in **Example 8**, part **b**.

position	1	2	3	4	5	6	7	8	9	10
score	12	13	13	14	15	15	15	16	17	18

i.e., the 4th score is 14

EXERCISE 5F.1

1 Below are the points scored by two basketball teams over a 12 match series:

Team A: 91, 76, 104, 88, 73, 55, 121, 98, 102, 91, 114, 82
Team B: 87, 104, 112, 82, 64, 48, 99, 119, 112, 77, 89, 108

Which team had the higher mean score?

2 Select the mode(s) for the following sets of numbers:

a 44, 42, 42, 49, 47, 44, 48, 47, 49, 41, 45, 40, 49

b 148, 144, 147, 147, 149, 148, 146, 144, 145, 143, 142, 144, 147

c 25, 21, 20, 24, 28, 27, 25, 29, 26, 28, 22, 25

3 Calculate the median value for the following data:

a 21, 23, 24, 25, 29, 31, 34, 37, 41

b 105, 106, 107, 107, 107, 107, 109, 120, 124, 132

c 173, 146, 128, 132, 116, 129, 141, 163, 187, 153, 162, 184

4 A survey of 50 students revealed the following number of siblings per student:

1, 1, 3, 2, 2, 2, 0, 0, 3, 2, 0, 0, 1, 3, 3, 4, 0, 0, 5, 3, 3, 0, 1, 4, 5,
1, 3, 2, 2, 0, 0, 1, 1, 5, 1, 0, 0, 1, 2, 2, 1, 3, 2, 1, 4, 2, 0, 0, 1, 2

a What is the modal number of siblings per student?

b What is the mean number of siblings per student?

c What is the median number of siblings per student?

5 The following table shows the average monthly rainfall for a city.

Month	*J*	*F*	*M*	*A*	*M*	*J*	*J*	*A*	*S*	*O*	*N*	*D*
Av. rainfall (mm)	16	34	38	41	98	172	166	159	106	71	52	21

Calculate the mean average monthly rainfall for this city.

6 The selling prices of the last 10 houses sold in a certain district were as follows:

$146 400, $127 600, $211 000, $192 500,
$256 400, $132 400, $148 000, $129 500,
$131 400, $162 500

a Calculate the mean and median selling prices of these houses and comment on the results.

b Which measure would you use if you were:

i a vendor wanting to sell your house

ii looking to buy a house in the district?

7 Find x if 5, 9, 11, 12, 13, 14, 17 and x have a mean of 12.

8 Towards the end of season, a basketballer had played 14 matches and had an average of 16.5 goals per game. In the final two matches of the season the basketballer threw 21 goals and 24 goals. Find the basketballer's new average.

9 A sample of 12 measurements has a mean of 16.5 and a sample of 15 measurements has a mean of 18.6. Find the mean of all 27 measurements.

10 15 of 31 measurements are below 10 cm and 12 measurements are above 11 cm. Find the median if the other 4 measurements are 10.1 cm, 10.4 cm, 10.7 cm and 10.9 cm.

11 The mean and median of a set of 9 measurements are both 12. If 7 of the measurements are 7, 9, 11, 13, 14, 17 and 19, find the other two measurements.

12 Seven sample values are: 2, 7, 3, 8, 4, a and b where $a < b$. These have a mean of 6 and a median of 5. Find **a** a and b **b** the mode.

MEASURES OF THE CENTRE FROM OTHER SOURCES

When the same data appear several times we often summarise the data in table form. Consider the data of the **given table**:

Data value	*Frequency*	$f \times x$
x = 3	f = 1	$3 \times 1 = 3$
4	1	$4 \times 1 = 4$
5	3	$5 \times 3 = 15$
6	7	$6 \times 7 = 42$
7	15	$7 \times 15 = 105$
8	8	$8 \times 8 = 64$
9	5	$9 \times 5 = 45$
Total	40	278

We can find the measures of the centre directly from the table.

The mode

The mode is 7. There are 15 occurances of this data value which is more than any other data value.

