3 DESCRIPTIVE STATISTICS

Objectives

After studying this chapter you should

- understand various techniques for presentation of data;
- be able to use frequency diagrams and scatter diagrams;
- be able to find mean, mode, median, quartiles and standard deviation.

3.0 Introduction

Before looking at all the different techniques it is necessary to consider what the **purpose** of your work is. The data you collected might have been wanted by a researcher wishing to know how healthy teenagers were in different parts of the country. The final result would probably be a written report or perhaps a TV documentary. A straightforward list of all the results could be presented but, particularly if there were a lot of results, this would not be very helpful and would be extremely boring.

The purpose of any statistical analysis is therefore to simplify large amounts of data, find any key facts and present the information in an interesting and easily understandable way. This generally follows three stages:

- sorting and grouping;
- illustration;
- summary statistics.

3.1 Sorting and grouping

The following table shows in the last two columns the average house prices for different regions in the UK in 1988 and 1989.

Clearly prices have increased but has the pattern of differences between areas altered?

	% dwellings owner occupied		Average dwelling price (£)	
	1988 (end)	**1989 (end)**	**1988**	**1989**
United Kingdom	65	67	49 500	54 846
North	58	59	30 200	37 374
Yorks. and Humbs.	64	66	32 700	41 817
East Midlands	69	70	40 500	49 421
East Anglia	68	70	57 300	64 610
South East	68	69	74 000	81 635
South West	72	73	58 500	67 004
West Midlands	66	67	41 700	49 815
North West	67	68	34 000	42 126

(Source: United Kingdom in Figures - Central Statistical Office)

One simple way you could look at the data is to place them all in order, e.g. for 1988 prices:

North	30 200
Yorks & Humbs.	32 700
North West	34 000
East Midlands	40 500
West Midlands	41 700
East Anglia	57 300
South West	58 500
South East	74 000

Even a simple exercise such as this shows clearly the range of values and any natural groups in the data and allows you to make judgements as to a typical house price.

However, with larger quantities of data, putting into order is both tedious and not very helpful. The most commonly used method of sorting large quantities of data is a **frequency** table. With qualitative or discrete quantitative data this is simply a record of how many of each type were present. The following frequency table shows the frequency with which **other types of vehicles** were involved in cycling accidents:

	Number	%
Motor Cycle	96	2.5
Motor Car	2039	52.3
Van	168	4.3
Goods Vehicle	126	3.2
Coach	49	1.3
Pedestrian	226	5.8
Dog	120	3.1
Cyclist	218	5.6
None - defective road surface	266	6.8
None - weather conditions	129	3.3
None - mechanical failure	65	1.7
Other	399	10.2
Total	3901	

(Source: Cycling Accidents - Cyclists' Touring Club)

Note: rounding errors mean that the total % is 100.1

With continuous data and with discrete data covering a wide range it is more useful to put the data into groups. For example, take the share prices in the information in the last chapter (see p32). This could be recorded as shown below:

Share Price (p)	Frequency
1 - 200	....
201 - 400	....
401 - 600	....
601 - 800	....
801 - 1000	....
1001 ormore	....
Total	

Note the following points:

- Group limits do not overlap and are given to the same degree of accuracy as the data is recorded.
- Whilst there is no absolute rule, neither too many nor too few groups should be used. A good rule is to look at the range of values, taking care with extremes, and divide into about six groups.
- If uneven group sizes are used this can cause problems later on. The only usual exception is that 'open ended' groups are often used at the ends of the range.

- The class boundaries are the absolute extreme values that could be rounded into that group, e.g. the upper class boundary of the first group is 200.5 (really 200.4999.....).

Stem and leaf diagrams

A new form of frequency table has become widely used in recent years. The **stem and leaf** diagram has all the advantages of a frequency table yet still records the values to full accuracy.

As an example, consider the following data which give the marks gained by 15 pupils in a Biology test (out of a total of 50 marks):

27, 36, 24, 17, 35, 18, 23, 25, 34, 25, 41, 18, 22, 24, 42

The stem and leaf diagram is determined by first recording the marks with the 'tens' as the **stem** and the 'units' as the **leaf**.

This is shown opposite.

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

The leaf part is then reordered to give a final diagram as shown. This gives, at a glance, both an impression of the spread of these numbers and an indication of the average.

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

Example

Form a stem and leaf diagram for the following data:

21, 7, 9, 22, 17, 15, 31, 5, 17, 22, 19, 18, 23,

10, 17, 18, 21, 5, 9, 16, 22, 17, 19, 21, 20.

Solution

As before, you form a stem and leaf, recording the numbers in the leaf to give the diagram opposite.

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

Exercise 3A

1. For each of the measurements you made at the start of Chapter 2 compile a suitable frequency table, or if appropriate a stem and leaf diagram.

2. The table below shows details of the size of training schemes and the number of places on the schemes. Notice that the table has used uneven group sizes. Can you suggest why this has been done?

Size of Training Schemes

Number of approved places	Number of schemes	Percentage of all schemes
1– 20	2167	51.4
21– 50	855	20.3
51– 100	581	13.8
101– 500	560	13.3
501– 1000	41	1.0
over 1000	14	0.3
	4218	

(Source: August 1985 Employment Gazette)

3. The table below shows the ages of registered drug addicts in the period 1971 -1976. What conclusions can you draw from this about the relative ages of drug users during this period?

Dangerous drugs: registered addicts United Kingdom

	1971	1972	1973	1974	1975	1976
Males	1133	1194	1369	1459	1438	1389
Females	416	421	446	512	515	492
Age distribution:						
Under 20 years	118	96	84	64	39	18
20 and under 25	772	727	750	692	562	411
25 and under 30	288	376	530	684	754	810
30 and under 35	112	117	134	163	219	247
35 and under 50	112	118	136	163	169	189
50 and over	177	165	180	197	193	188
Age not stated	20	16	1	8	17	18

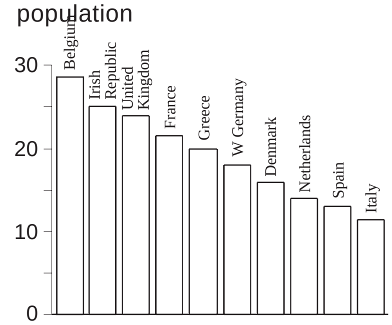
3.2 Illustrating data - bar charts

In the last question of the previous exercise you would have to look at the different figures and make size comparisons to interpret the data; e.g. in 1976 there were twice as many in the 25-30 age group as were in the 20-25 age group. Using diagrams can often show the facts far more clearly and bring out many important points.

The most commonly used diagrams are the various forms of **bar chart**. A true bar chart is strictly speaking only used with qualitative data, as shown opposite.

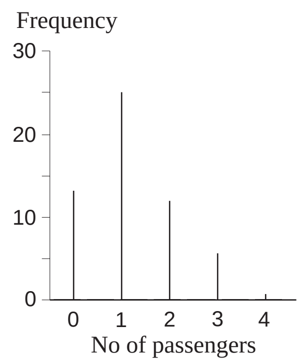
Note that there is no scale on the horizontal axis and gaps are left between bars.

Child pedestrians killed in Europe: deaths per million population



With quantitative discrete data a frequency diagram is commonly used. In a school survey on the number of passengers in cars driving into Norwich in the rush hour the following results were obtained.

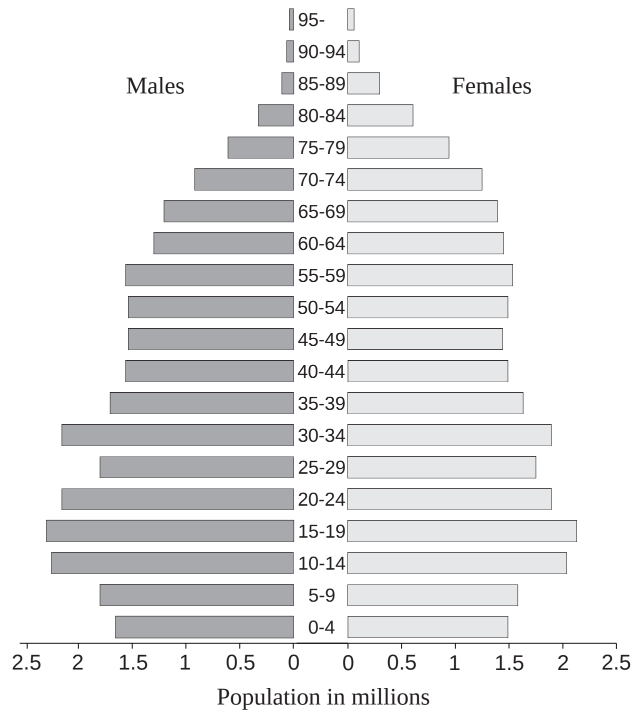
No. of passengers	Frequency
0	13
1	25
2	12
3	6
4	1



Strips are used rather than bars to emphasise discreteness. In practice, however, many people use a bar as this can be made more decorative. It is again usual to keep the bars separate to indicate that the scale is not continuous.