The mean

Adding an $f \times x$ **column** to the table helps to add all scores. For example, there are 15 occurances of the data value 7, these add to $15 \times 7 = 105$.

So, mean $= \dfrac{278}{40} = 6.95$ i.e., mean $= \dfrac{\sum f \times x}{\sum f}$.

$\sum$ is the Greek letter sigma, which we use to represent '*the sum of*'.

The median

There are 40 data values, an even number, so there are *two middle* data values. What are they? How do we find them from the table?

As the sample size $n = 40$, $\dfrac{n+1}{2} = \dfrac{41}{2} = 20.5$

$\therefore$ the median is the average of the 20th and 21st data values.

In the table, the blue numbers show us accumulated values.

Data Value	*Frequency*	*Cumulative frequency*	
3	1	1	one number is 3
4	1	2	two numbers are 4 or less
5	3	5	five numbers are 5 or less
6	7	12	12 numbers are 6 or less
7	15	27	27 numbers are 7 or less
8	8	35	35 numbers are 8 or less
9	5	40	40 numbers are 9 or less
Total	40		

We can see that the 20th and 21st data values (in order) are both 7's,

$\therefore$ median $= \dfrac{7-7}{2} = 7$

Notice that in this example the distribution is clearly skewed even though the mean, median and mode are nearly equal. So, we must be careful in saying that equal values of these measures of the middle enable us to say with certainty that the distribution is symmetric.

GROUPED DISCRETE DATA

The mean, $\overline{x}$, is calculated by:

- multiplying each score (x) by its frequency (f)
- finding the sum of all the values of $f \times x$,
- using the formula, $\overline{x} = \dfrac{\text{total of all } f \times x}{\text{total frequency}} = \dfrac{\sum f \times x}{\sum f}$

Example 9

For the following distribution find:

a the mean

b mode

c median

Score	*Frequency*
1	6
2	9
3	4
4	7

6

8

Score (x)	*Frequency (f)*	$f \times x$	*Cumu. freq.*
1	6	6	6
2	9	18	15
3	4	12	19
4	7	28	26
	26	64	

Note:
6 scores of 1.
15 scores of 1 or 2
$\therefore$ 7th, 8th, ..., 15th are all 2's.

a mean $= \dfrac{\sum f \times x}{\sum f}$

$= \dfrac{64}{26}$

$\doteqdot 2.46$

b mode = 2 {occurs 9 times}

c Since there are 26 scores, there are two 'middle' scores

$$\therefore \text{ median } = \frac{\text{13th score } + \text{ 14th score}}{2}$$

$$= \frac{2+2}{2}$$

$$= 2$$

Example 10

The distribution obtained by counting the contents of 25 match boxes is shown:

Number of matches	*Frequency*
47	2
48	4
49	7
50	8
51	3
53	1

Find the:

a mean

b mode

c median number of matches per box.

Number of matches (x)	*Frequency (f)*	$f \times x$	*Cumulative frequency*
47	2	94	2
48	4	192	6
49	7	343	13
50	8	400	21
51	3	153	24
53	1	53	25
Total	25	1235	-

Note:
6 scores are 47 or 48.
13 scores are 47, 48 or 49.
$\therefore$ 7th, 8th, ..., 13th are all 49s.

a mean $= \dfrac{\sum f \times x}{\sum f}$

$= \dfrac{1235}{25}$

$= 49.4$

b mode = 50

c median is the 13th score

$= 49$

$\{\frac{25+1}{2} = 13,$ i.e., 13th$\}$

EXERCISE 5F.2

1 A hardware store maintains that packets contain 60 nails. To test this, a quality control inspector tested 100 packets and found the following distribution:

Number of nails	*Frequency*
56	8
57	11
58	14
59	18
60	21
61	8
62	12
63	8
Total	100

a Find the mean, mode and median number of nails per packet.

b Comment on these results in relation to the store's claim.

c Which of these three measures is most reliable? Comment on your answer.

2 51 packets of chocolate almonds were opened and their contents counted. The following table gives the distribution of the number of chocolates per packet sampled.

Find the mean, mode and median of the distribution.