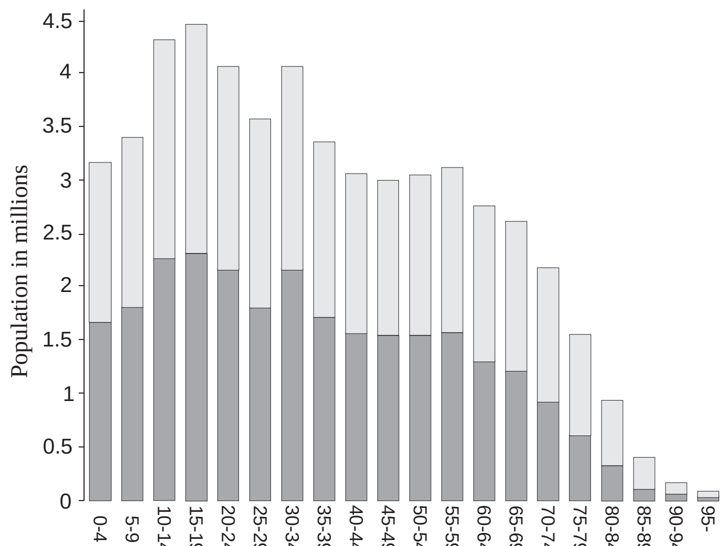
Composite bar charts

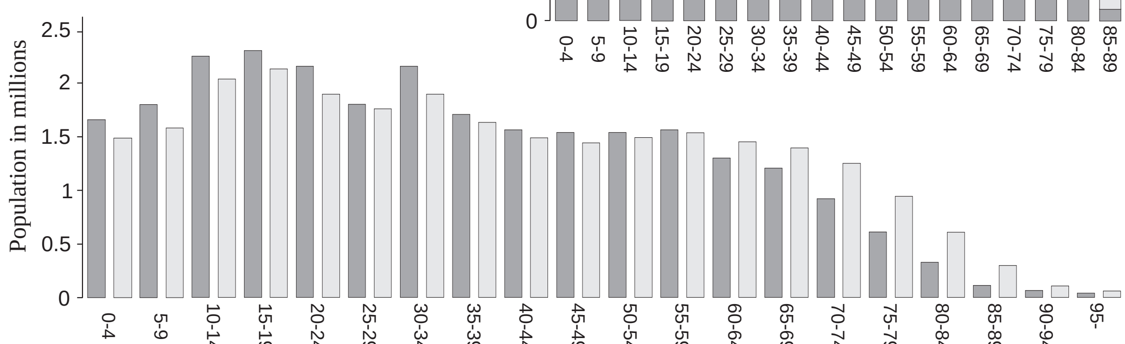
Composite bar charts are often used to show sets of comparable information side by side, as shown opposite.



Age group distribution, Great Britain, 1981

There are alternative ways this could have been shown, as illustrated opposite and below.



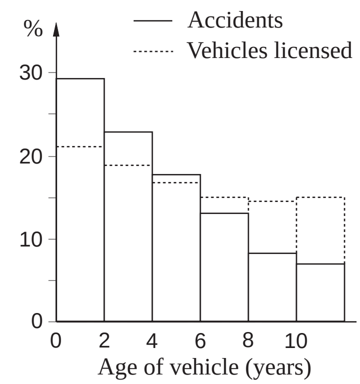


Activity 1 Interpreting the graph

Working in groups, consider these questions about the previous composite bar charts.

What are the main differences between the age distributions of men and women? Can you explain why there are more people in their 50's than 40's? What are the main advantages and disadvantages of each of the different methods of presenting the data?

Histograms

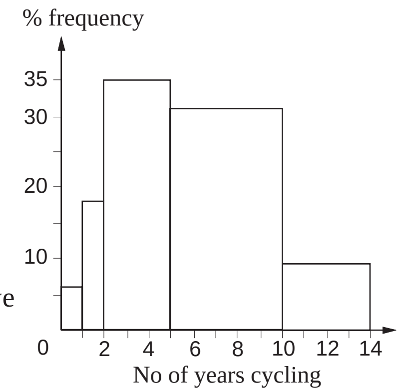


A **histogram** is generally used to describe a bar chart used with continuous data.

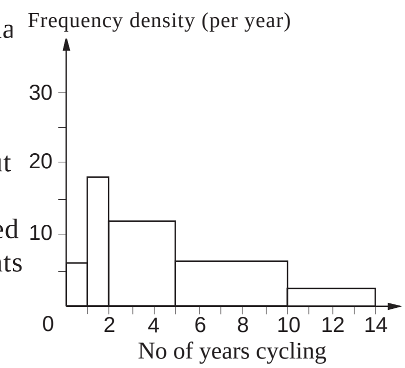
Note that the horizontal axis is a proper numerical scale and that no gaps are drawn between bars. Bars are technically speaking drawn up to the class boundaries though in practice this can be hard to show on a graph. Care must be taken however if there are uneven group sizes. For example the following table shows the percentages of cyclists divided into different age groups and sexes.

Number of years cycling	Age 0-16	Age 16-25	Age 25+	Sex Male	Sex Female
0-1	6%	4%	1%	2%	3%
1-2	18%	8%	3%	4%	8%
2-5	35%	25%	10%	12%	21%
5-10	31%	29%	9%	13%	15%
10-14	9%	33%	77%	69%	52%

(Source: Cycling Accidents - Cyclist's Touring Club.)



If you use the pure frequency values from the table to draw a histogram showing the percentages of children aged 0-16 who have been cycling for different numbers of years, you get the diagram opposite. This, though, is incorrect .



The fact that the groups are of different widths makes it appear tha children are more likely to have been cycling for longer periods. This is because our eyes look at the proportion of the **areas**. To overcome this you need to consider a standard unit, in this case a year. The first two percentage frequencies would be the same, but the next would be 35/3 = 11.7% as it covers a three year period. This is called the **frequency density**; that is, the frequency divided by the class width. Similarly, dividing by 5 and 4 gives the heights for the remaining groups. The correct histogram is shown opposite.

Note the labelling of the vertical scale.

Example

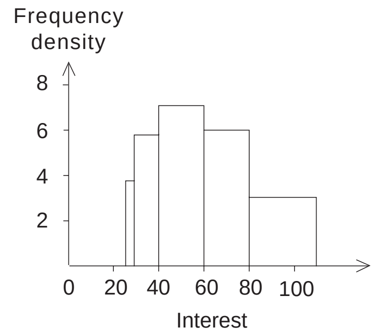
The table shows the distribution of interest paid to investors in a particular year.

Interest (£)	25-	30-	40-	60-	80-	110-
Frequency	18	55	140	124	96	0

Draw a histogram to illustrate the data.

Solution

Interest	Class widths	Frequency	Frequency density
25-	5	18	3.6
30-	10	55	5.5
40-	20	140	7.0
60-	20	124	6.2
80-	30	96	3.2

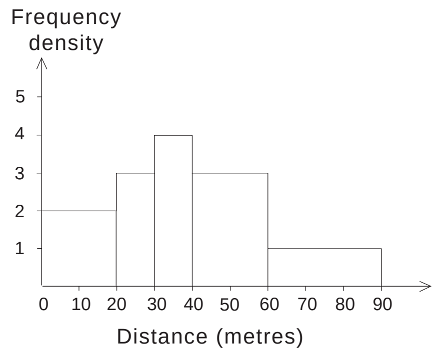


Example

The histogram opposite shows the distribution of distances in a throwing competition.

(a) How many competitors threw less than 40 metres?

(b) How many competitors were there in the competition?



Solution

Using the formula

$$\text{class width} \times \text{frequency density} = \text{frequency}$$

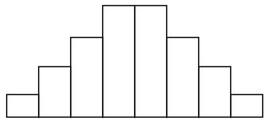
gives the following table.

Interval	Class width	Frequency density	Actual frequency
0-20	20	2	$2 \times 20 = 40$
20-30	10	3	$3 \times 10 = 30$
30-40	10	4	$4 \times 10 = 40$
40-60	20	3	$3 \times 20 = 60$
60-90	30	1	$1 \times 30 = 30$

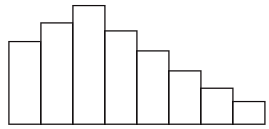
(a) $40 + 30 + 40 = 110$

(b) $40 + 30 + 40 + 60 + 30 = 200$

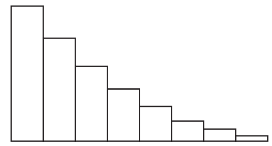
There are a number of common shapes which appear in histograms and these are given names:



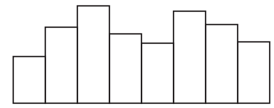
Symmetrical or *Bell Shaped*
e.g. exam results



Positively (or right) *Skewed*
e.g. earnings of people in the UK



Reverse J Shaped
e.g. lifetimes of light bulbs



Bimodal (i.e. twin peaks)
e.g. heights of 14 yr old boys and girls

When a histogram is drawn with continuous data it appears that there are shifts in frequency at each class boundary. This is clearly not true and to show this you can often draw a line joining the middles of the tops of the bars, either as a series of straight lines to form a **frequency polygon**, or more realistically with a curve to form a **frequency curve**. These also show the shape of the distribution more clearly.

Exercise 3B

1. Draw appropriate bar charts for the data you collected at the start of Chapter 2.
2. Use the information on the ages of sentenced prisoners in the table opposite to draw a composite bar chart. Ignore the uneven group sizes.

   Explain why you have used the particular type of diagram you have.

Age and sex of prisoners, England and Wales 1981

Age	Men	Women
14-16	1637	129
17-20	9268	238
21-24	7255	235
25-29	5847	188
30-39	7093	236
40-49	3059	132
50-59	1128	35
60 and over	262	7

3. The information below and opposite relates to people taking out mortgages. Draw an appropriate bar chart for the All buyers information in each case.

By type of dwelling (%)

Type	All buyers
Bungalow	10
Detached house	19
Semi-detached house	31
Terraced house	31
Purpose built flat	7
Converted flat	3

By age of borrowers (%)

Age	All buyers
Under 25	22
25-29	26
30-34	21
35-44	20
45-54	8
55 & over	3

By mortgage amounts(%)

Amount	All buyers
Under £8000	16
£ 8000 - £ 9999	10
£10000 - £11999	16
£12000 - £13999	17
£14000 - £15999	17
£16000 & over	24

4. 100 people were asked to record how many television programmes they watched in a week. The results are shown opposite.

Draw a histogram to illustrate the data.