Number in packet	*Frequency*
32	6
33	8
34	9
35	13
36	10
37	3
38	2

3 The table alongside compares the mass at birth of some guinea pigs with their mass when they were two weeks old.

a What was the mean birth mass?

b What was the mean mass after two weeks?

c What was the mean increase over the two weeks?

Guinea Pig	*Mass (g) at birth*	*Mass (g) at 2 weeks*
A	75	210
B	70	200
C	80	200
D	70	220
E	74	215
F	60	200
G	55	206
H	83	230

GROUPED CONTINUOUS DATA

When information has been gathered in classes we use the **midpoint** of the class to represent all scores within that interval.

The **midpoint** of a class interval is the average of its endpoints.

For example, the midpoint of $0 - < 50$ is $\dfrac{0+50}{2} = 25$.

We are assuming that the scores within each class are evenly distributed throughout that interval. The mean calculated will therefore be an **approximation** to the true value.

The mode is simply the **modal class** (the class which occurs most frequently).

Example 11

For the following distribution find:

a mean

b mode

Class Interval	*Freq.*
0 - 49.99	12
50 - 99.99	20
100 - 149.99	24
150 - 199.99	23
200 - 249.99	17
250 - 299.99	6

Class interval	*Freq. (f)*	*Midpt. (x)*	$f \times x$
0 - < 50	12	25	300
50 - < 100	20	75	1500
100 - < 150	24	125	3000
150 - < 200	23	175	4025
200 - < 250	17	225	3825
250 - < 300	6	275	1650
Total	102		14 300

Note: midpoint of 150 - < 200 is $\dfrac{150+200}{2}$

i.e., 175

a mean $= \dfrac{\sum f \times x}{\sum f}$

$\doteqdot \dfrac{14300}{102}$

$\doteqdot 140$

b modal class is 100 - < 150

Note: In **Example 11**, the **true class boundary** of, say, the class interval 50-99.99 is 49.995 to 99.995 and the class interval length is 50. The lower limit of this class is therefore 49.995, not 50.

EXERCISE 5F.3

1 Find the approximate mean for each of the following distributions:

a

Score (x)	*Frequency* (f)
1-5	7
6-10	12
11-15	15
16-20	10
21-25	11

b

Score (x)	*Frequency* (f)
40-42	2
43-45	1
46-48	5
49-51	6
52-54	12
55-57	3

2 50 students sit a mathematics test and the results are as follows:

Score	0-9	10-19	20-29	30-39	40-49
Frequency	2	5	7	27	9

Find an estimate of the mean score.

3 Following is a record of the number of goals Chloë has scored in her basketball matches.

15 8 6 10 0 9 2 16 11 23 14 13 17 16 20 12 13
12 10 3 13 5 18 14 19 4 15 15 19 19 14 6 11 29
8 9 3 20 9 25 7 15 19 21 23 12 17 22 14 26

a Find the mean number of goals per match.

b Estimate the mean by grouping the data into:
  i intervals 0-4, 5-9, 10-14, etc.   ii intervals 0-8, 9-16, 17-24, 25-30.

c Comment on your answers from a and b.

4 The table shows the length of newborn babies at a hospital over a one week period.
Find the approximate mean length of the newborn babies.

Length (mm)	*frequency*
400 to 424	2
425 to 449	7
450 to 474	15
475 to 499	31
500 to 524	27
525 to 549	12
550 to 574	4
575 to 599	1

5 The table shows the petrol sales in one day by a number of city service stations.

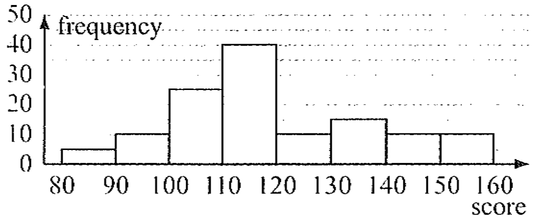
a How many service stations were involved in the survey?

b Estimate the total amount of petrol sold for the day by the service stations.

c Find the approximate mean sales of petrol for the day.