No. of programmes	0-	10-	18-	30-	35-	45-	50-	60-
No. of viewers	3	16	36	21	12	9	3	0

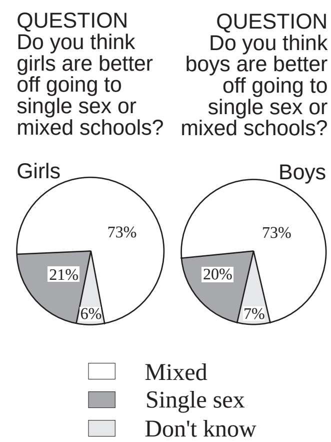
5. 68 smokers were asked to record their consumption of cigarettes each day for several weeks. The table shown opposite is based on the information obtained.

Illustrate these data by means of a histogram.

Average no. of cigarettes smoked per day	0-	8-	12-	16-	24-	28-	34-50
No. of smokers	4	6	12	28	8	6	4

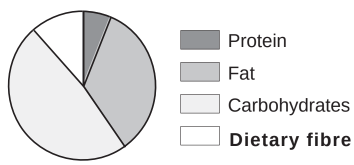
3.3 Illustrating data - pie charts

Another commonly used form of diagram is the **pie chart**. This is particularly useful in showing how a total amount is divided into constituent parts. An example is shown opposite.



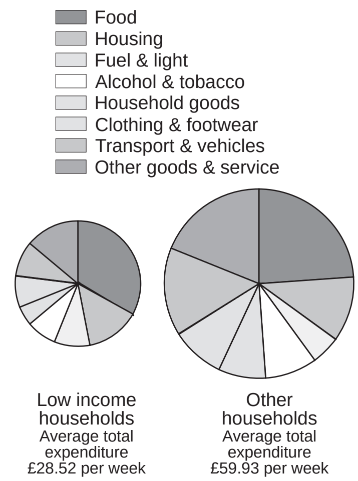
To construct a pie chart it is usually easiest to calculate percentage frequencies. Look at the contents list for the packet of 'healthy' crisps:

Nutrient	Per 100 g
Protein	6.1 g
Fat	34.2 g
Carbohydrates	48.1 g
Dietary Fibre	11.6 g



There are now percentage pie chart scales which can be used to draw the charts directly. Using a traditional protractor method you need to find 6.1% of 360° etc. This gives the pie chart shown above.

When two sets of information with different totals need to be shown, the comparative pie charts are made with sizes proportional to the totals. However, as was discussed with histograms, it is the relative area that the mind uses to make comparisons. The radii therefore have to be in proportion to the **square root** of the total proportion. For example, in the graph opposite the pie charts are drawn in proportion to the 'average total expenditure' i.e. 59.93/28.52 = 2.10.



The radii are therefore in the proportion $\sqrt{2.10} \approx 1.45$. Smaller radius $= 1.7$ cm, then the larger radius $= 1.7 \times 1.45 = 2.5$ cm.

In general, when the total data in the two cases to be illustrated are given by A_1 and A_2, then the formula for the corresponding radii is given by

$$\frac{A_1}{A_2} = \frac{\pi r_1^2}{\pi r_2^2} = \left(\frac{r_1}{r_2}\right)^2$$

Alternatively,

$$\frac{r_1}{r_2} = \sqrt{\frac{A_1}{A_2}}$$

Exercise 3C

1. Draw pie charts for hair colour and eye colour from the results of your survey at the start of Chapter 2.

2. During the 1983 General Elections the % votes gained by each party and the actual number of seats gained by each party are shown opposite.

   (a) Draw separate pie charts, using the same radius, for votes and seats won.

   (b) Calculate the number of seats that would have been gained if seats were allocated in proportion to the % votes gained. Show this and the actual seats gained on a composite bar chart.

   (c) Show how this information could be used to argue the case in favour of proportional representation.

	Conservative	Labour
% Votes	43.5	28.3
Seats won	397	209

	Liberal/Democrats	Other
% Votes	26.0	2.2
Seats won	23	21

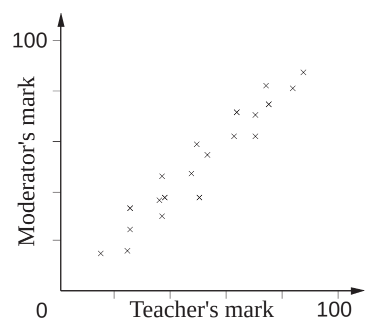
3. According to a report showing the differences in diet between the richest and poorest in the UK the figures opposite were given for the consumption of staple foods (ounces per person per week).

   Draw comparative pie charts for this information. What differences in dietary pattern does this information show?

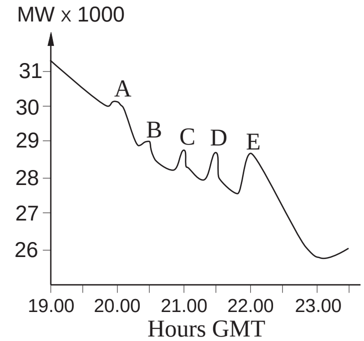
	Poorest 10%	Richest 10%
White bread	26	12.3
Sugar	11.5	8
Potatoes	48.3	33.4
Fruit	13	25.3
Vegetables	21.5	30.7
Brown bread	5.2	8

3.4 Illustrating data – line graphs and scattergrams

Where there is a need to relate one variable to another a different form of diagram is required. When a link between two different quantities is being examined a **scattergram** is used. Each pair of values is shown as a point on a graph, as shown opposite.

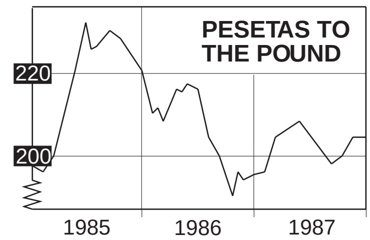


In other cases where the scale on the *x*-axis shows a systematic change in a particular time period, a line graph can be used as shown in the graph opposite.



The effect of a popular television programme on electricity demand is shown in this curve, which shows typical demand peaks. Peaks A and E coincide with the start and finish of the programme; peaks B, C and D coincide with commercial breaks.

Care needs to be taken over vertical scales. In the graph opposite it appears that the value of the peseta has varied dramatically in relation to the pound. However, looking at the scale shows that this has at most varied by 20 pesetas $(\pm 5\%)$. To start the scale at 0 would clearly be unreasonable so it is usual to use a zig-zag line at the base of a scale to show that part of the scale has been left out.



Exercise 3D

1. By drawing scattergrams of your data from Activity 1 at the start of Chapter 2 examine the following statements:
   (a) Taller people tend to have faster pulses.
   (b) People with faster pulses tend to have quicker reaction times.
   (c) High blood pressure is more common in heavier people.

2. The next page shows details of statistics published by Devon County Council on road accidents in 1991. Use this information to write a newspaper report on accidents in the county that year. Include in your report any of the tables and diagrams shown or any of your own which you think would be suitable in an article aimed at the general public.

3.5 Using computer software

There are many packages available on the market which are able to do all or most of the work covered here. These fall into two main categories:

(a) Specific statistical software where a program handles a particular technique and data are fed in directly.

(b) Spreadsheet packages, where data are stored in a matrix of rows and columns; a series of instructions can then carry out any technique which the particular package is able to do.

In the commercial/research world very little work is now carried out by hand; the large quantities of data would make this very difficult.

Activity 2

If you have access to a computer, find out what software you have available and use this to produce tables and diagrams for the data you have collected.

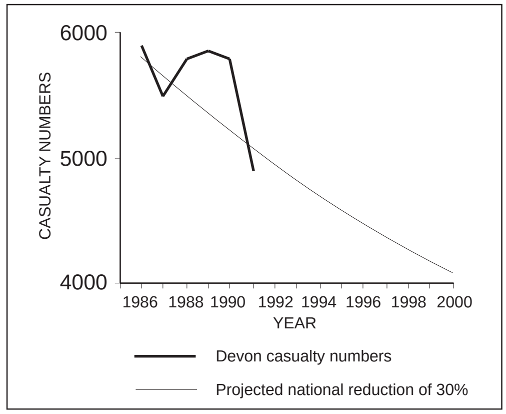
How many?

Reported injury accidents have decreased by 11% compared with last year. Traffic flows also show a small decrease in numbers in urban areas.

Accidents by year and severity

Year	Fatal	Serious	Slight	Total injury accidents
82	91	1 521	2 680	4 292
83	87	1 453	2 808	4 348
84	78	1 486	2 868	4 432
85	65	1 432	3 003	4 500
86	78	1 424	2 950	4 452
87	81	1 243	2 891	4 215
88	74	1 188	3 056	4 318
89	80	1 120	3 199	4 399
90	67	1 048	3 124	4 239
91	76	866	2 814	3 756

Target reduction



The government has set a target of 30% reduction in casualties by the year 2000 using a base of an average figure for 1981 - 1985.

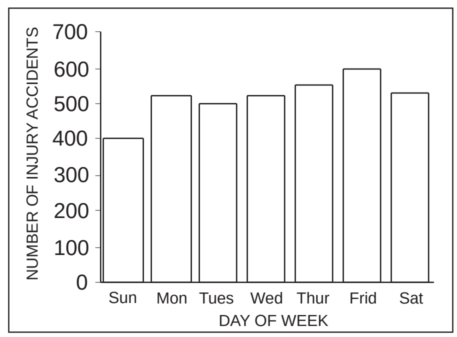
Who?

This table shows the number of people killed and injured in 1991.

Casualties by road user type

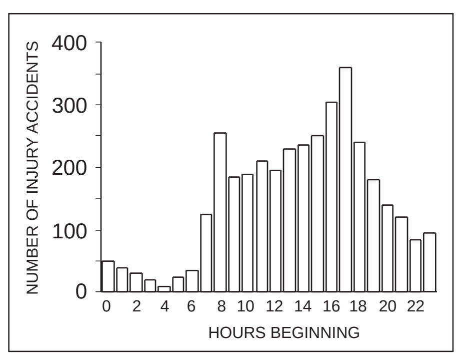
	1991 Fatal	Serious	Slight	Total
Pedestrians	21	216	497	734
Pedal Cyclists	2	69	257	328
Motorcycle Riders	21	234	431	686
Motorcycle Passengers	0	14	50	64
Car Drivers	20	265	1387	1672
Front Seat Car Passengers	7	110	590	707
Rear Seat Car Passengers	6	61	325	392
Public Service Vehicle Passengers	0	4	67	71
Other Drivers	4	26	117	147
Other Passengers	2	14	43	59
Totals	83	1013	3764	4860

Injury accidents by day of week 1991



Accident levels are highest towards the end of the week. This reflects the increased traffic on those days during holiday periods as well as weekend 'evenings out' throughout the year.

Injury accidents by time of day 1991



Accidents plotted by hours of day clearly shows the peaks during the rush hours particularly in the evening. Traffic flows decrease during the rest of the evening but the accident levels remain high.

Accidents involving children

The table shows the number of children killed and injured in Devon for the years 1989 - 1991.

	Age group (years) 0 - 4			5 - 9			10 - 15			Total 0 - 15		
	89	90	91	89	90	91	89	90	91	89	90	91
Pedestrians	41	48	49	96	105	89	139	125	112	276	278	250
Pedal cycles	1	1	2	25	20	27	134	115	105	160	136	134
Car passengers	38	71	38	72	54	49	107	93	88	217	218	175
Others	2	12	4	4	16	5	68	46	18	74	74	27
Totals	82	132	93	197	195	170	448	379	323	727	706	586

3.6 What is typical?

At the beginning of Chapter 2 a question was posed concerning the normal blood pressure for someone of your age. If you did this experiment you will perhaps have a better idea about what kind of value it is likely to be. Another question you might ask is 'Are women's blood pressures higher or lower than men's?'

If you just took the blood pressure of one man and one woman this would be a very poor comparison. What you need, therefore, is a single representative value which can be used to make such comparisons.

Activity 3

Obtain about 30 albums of popular music where the playing time of each track is given. Write down the times in decimal form (most calculators have a button which converts minutes and seconds to decimal form) and the total time of the album. Also write down the number of tracks on the album.

There are two questions that could be asked:

(a) What is a typical track/album length?

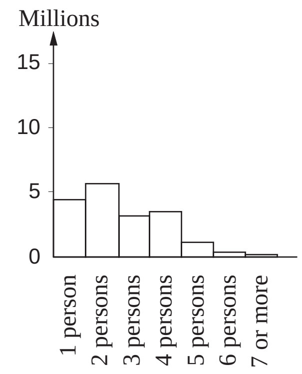
(b) What is a typical number of tracks on an album?

Using the mode and median

The easiest measure of the average that could be given is the **mode**. This is defined as the item of data with the **highest frequency**.

Activity 4 Census data

An extract from the 1981 census is shown opposite. What does it show?



SIZE OF HOUSEHOLDS

The most common size of household in 1981 was two people. There were just under 20 million households in total.
In 4.3% of households in Great Britain there was more than one person per room compared with 7.2% in 1971.

When data are grouped you have to give the **modal group**. In the following example the modal group is 1500 cc - 1750 cc.

Engine size : Private cars involved in accidents

-1000 cc	7.7%
-1250 cc	13.9%
-1500 cc	25.4%
-1750 cc	27.2%
-2000 cc	12.6%
-2500 cc	9.3%
Over 2500 cc	3.9%

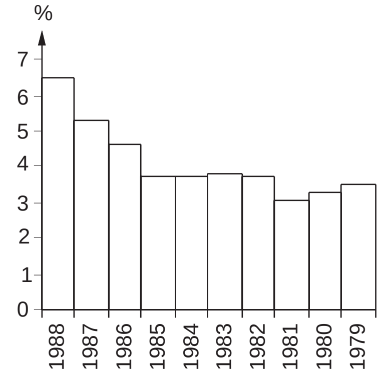
(Source - Analysis of accidents - Assn. of British Insurers)

There are, however, problems with using the mode:

(a) The mode may be at one extreme of the data and not be typical of all the data. It would be wrong to say from the data opposite that accidents were typically caused by people who had passed their test in the last year.

(b) There may be no mode or more than one mode (bimodal).

(c) Some people use a method with grouped data to find the mode more precisely within a group. However, the way in which data were grouped can affect in which group the mode lies.



Distribution of accidents in 1989 by year in which driving test was passed.

The mode has some practical uses, particularly with discrete data (e.g. tracks on an album) and you can even use the mode with qualitative data. For example, a manufacturer of dresses wishing to try out a new design in one size only would most likely choose the modal size.

The **median** aims to avoid some of the problems of the mode. It is the value of the **middle item** of data when they are all placed in order. For example, to find the median of a group of seven people's weights in kg: 75.3, 82.1, 64.8, 76.3, 81.8, 90.1, 74.2, you first put them in order and then identify the middle one.

64.8, 74.2, 75.3, 76.3, 81.8, 82.1, 90.1,
↑
median

Example

Find the median mark for the following exam results (out of 20). Compare this to the mode.

2, 3, 7, 8, 8, 8, 9, 10, 10, 11, 12, 12, 14, 14, 16, 17, 17, 19, 19, 20

Solution

There are 20 items of data, so the median is the $\frac{21}{2} = 10\frac{1}{2}$th item;

i.e. you take the average of the 10th and 11th items, giving

$$\text{median} = \frac{11+12}{2} = \frac{23}{2} = 11.5.$$

The mode is 8, since there are three results with this value.

For these data, the median gives a more representative mark than does the mode.

In general, if there are n items of data, the median is the

$$\frac{(n+1)}{2}\text{th item.}$$

Where there is an even number of data the median will be in between two actual values of data, and so the two values are averaged.

Exercise 3E

1. Find the median length of track time for each of your albums.
2. The data opposite show the cost of various medical insurance schemes for people living in London or provincial areas. Find the median cost of insurance for a single person aged 25 in

   (i) London (ii) Provincial areas.

   What is the approximate extra paid by a person living in London?

Company	Maximum benefits yearly per person	Yearly premium for single person (age 25)	
		London rates	Provincial rates
	£	£	£
AMA	40 000	222	153
BCWA	No limit	190	139
BUPA	No limit	316	205
Crown Life	45 000	258	172
Crusader	No limit	279	195
EHAS	No limit	292	236
Health First	No limit	255	166
Holdcare	No limit	180	134
Orion	50 000	182	182
PPP	No limit	288	156
WPA	45 000	271	188

3.7 Grouped data

With grouped data a little more work is required. An example concerning yearly cycling in miles is shown opposite.

The median is the

$$\frac{(8552+1)}{2} = 4276.5\text{th item.}$$

There are two commonly used methods for finding this:

Miles cycled in 1980

Miles	Number	%
0-500	1252	15
500-1000	1428	17
1000-1500	1231	14
1500-2000	1016	12
2000+	3625	42
TOTAL	8552	100

(a) **Linear interpolation**. This assumes an even spread of data within each group.

By adding up the frequencies:

$$1252+1428+1231=3911$$

but $3911+1016=4927$

You can deduce that the 4276.5 th piece of data is therefore in the 1500–2000 group and in the bottom half.

More precisely this is $4276.5-3911=365.5$ items along that group. Since there are 1016 item in this group you need to go 365.5/1016 = 0.36 of the way up this group.

This will be

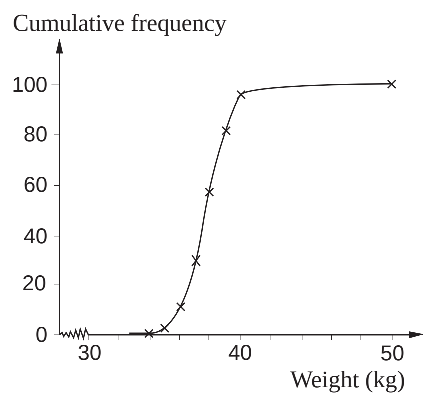
$$1500+(0.360\times 500)=1680.$$

It should be remembered this is only an approximate result and should not be given to excessive accuracy.

(b) **Cumulative frequency curves**. This is a graphical method and therefore of limited accuracy, but assumes a more realistic nonlinear spread in each group. Other information apart from the median can also be obtained from them.

The cumulative frequencies are the frequencies that lie below the upper class boundaries of that group. For example in a large survey on people's weights in kg the following results were obtained:

Weight (kg)	Frequency	Cumulative frequency
< 33.0	1	1
33.0 - 33.9	0	1
34.0 - 34.9	2	3
35.0 - 35.9	8	11
36.0 - 36.9	19	30
37.0 - 37.9	27	57
38.0 - 38.9	25	82
39.0 - 39.9	14	96
40.0 - 49.9	3	99
≥ 50.0	1	100



For example, the cumulative frequency 30 tells you that 30 people weighed less than 36.95 kg. These are then plotted using the **upper class boundaries** (U.C.B.) on the *x*-axis.

The median is at the 50.5th item and can be read from the graph. The graph can also be used to answer such questions as, 'How many people weighed 38.5 kg or less?

Note the 'S' shape of the graph, which will occur when the distribution is bell shaped.

Activity 5

Use the cumulative frequency graph on page 63 to estimate

(a) the percentage of people with weight

(i) less than 38.5 kg,

(ii) greater than 37.5 kg;

(b) the weight which is exceeded by 75% of people.

Exercise 3F

1. Draw up a frequency table of the track times for all the albums in the survey conducted in Activity 3. Draw a cumulative frequency curve of the results and use this to estimate the median playing time.

2. The data below show the monthly rainfall at various weather stations in Norfolk one September. Compile a frequency table and draw a cumulative frequency curve to find the median monthly rainfall.