Thousands of litres (l)	*frequency*
2000 to 2999	4
3000 to 3999	4
4000 to 4999	9
5000 to 5999	14
6000 to 6999	23
7000 to 7999	16

6 This histogram illustrates the results of an aptitude test given to a group of people seeking positions in a company.



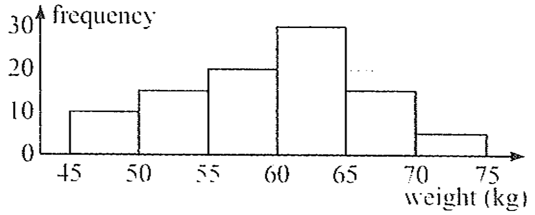
a How many people sat for the test?

b Find an estimate of the mean score for the test.

c What fraction of the people scored less than 100 for the test?

d If the top 20% of the people are offered positions in the company, estimate the minimum mark required.

7



The histogram shows the weights (in kg) of a group of year 10 students at a country high school.

a How many students were involved in the survey?

b Calculate the mean weight of the students.

c How many students weigh less than 56 kg?

d What percentage of students weigh between 50 and 60 kg?

e If a student was selected at random, what would be the chance that the student weighed less than 60 kg?

For grouped continuous data, the **median** can be determined by either of two methods:

- by drawing a cumulative frequency graph and finding the 50th percentile, or
- by examining the cumulative frequency table, then
  - ► finding the interval in which the median lies, (this interval is called the median class)
  - ► using the formula,

$$\textbf{median} = L + \left(\frac{i}{f} \times C\right),$$

where L is the **lower limit of the median class**
i is the **number of scores in the median class needed to arrive at the middle score**
f is the **number of scores in the median class**
C is the **length of the class interval.**

Example 12

Use the formula to find the median of the mathematics test marks:

Test Mark	*Number of students*
30 - 39	6
40 - 49	20
50 - 59	64
60 - 69	87
70 - 79	51
80 - 89	19
90 - 99	10

Test mark	*Lower boundary*	*Number of students*	*Cumulative frequency*
30 - < 39	29.5	6	6
40 - < 49	39.5	20	26
50 - < 59	49.5	64	90
60 - < 69	59.5	87	177
70 - < 79	69.5	51	228
80 - < 89	79.5	19	247
90 - < 99	89.5	10	257
	Total	257	

90 scores $\leqslant 59$
177 scores $\leqslant 69$

Since there are 257 scores, the median is the 129th score $\{\frac{257+1}{2} = 129\}$
$\therefore$ median class is 60 - 69.

Now median $= L + \dfrac{i}{f} \times C$
$= 59.5 + \frac{39}{87} \times 10$
$\doteqdot 64.0$

$L = 59.5$ halfway between 59 and 60
$i = 129 - 90 = 39$
$f = 87$
$C = 10$

PERCENTILES

A **cumulative frequency graph** can be used to find the percentages of the data **below** certain values.

These values are known as **percentiles**.

Percentiles are those values of the data below/above which certain percentages of the frequencies lie.

For instance, the 25th percentile is the value that has 25% of the total frequencies below it.

The 50th percentile is the middle value or **median value** of the data.

The 25th percentile is referred to as the **lower quartile** of the data.

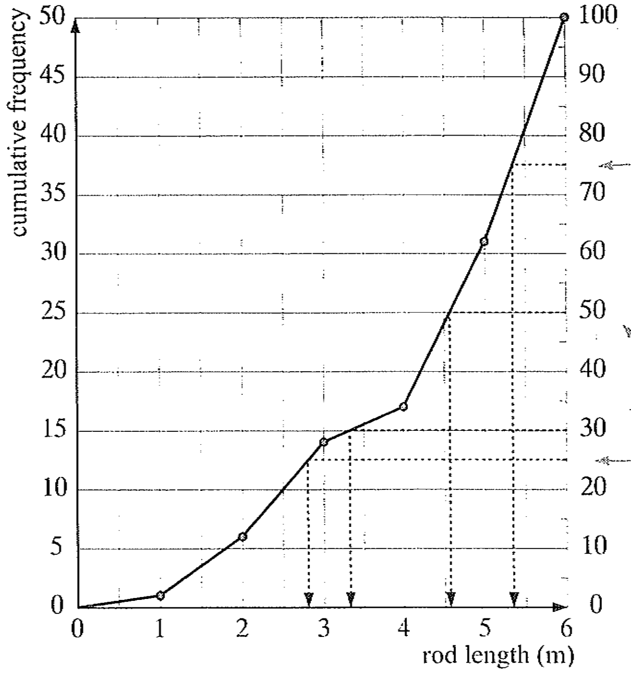
The 75th percentile is referred to as the **upper quartile** of the data.