Acle	91.6	Dunton	67.6	Lingwood	79.2	U.Sheringham	71.4
Ashi	80.8	Edgefield	H108.4	Loddon	74.0	Shotesham	82.0
Ayylebridge	74.8	Fakenham	84.3	Lyng	74.8	Shropham	85.6
Aylsham	91.4	Felmingham	85.9	Marham R.A.F.	59.5	Snettisham	82.3
Barney	82.5	Feltwell	71.6	Morley	78.7	Snoring Little	79.0
Barton	84.7	Foulsham	78.76	Mousehold	74.8	Spixworth	72.0
Bawdeswell	73.2	Framingham C	69.6	Norton Subcourse	69.3	Starston	78.5
Beccles	73.7	Fritton	82.0	Norwich Cemetery	84.8	S.Strawless	77.2
Besthorpe	73.5	Great Fransham	75.5	Nch.G Borrow Road	85.3	Swaffham	87.9
Blakeney	76.1	Gooderstone	75.1	Ormesby	94.7	Syderstone	88.2
Braconash	57.9	Gressehall	71.4	Paston School	81.9	Taverham	83.4
Bradenham	58.4	Heigham WW	87.7	Pulham	68.5	North Thorpe	78.6
Briston	91.5	Hempnall	66.9	Raveningham	44.7	Thurgarton	70.0
Brundall	68.6	Hempstead Holt	105.5	E.Raynham	70.5	Tuddenham E	79.8
Burgh Castle	76.9	Heydon	76.2	S.Raynham	78.1	Tuddenham N	81.5
Burnham Market	63.0	Hickling	63.2	Rougham	72.9	Wacton	61.6
Burnham Thorpe	L42.2	Hindringham	65.8	North Runeton	61.7	North Walsham	75.2
Buxton	85.3	Holme	69.3	Saham Toney	84.3	West Winch	65.9
Carbrooke	93.1	Hopton	84.9	Salle	75.0	Gt. Witchingham	74.7
Clenchwarton	56.0	Horning	87.7	Sandringham	76.5	Wiveton	78.2
Coltishall R.A.F.	87.0	Houghton St. Giles	89.2	Santon Downham	89.4	Wolferton	59.0
Costessey	74.6	Ingham	75.2	Scole	71.3	Wolterton Hall	89.8
North Creake	80.2	High Kelling	93.5	Sedgeford	65.8	Woodrising	82.9
Dereham	85.8	Kerdiston	73.2	Shelfanger	76.6	Wymondham	68.2
Ditchingham	67.6	King's Lynn	63.5	L.Sheringham	72.8	Taverh'm 46-yr av.	53.6
Downham Market	59.7	Kirstead	79.2				

H - highest, L - lowest

(Source : Eastern Daily Press)

3. The distribution of ordinary shares for Cable & Wireless PLC in 1987 is shown opposite. Find the median amount of shares using interpolation. Comment critically on the use of the median as a typical value in this case.

The distribution of ordinary shares at 31 March, 1987	Number of holdings
1 - 250	50 268
251 - 500	69 443
501 - 1 000	25 705
1 001 - 10 000	32 730
10 001 - 100 000	2 086
100 001 - 999 999	669
1 000 000 and over	166
	181 067

(Source: Cable & Wireless PLC - Report 1987)

3.8 Interpreting the mean

One criticism of the median is that it does not look at **all** the data. For example a pupil's marks out of 10 for homework might be:

3, 4, 4, 4, 9, 10, 10.

The pupil might think it unfair that the median mark of 4 be quoted as **typical** of his work in view of the high marks obtained on three occasions.

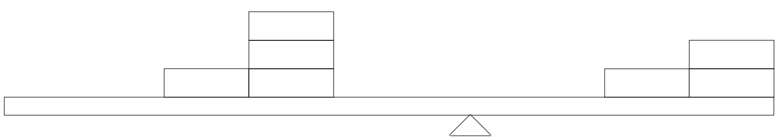
The **mean** though is a measure which takes account of every item of data. In the example above the pupil has clearly been inconsistent in his work. If he had been consistent in his work what mark would he have had to obtain each time to achieve the same total mark for all seven pieces?

$$\text{Total mark} = 3+4+4+4+9+10+10 = 44$$

$$\text{Consistent mark} = \frac{44}{7} \approx 6.3$$

This is in fact the **arithmetic mean** of his marks and is what most people would describe as the **average mark**.

But what does the **mean** actually mean? The mean is the most commonly used of all the 'typical' values but often the least understood. The mean can be basically thought of as a balancing device. Imagine that weights were placed on a 10 cm bar in the places of the marks above. In order to balance the data the pivot would have to be placed at 6.3



This is both the strength and weakness of the mean; whilst it uses all the data and takes into account end values it can easily be distorted by extreme values. For example, if in a small company the boss earns £30 000 per annum and his six workers £5000, then

$$\text{mean} = \frac{1}{7}(30\,000+5000+5000+5000+5000+5000+5000)$$
$$= £8571$$

The workers might well argue however that this is **not** a typical wage at the company!

In general though, the mean of a set of data x_i i.e. $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{\Sigma x_i}{n}$$

The summation is over i, but often for shorthand it is simply written as

$$\bar{x} = \frac{\Sigma x}{n}$$

Activity 6 What do you mean?

In the BBC 'Yes Minister' programme the Prime Minister instructs his Private Secretary to give the Press the average wage of a group of workers. The Private Secretary asks, 'Do you mean the wage of the average worker or the average of all the workers' wages?' The PM replies, 'But they are the same thing, aren't they?' Do you agree?

Exercise 3G

Employment in manufacturing

% of total civilian employment

	1960	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983
Canada	23.7	22.3	21.8	21.8	22.0	21.7	20.2	20.3	19.6	19.6	19.9	19.7	19.3	18.1	17.5
US	27.1	26.4	24.7	24.3	24.8	24.2	22.7	22.8	22.7	22.7	22.7	2.1	21.7	20.4	19.8
Japan	21.5	27.0	27.0	27.0	27.4	27.2	25.8	25.5	25.1	24.5	24.3	24.7	24.8	24.5	24.5
France	27.5	27.8	28.0	28.1	28.3	28.4	27.9	27.4	27.1	26.6	26.1	25.8	25.1	24.7	24.3
W. Germany	37.0	39.4	37.4	36.8	36.7	36.4	35.6	35.1	35.1	34.8	34.5	34.3	33.6	33.1	32.5
Italy	23.0	27.8	27.8	27.8	28.0	28.3	28.2	28.0	27.5	27.1	26.7	26.7	26.1	25.7	24.7
Netherlands	30.6	26.4	26.1	25.6	25.4	25.6	25.0	23.8	23.2	23.0	22.3	21.5	20.9	20.5	20.3
Norway	25.3	26.7	25.3	23.8	23.5	23.6	24.1	23.2	22.4	21.3	20.5	20.3	20.2	19.7	18.2
UK	36.0	34.5	33.9	32.8	32.2	32.3	30.9	30.2	30.3	30.0	29.3	28.1	26.2	25.3	24.5

1. The information in the table above gives the percentage of workers employed in the manufacturing industry in the major industrial nations. Find the average percentage employed for 1960, 1975 and 1983. What does this tell you about the involvement of people in manufacturing industry in this period?

2. The results shown opposite are the final positions in the First Division Football in the 1990/91 season.

   (a) Total the goals scored both home and away and hence find the mean number of goals scored per match for each team.

   (b) Plot a scattergram of x, position in league, against y, average goals scored. How true is it that a high goal scoring average leads to a higher league position?

Division One

Pos		P	Home W	Home D	Home L	Home F	Home A	Away W	Away D	Away L	Away F	Away A	Pts
1	Arsenal	38	15	4	0	51	10	9	9	1	23	8	83
2	Liverpool	38	14	3	2	42	13	9	4	6	35	27	76
3	Crystal Pal	38	11	6	2	26	17	9	3	7	24	24	69
4	Leeds Utd	38	12	2	5	46	23	7	5	7	19	24	64
5	Man City	38	12	3	4	35	25	5	8	6	29	28	62
6	Man Utd	37	11	3	4	33	16	5	8	6	24	28	58
7	Wimbledon	38	8	6	5	28	22	6	8	5	25	24	56
8	Nottm For	38	11	4	4	42	21	3	8	8	23	29	54
9	Everton	38	9	5	5	26	15	4	7	8	24	31	51
10	Chelsea	38	10	6	3	33	25	3	4	12	25	44	49
11	Tottenham	37	8	9	2	35	22	3	6	9	15	27	48
12	QPR	38	8	5	6	27	22	4	5	10	17	31	46
13	Sheff Utd	38	9	3	7	23	23	4	4	11	13	32	46
14	Southptn	38	9	6	4	33	22	3	3	13	25	47	45
15	Norwich	38	9	3	7	27	32	4	3	12	14	32	45
16	Coventry	38	10	6	3	30	16	1	5	13	12	33	44
17	Aston Villa	38	7	9	3	29	25	2	5	12	17	33	41
18	Luton	38	7	5	7	22	18	3	2	14	20	43	37

(c) The table below gives, amongst other information, the mean 'Goals Scored' and 'Goals Conceded' for the successful years of Arsenal. What do these 'averages' tell you about the scores in matches of earlier years?

Seasons of success: How Arsenal's past and present League triumphs measure up

	Games							Average goals per match	
Season	**P**	**W**	**D**	**L**	**Pts**	**F**	**A**	**Scored**	**Conceded**
1990 - 91	38	24	13	1	83	74	18	1.95	0.47
1988 - 89	38	22	10	6	76	73	36	1.92	0.95
1970 - 71	42	29	7	6	85	71	29	1.69	0.69
1932 - 33	42	25	8	9	75	118	61	2.81	1.45

3. Find the mean playing time of the tracks of one of your albums. How does this compare with your median time? Which do you think is a better measure?

3.9 Using your calculator

Most modern calculators have a statistical function. This enables a running check to be kept on the total and number of results entered. Check your instruction booklet on how to do this. It is good practice when entering a set of values always to check the n memory to ensure you haven't missed a value out or put in too many. A common fault is to forget to clear a previous set of results.

When dealing with large amounts of data it is easy to make a mistake in adding up totals or entering. For example, the number of children in families for a class of children was recorded opposite:

No. of children (x)	Frequency (f)
1	8
2	11
3	6
4	4
5	1

The total could be found by repeated addition,

i .e $1+1+1+1+1+1+1+1+2+2 \ldots +4+4+4+4+5.$

However, it is far simpler to multiply the x values by the frequencies,

i.e. $(1\times 8)+(2\times 11)+(3\times 6)+(4\times 4)+(5\times 1).$

So if n is the sum of the frequencies, in general

$$\bar{x} = \frac{\Sigma\, x_i\, f_i}{\Sigma\, f_i} \text{ when } n = \Sigma\, f_i$$

Most calculators can automatically enter frequencies - check your calculator instructions carefully.

With grouped frequency tables the same principle applies except that for the x value the mid-mark of the group is used (i.e. the value half way between the class limits). This is not entirely accurate as it assumes an **even spread** of data within the group. Usually differences above and below will cancel out but beware of quoting values with too high a degree of accuracy. The ages of people injured in road accidents in Cornwall in 1988 are shown opposite.

Age	Mid-mark	Frequency	$x \times f$
1 -10	6	199	1194
11-20	16	895	14320
21-30	26	625	16250
31-40	36	388	13968
41-50	46	261	12006
51-60	56	153	8568
61-70	66	141	9306
71+	76	140	10640
		2802	86252

Since an age of 1 – 10 really means from 1 right up to (but not including) 11, its midpoint is 6. Similarly for the other intervals.

This gives

$$\bar{x} = \frac{86252}{2802} \approx 31$$

Note that in the last open ended group a mid-mark of 76 was used to tie in with other groups. However, as this has a high frequency it could be a cause of error if there were, in fact, a significant number of over 80-year-olds involved in accidents.

Exercise 3H

1. The table opposite shows the wages earned by YTS trainees in 1984. Do you think that the mean of £28.10 is a fair figure to quote in these circumstances? What figure would you quote and why?
2. Find the mean number of shares issued by Cable & Wireless PLC as given in Exercise 3F, Q3. Why is there such a difference between the median and the mean? What information might be useful in obtaining a more accurate estimate of the mean?

Weekly income of trainees (March 1984)

Income	Per cent of trainees
£25.00	84
Over £25.00 up to £30.00	3
Over £30.00 up to £35.00	3
Over £35.00 up to £40.00	1
Over £40.00 up to £50.00	4
Over £50.00 up to £60.00	3
Over £60.00	2
	100

Mean £28.10

3.10 How spread out are the data?

Activity 7 Do differences in height even out as you get older?

Earlier you collected heights of people in your own age group. Collect at least 20 heights of people in an age group four or five years younger. Is there more difference in heights in the younger age group than in the older?

This section will examine ways of looking at this.

Example

Multiple discipline endurance events have gained in popularity over the last few years. The data on the next page gives the results of the first 50 competitors in a biathlon race consisting of a 15 mile bike ride followed by a 5 mile run. Some competitors argued that the race was biased towards cyclists as a good cyclist could make up more time in the cycling event which she or he would not lose on the shorter event. What you need to consider here is whether cycling times are more varied than running times.

Solution

The simplest way this could be done would be to look at the difference between the fastest and slowest times for each part. This is the **range**.

For cycling

$$\text{range} = 1\text{h } 9\text{s} - 44\text{min } 50\text{s} = 15\text{min } 19\text{s}$$

and for running

$$\text{range} = 48\text{min } 51\text{s} - 32\text{min } 23\text{s} = 16\text{min } 28\text{s}.$$

So, on the face of it, running times are more spread out than cycling times. However, in both sets of figures there are unrepresentative results at the end of the range which can on their own account for the difference in ranges. The range is therefore far too prone to effects of extremes, called **outliers**, and is of limited practical use.

Some statisticians use $\frac{n}{2}$ for the median, $\frac{n}{4}$, $\frac{3n}{4}$ for the quartiles when using grouped data – this is acceptable, and would not be penalised in the AEB Statistics Examination.

To overcome this, the **inter-quartile range** (IQR) attempts to miss out these extremes. The **quartiles** are found in the same way as the median but at the $\frac{(n+1)}{4}$ th and $\frac{3(n+1)}{4}$ th item of data. Taking just the fastest seven items of cycling data, look for the quartiles at the 2nd and 6th item:

44:50	45:25	47:15	47:16	48:07	48:07	48:18
	↑		↑		↑	
	lower quartile (LQ)		median		upper quartile (UQ)	

The inter-quartile range $= 48.07 - 45.25 = 2\text{min } 42\text{s}.$

This tells you the range within which the middle 50% of data lies. In some cases, where the data are roughly symmetrical, the **semi inter-quartile range** is used. This gives the range either side of the median which contains the middle 50% of data.

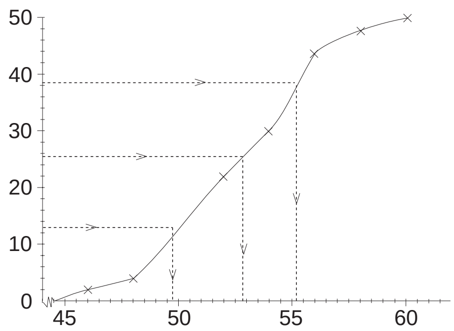
Mildenhall C.C.
Biathalon 30.8.87
Results

Finishing order Position	**No**	**Name**	**Club**	**Cycle Time**	**Run Time**	**Total Time**
1	157	Roy E. Fuller	Ely & Dist C.C.	48.18	33.55	1.22.13
2	106	Clive Catchpole	Fitness Habit (Ipswich)	45.25	36.59	1.22.24
3	108	Robert Quarton	Fitness Habit (Ipswich)	48.50	33.45	1.22.35
4	26	Michael Bennett	Fitness Habit (Ipswich)	47.15	35.47	1.23.02
5	110	David Minns	West Suffolk A.C.			
			Mildenhall C.C/Dairytime	51.00	32.32	1.23.32
6	30	Christopher Neale	Surrey Road C.C.	48.07	36.33	1.24.40
7	46	Roger Jackerman	Met Police A.A.	50.15	35.14	1.25.29
8	60	David Chamborlain	Scalding C.C.			
			Holbeach A.C.	48.07	37.39	1.25.46
9	66	Nigel Morrison	Halstead Roadrunners	48.50	37.15	1.26.05
10	80	Michael Meyer		49.50	37.04	1.26.54
11	143	Paul Chapman	Bishop Stortford C.C.	50.00	37.10	1.27.10
12	120	Chris Carter	North Bucks R.C.	47.16	39.57	1.27.13
13	123	Ian Coles	Colchester Rovers	49.55	37.43	1.27.38
14	102	Stephen Nobbs	North Norfolk			
			Beach Runners	53.12	34.42	1.27.54
15	171	David Smith	Ipswich Jaffa	55.46	32.23	1.28.09
16	129	Don Hutchinson	Sir M. McDonald & Partners			
			Running Club	52.03	36.08	1.28.11
17	50	Bill Morgan	Diss & Dist Wheelers	49.15	37.46	1.29.01
18	169	C. Willmets	Cambridge Triathlon	50.45	38.32	1.29.51
19	155	John Wright	Duke St. Runners	55.25	34.11	1.29.36
20	58	R. F. Williams	North Norfolk			
			Beach Runners	52.50	37.01	1.29.51
21	187	Jon Trevor	East London Triathletes			
			Unity C.C.	51.30	38.22	1.29.52
22	18	Julian Tomkinson		55.12	34.55	1.30.07
23	181	G. Carpenter		58.15	32.38	1.30.53
24	56	Duncan Butcher	St. Edmund Pacers	55.42	35.18	1.31.00
25	147	H. D. Ward	Colchester Rovers	49.45	41.39	1.31.24
26 =	40	Jeffrey P. Hathaway	North Bucks R.C.	44.50	46.51	1.31.41
26 =	12	Steven Elvin		55.15	36.26	1.31.41
28	165	Geoffrey Davidson	Wymondham Joggers	53.00	38.43	1.31.43
29	175	Mike Parkin	Deeping C.C.	50.35	41.50	1.32.35
30	149	Pete Cotton	Mildenhall C.C./Dairytime	54.25	38.21	1.32.46
31	84	Barry Parker	Thetford A.C.			
			Wymondham Joggers	53.48	39.17	1.33.05
32	90	Keith Tyler	Wisbech Wheelers			
			Cambs Speed Skaters	48.45	44.54	1.33.39
33	36	Derek Ward	Duke St. Runners	54.10	39.41	1.33.51
34	38	Gordon Bidwell	West Norfolk A.C.	55.17	38.36	1.33.53
35	139	John M. Chequer	Granta Harriers	54.35	39.55	1.34.30
36	59	Jeremy Hunt	ABC Centerville	53.20	41.5	1.34.35
37	133	W. E. Clough	Cambridge			
			Town & County C.C.	52.32	42.22	1.34.54
38	163	Bruce Short	West Norfolk Rugby Union	51.10	44.02	1.35.12
39	185	Kate Byrne	East London Triathletes			
			Unity C.C.	54.05	41.17	1.35.22
40	29	Justin Newton	Mildenhall C.C./Dairytime	56.20	40.54	1.37.14
41	127	S. Kennett		58.40	38.45	1.37.25
42	14	David J. Cassell	Bungay Black Dog	57.59	40.11	1.38.10
43	78	Roger Temple		54.27	44.26	1.38.53
44	141	Lulu Goodwin		53.37	45.37	1.39.14
45	48	Patrick Ash	North Norfolk			
			Beach Runners			
			North Norfolk Wheelers	55.27	44.06	1.39.33
46	62	Philip Mitchell		55.54	43.44	1.39.38
47	76	Parry Pierson Cross	Havering C. T. C.	50.48	48.51	1.39.39
48	118	Geoff Holland	Wymondham Joggers	57.12	42.44	1.39.56
49	197	Terry Scott		1.00.09	40.01	1.40.10
50	137	Nigel Chapman	Bishop Stortford C.C.	57.45	42.33	1.40.18

With grouped data you can use either the interpolation method or a cumulative frequency curve to find the quartiles and hence the IQR. For cycling, the graphed data are summarised opposite.