Example 13

From the cumulative frequency graph in **Example 7**, find:

a the 75th percentile
b the 30th percentile
c the median value
d the 25th percentile.

Cumulative frequency graph of 'Steel rod' data



a Total frequency is 50,
$\therefore$ 75th percentile is the (75% of 50)th score length
$= 37\frac{1}{2}$th score length
$\doteqdot 5.3$

b The 30th percentile is the 15th score $\doteqdot 3.3$

c The median is the 50th percentile, i.e., the 25th score length $\doteqdot 4.6$

d The 25th percentile is the $12\frac{1}{2}$th score length $\doteqdot 2.8$

EXERCISE 5F.4

1 Calculate the median of the following distributions:

a

score	1	2	3	4	5	6
frequency	25	11	8	5	4	1

b

score	5	6	7	8	9	10
frequency	1	3	11	12	8	2

2 This table indicates the number of errors in randomly chosen pages of a telephone directory:

number of errors	0	1	2	3	4	5	6
frequency	67	35	17	8	11	2	1

Find the median number of errors.

3 The following data shows the lengths of 30 trout caught in a lake during a fishing competition. Measurements are to the nearest centimetre.

31	38	34	40	24	33	30	36	38	32	35	32	36	27	35
40	34	37	44	38	36	34	33	31	38	35	36	33	33	28

a Construct a cumulative frequency table for trout lengths, x cm, using the following intervals $24 \leqslant x < 27$, $27 \leqslant x < 30$, ... etc.

b Draw a cumulative frequency graph.

c Use **b** to find the median length.

d Use the original data to find its median and compare your answer with **c**. Comment!

4 In an examination the following scores were achieved by a group of students:

Draw a cumulative frequency graph of the data and use it to find:

a the median examination mark

b how many students scored less than 65 marks

c how many students scored between 50 and 70 marks

d how many students failed, given that the pass mark was 45

e the credit mark, given that the top 16% of students were awarded credits.

Score	*frequency*
$10 \leqslant x < 20$	2
$20 \leqslant x < 30$	5
$30 \leqslant x < 40$	7
$40 \leqslant x < 50$	21
$50 \leqslant x < 60$	36
$60 \leqslant x < 70$	40
$70 \leqslant x < 80$	27
$80 \leqslant x < 90$	9
$90 \leqslant x < 100$	3

5 In a cross-country race, the times (in minutes) of 80 competitors were recorded as follows:

Draw a cumulative frequency graph of the data and use it to find:

a the median time

b the 75th percentile

c the 30th percentile

Score	*frequency*
$20 \leqslant t < 25$	15
$25 \leqslant t < 30$	33
$30 \leqslant t < 35$	21
$35 \leqslant t < 40$	10
$40 \leqslant t < 45$	1

6 The following table gives the age groups of car drivers involved in an accident in a city for a given year.

Draw a cumulative frequency graph of the data and use it to find:

a the median age of the drivers involved in the accidents

b the percentage of drivers, with ages of 23 or less, involved in accidents.

c Estimate the probability that a driver involved in an accident is:

i aged less than or equal to 27 years

ii aged 27 years.

Age (in years)	*No. of accidents*
$16 \leqslant x < 20$	59
$20 \leqslant x < 25$	82
$25 \leqslant x < 30$	43
$30 \leqslant x < 35$	21
$35 \leqslant x < 40$	19
$40 \leqslant x < 50$	11
$50 \leqslant x < 60$	24
$60 \leqslant x < 80$	41