Cycling Times	Frequencies	Cumulative Frequency
44:00-45:59	2	2
46:00-47:59	2	4
48:00-49:59	10	14
50:00-51:59	8	22
52:00-53:59	8	30
54:00-55:59	13	43
56:00-57:59	4	47
58:00 +	3	50

The cumulative frequency curve is shown below. Note that you plot (46, 2), (48, 4), etc. but that the last point cannot from this grouped data be plotted.



The median is given by the

$$\frac{(50+1)}{2} = 25.5 \text{ th}$$

item of data. So drawing across to the cumulative frequency curve and then downwards gives an estimate of the median as 52.7.

Similarly estimates for the quartiles are given by the

$$\frac{(50+1)}{4} = 12.75\text{th item}$$

and the $$\frac{3(50+1)}{4} = 38.25\text{th item.}$$

This gives estimates

$$\text{LQ} = 49.7 \text{ min}, \quad \text{UQ} = 55.2 \text{ min}$$

with an inter-quartile range of $55.2 - 49.7 = 5.5$ min.

Using interpolation, the lower quartile is at the 12.75th item, and an estimate for this, since there are 4 items up to 48:00 and 10 items in the next group which has class width 2, is given by

$$\text{LQ} = 48.0 + \left[\frac{(12.75-4)}{10} \times 2\right]$$

$$= 49.8 \text{ min}.$$

Similarly the upper quartile is the 38.25 th item, and an estimate is

$$\text{UQ} = 54.00 + \left[\frac{(38.25-30)}{13}\times 2\right]$$

$$= 55.3 \text{ min}.$$

Hence the inter-quartile range is given by

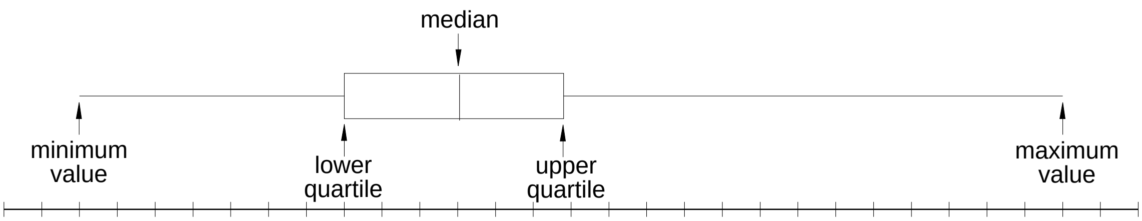
$$\text{IQR} = 55.3 - 49.8 = 5.5 \text{ min}.$$

If a stem and leaf diagram has been used, the median and quartiles can be taken from the data directly. To assist in this, the cumulative frequencies are calculated working from both ends to the middle. The stem and leaf diagram for the **rounded decimal times** is shown opposite. The stem is in minutes, and the leaf is rounded to one d.p. of a minute.

(1)	44	8	
(2)	45	4	
(2)	46		
(4)	47	33	
(10)	48	113888	
(14)	49	3(88)9	Lower quartile
(19)	50	03688	
(22)	51	025	
(25)	52	15(8)	
(25)	53	(0)368	Median
(21)	54	12456	
(16)	55	233(45)7899	Upper quartile
(7)	56	3	
(6)	57	28	
(4)	58	137	
(1)	59		
(1)	60	2	

A new form of diagram, using the median and quartiles, is becoming increasingly popular. The **box and whisker plot** shows the data on a scale and is very useful for comparing the 'distribution' of several sets of data drawn on the same scale.

The box is formed by using the two quartiles, and the median is illustrated by a line. The whiskers are found by using minimum and maximum values, as illustrated below.



Example

Use a box and whisker plot to illustrate the following two sets of data relating to exam results of 11 candidates in Mathematics and English.

Pupil	A	B	C	D	E	F	G	H	I	J	K
Maths	62	91	43	31	57	63	80	37	43	5	78
English	65	57	55	37	62	70	73	49	65	41	64

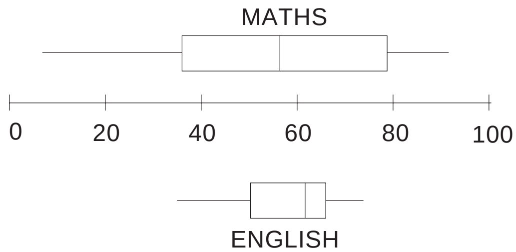
Solution

Rearrange each set of data into increasing order.

Maths 5 31 37 43 43 57 62 63 78 80 91

↑ LQ ↓ (37 / 49) ↑ median ↓ (57 / 62) ↑ UQ ↓ (78 / 65)

English 37 41 49 55 57 62 64 65 65 70 73



This diagram helps you to see quickly the main characteristics of the data distribution for each set. It does not, however, enable comparisons to be made of the relative performances of candidates.

Exercise 3I

1. Using any method find the IQR of the running times shown in the table of biathlon results at the start of this section. Are the competitors justified in their complaint?
2. Find the median and IQR for the heights of both age groups measured in earlier activities. Are heights more varied at a particular age?
3. When laying pipes, engineers test the soil for 'resistivity'. If the reading is low then there is an increasing risk of pipes corroding. In a survey of 159 samples the following results were found:

Resistivity (ohms/cm)	Frequency
400 - 900	5
901 - 1500	9
1501 - 3500	40
3501 - 8000	45
8001 - 20000	60

Find the median and inter-quartile range of this data.

3.11 Standard deviation

Like the median, the quartiles fail to make use of all the data. This can of course be an advantage when there are extreme items of data. There is a need then for a measure which makes use of **all** data. There is also a need for a measure of **spread** which relates to a central value. For example, two classes who sat the same exam might have the same mean mark but the marks may vary in a different pattern around this. It seems sensible if you are using all the data that the measure of spread ought to be related to the mean.

One method sometimes used is the **mean deviation from the mean**.

For example, take the following data:

6, 8, 8, 9, 14, 15,

the mean of which is 10.

The differences, or deviations, of these from the mean are given by

$$-4,\ -2,\ -2,\ -1,\ +4,\ +5.$$

To find a summary measure you first need to combine these, but by simply adding them together you will always get zero.

Why is the sum of the deviations always zero?

The mean deviation simply ignores the sign, using what is known in mathematics as the **modulus**, e.g. $|-3| = 3$ and $|3| = 3$. In order that the measure is not linked to the size of sample, you then average the deviations out:

$$\text{mean deviation from the mean} = \frac{1}{n}\Sigma\left|x_i - \bar{x}\right|$$

In the example, this has value $\frac{1}{6}(4+2+2+1+4+5) = 3$.

However, just ignoring signs is not a very sound technique and the mean deviation is not often used in practice.

Activity 8 Pulse rates

The pulse rates of a group of 10 people were:

72, 80, 67, 68, 80, 68, 80, 56, 76, 68.

The mean of this data is about 70. Now calculate the deviations of all the values from this 'assumed' mean. Instead of just ignoring the signs however, square the deviations and add these together,

i.e $\quad 2^2 + 10^2 + 3^2 + 2^2 + 10^2 + 2^2 + 10^2 + 14^2 + 6^2 + 2^2 = 557$

Note how the sign now becomes irrelevant.

Repeat this with other assumed means around the same value and put the results in a table (it will save time to work in a group):

Assumed mean	67	68	69	69.5	70	70.5	71	72	73
Σd^2					557				

Now plot a graph of these results.

What you should find in this activity is that the results form a quadratic graph. The value of assumed mean at the bottom of the graph is the value for which the sum of the squared deviations is the least. Find the arithmetic mean of your data and you may not be surprised to find that this is the same value. This idea is an important one in statistics and is called the **'least squares method'**.

Squaring the deviations then is an alternative to using the modulus and the result can be averaged out over the number of items of data. This is known as the **variance**. However, the value can often be disproportionately large and it is more common to square root the variance to give the **standard deviation** (SD). So

$$\text{variance} \quad s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

$$\text{standard deviation} \quad s = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$$

Example

Find the standard deviation of the pulse rates in Activity 8.

Solution

$\bar{x} = 71.6$, so you have the following table:

	72	80	67	68	80	68	80	56	76	69
$\lvert x - \bar{x} \rvert$	0.4	8.4	4.6	3.6	8.4	3.6	8.4	15.6	4.4	2.6
$(x - \bar{x})^2$	0.16	70.56	21.16	12.96	70.56	12.96	70.56	243.36	19.36	6.76

giving $\quad \Sigma(x - \bar{x})^2 = 528.40$.

Hence variance, $\quad s^2 = \dfrac{528.40}{10} = 52.84$

and standard deviation, $\quad s \approx 7.27$.

It is very tedious to calculate by this method – even using a calculator you would have problems, as the calculator would have to memorise all the data until the mean could be calculated. An alternative formula often used is

$$s^2 = \left(\frac{1}{n}\Sigma x^2\right) - \bar{x}^2$$

You can derive this result by noting that

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

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

$$= \frac{1}{n}\Sigma x_i^2 - \frac{2\bar{x}}{n}\Sigma x_i + \frac{\bar{x}^2}{n}\Sigma 1.$$

But $\frac{1}{n}\Sigma x_i = \bar{x}$ and $\Sigma 1 = n$,

giving $s^2 = \frac{1}{n}\Sigma x_i^2 - 2\bar{x}^2 + \bar{x}^2$

or $s^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$.

Calculators use this method and keep a running total of

(a) n the quantity of data entered,

(b) Σx the running total,

(c) Σx^2 the sum of the values squared.

This is illustrated opposite, and

x	Σx	Σx^2
72	72	5184
80	152	11584
67	219	16073
..	..	..
..	..	..
..	..	..
69	716	51794

$$\bar{x} = \frac{716}{10} = 71.6$$

$$s = \sqrt{\frac{51794}{10} - 71.6^2} = 7.27.$$

Find out how to use your calculator to calculate the standard deviation (SD). Most will give you all the values in the above formula too.

What does the standard deviation stand for?

Whereas you were able to say that the IQR was the range within which the middle 50% of a data set lies there is no absolute meaning that can be given to the SD. On its own then it can be difficult to judge the significance of a particular SD.

It is of more use to compare two sets of data.

Example

Compare the means and standard deviation of the two sets of data

(a) 3, 4, 5, 6, 7

(b) 1, 3, 5, 7, 9

Solution

(a) $\bar{x} = \dfrac{3+4+5+6+7}{5} = 5,$

and $s^2 = \dfrac{1}{5}(9+16+25+36+49) - 25$

$= 27 - 25 = 2,$

giving $s \approx 1.414$.

(b) As in (a), $\bar{x} = 5,$

but $s^2 = \dfrac{1}{5}(1+9+25+49+81) - 25$

$= 33 - 25 = 8,$

giving $s \approx 2.828$.

Thus the two sets of data have equal means but since the spread of the data is very different in each set, they have different SDs. In fact, the second SD is double the first.

Activity 9

Construct a number of data sets similar to those in the example, which all have the same means. Estimate what you think the standard deviation will be. Now calculate the values and see if they agree with your intuitive estimate.

Activity 10

Find the standard deviation of the album track length data used earlier. Do some albums have more varied track lengths than others?

With grouped frequency tables the SD can be calculated as follows. Find Σx and Σx^2 by multiplying the frequency by the mid-marks and the mid-marks squared respectively.

e.g.

Height	Frequency	Σx	Σx^2
140-149	5	5×144.5	$5 \times (144.5)^2$

As with means, most modern calculators can perform these operations in statistical mode.

Example

The lengths of 32 fish caught in a competition were measured correct to the nearest mm. Find the mean length and the standard deviation.

Length	20-22	23-25	26-28	29-31	32-34
Frequency	3	6	12	9	2

Solution

Group	Mid-point (x)	Frequency (f)	fx	$f(x^2)$
20-22	21	3	63	1323
23-25	24	6	144	3456
26-28	27	12	324	8748
29-31	30	9	270	8100
32-34	33	2	66	2178
		$\Sigma f = 32$	$\Sigma fx = 867$	$\Sigma fx^2 = 23805$

So $$\bar{x} = \frac{\Sigma x_i}{n} = \frac{\Sigma fx}{\Sigma f} = \frac{867}{32} \approx 27.1$$

and $$s^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2 = \frac{\Sigma fx^2}{\Sigma f} - \bar{x}^2$$

$$= \frac{23805}{32} - \left(\frac{867}{32}\right)^2 \approx 9.835$$

$$\Rightarrow \quad s \approx 3.14$$

Note that, for grouped data, the general formulae for mean and standard deviation became

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}, \quad s^2 = \frac{\Sigma fx^2}{\Sigma f} - \bar{x}^2.$$

Exercise 3J

1. From the frequency tables drawn up earlier for the biathlon race find the standard deviations of the running and cycling times. Are cycling times more varied?
2. The data opposite give the age of mothers of children born over the last 50 years. Find the mean and SD of the ages for 1941, 1961 and 1989. What does this tell you about the change in the age at which women are tending to have children?

Live births: by age of mother

Great Britain — **Percentages**

Age of mother	1941	1951	1961	1971	1981	1989
15-19	4.3	4.3	7.2	10.6	9.0	8.2
20-24	25.4	27.6	30.8	36.5	30.9	26.9
25-29	31.0	32.2	30.7	31.4	34.0	35.4
30-34	22.1	20.7	18.8	14.1	19.7	21.1
35-39	12.7	11.5	9.6	5.8	5.3	7.0
40-44	4.2	3.4	2.7	1.5	1.0	1.3
45-49	0.3	0.2	0.2	0.1	0.1	0.1

(Source: Population Censuses and Surveys Scotland)

3. The data below give the usual working hours of men and women, both employed and self-employed. Find the mean and standard deviation of the four groups and use this information to comment on the differences between men and women and employed/self-employed people.

Basic usual hours worked: by sex and type of employment, 1989

Great Britain				Percentages
	Males		**Females**	
	Employees	**Self employed**	**Employees**	**Self employed**
Hours per week				
Less than 5	0.4	1.0	2.2	6.0
5 but less than 10	1.1	0.9	6.5	7.3
10 but less than 15	1.0	1.1	7.8	9.2
15 but less than 20	0.7	0.9	9.4	7.4
20 but less than 25	0.9	1.6	10.9	8.5
25 but less than 30	1.0	1.3	5.9	5.4
30 but less than 35	2.6	3.2	6.9	7.7
35 but less than 40	50.7	8.6	38.7	9.1
40 but less than 45	28.6	26.0	9.1	13.1
45 but less than 50	5.2	12.5	1.0	6.3
50 but less than 55	3.0	12.7	0.6	4.4
55 but less than 60	1.3	4.6	0.2	2.4
60 and over	3.2	25.2	0.6	12.8

(Source: Labour Force Survey Employment Department)

(NB Column totals do not sum exactly to 100 due to rounding errors in individual entries.)

3.12 Miscellaneous Exercises

1. The data below show the length of marriages ending in divorce for the period 1961-1989. Using the data for 1961, 1971, 1981 and 1989:

   (a) draw any diagrams which you think useful to illustrate the pattern of marriage length;

   (b) calculate any measures which you think appropriate;

   (c) write a short report on the pattern of marriage breakdowns over this period.

Percentages and thousands

Year of divorce	**1961**	**1971**	**1976**	**1981**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**
Duration of marriage (percentages)											
0-2 years	1.2	1.2	1.5	1.5	1.3	1.2	8.9	9.2	9.3	9.5	9.8
3-4 years	10.1	12.2	16.5	19.0	19.5	19.6	18.8	15.3	13.7	13.4	13.4
5-9 years	30.6	30.5	30.2	29.1	28.7	28.3	36.2	27.5	28.6	28.0	28.0
10-14 years	22.9	19.4	18.7	19.6	19.2	18.9	17.1	17.5	17.5	17.5	17.6
15-19 years	13.9	12.6	12.8	12.8	12.9	13.2	12.2	12.8	13.0	13.2	13.0
20-24 years		9.5	8.8	8.6	8.6	8.7	7.9	8.4	8.7	9.1	9.0
25-29 years	21.2	5.8	5.6	4.9	5.2	5.3	4.7	4.8	4.9	4.9	4.9
30 years and over		8.9	5.9	4.5	4.7	4.6	4.2	4.3	4.3	4.3	4.3
All durations (= 100%) (thousands)	27.0	79.2	134.5	155.6	160.7	156.4	173.7	166.7	163.1	164.1	162.5

2. As a result of examining a sample of 700 invoices, a sales manager drew up the grouped frequency table of sales shown opposite.

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

   (b) Explain why the mean and the standard deviation might not be the best summary statistics to use with these data.

   (c) Calculate estimates of alternative summary statistics which might be used by the sales manager. Use these estimates to justify your comment in (b). (AEB)

Amount on invoice (£)	**Number of invoices**
0-9	44
10-19	194
20-49	157
50-99	131
100-149	69
150-199	40
200-499	58
500-749	7

3. Using the number of incomes in each category, calculate the mean income in 1983/4 and 1984/5.

   Do you think these are the best measures to use here? Give your reasons and suggest alternative measures.

1983/84 Annual Survey

Lower limit of range of income	Number of incomes (Thousands)
All incomes	22 015
Income before tax £	
1 500	509
2 000	1 230
2 500	1 070
3 000	1 200
3 500	1 220
4 000	1 240
4 500	1 130
5 000	1 140
5 500	1 100
6 000	1 890
7 000	1 710
8 000	2 810
10 000	2 040
12 000	1 740
15 000	1 120
20 000	645
30 000	169
50 000	44
100 000 and over	8

1984/85 Annual Survey

Lower limit of range of income	Number of incomes (Thousands)
All incomes	22 164
Income before tax £	
2 000	1 340
2 500	1 000
3 000	1 060
3 500	1 090
4 000	1 210
4 500	1 090
5 000	1 060
5 500	1 985
6 000	1 190
7 000	1 690
8 000	2 930
10 000	2 090
12 000	1 990
15 000	1 340
20 000	780
30 000	246
50 000	62
100 000 and over	11

4. The table opposite shows the lifetimes of a random sample of 200 mass produced circular abrasive discs.

   (a) Without drawing the cumulative frequency curve, calculate estimates of the median and quartiles of these lifetimes.

   (b) One method of estimating the skewness of a distribution is to evaluate

   $$\frac{3\,(\text{mean} - \text{median})}{\text{standard deviation}}.$$

   Carry out the evaluation for the above data and comment on your result.

   Use the quartiles to verify your findings.

   (AEB)

Lifetime (to nearest hour)	Number of discs
690-709	3
710-719	7
720-729	15
730-739	38
740-744	41
745-749	35
750-754	21
755-759	16
760-769	14
770-789	10

5. The following information is taken from a government survey on smoking by schoolchildren.

Cigarette consumption (per week)	England and Wales 1982	1984	1986
Boys	%	%	%
None	12	13	12
1-5	24	24	25
6-40	33	31	30
41-70	16	16	18
71 and over	16	14	15
Mean	33	31	33
Median	15	16	20
Base (= 100%)	*272*	*419*	*210*
Girls			
None	13	10	10
1-5	29	26	21
6-40	32	34	38
41-70	14	15	16
71 and over	11	14	15
Mean	26	30	32
Median	11	14	17
Base (= 100%)	*289*	*373*	*266*

(a) Both the mean and median have been calculated for each category. Why do these differ so much? Which would you prefer as a suitable measure in this survey?

(b) Write a short report using suitable illustrations on the pattern of teenage smoking over the years 1982-1986.

6. The data below form part of a survey on the TV watching habits of schoolchildren.

(a) Find the mean and SD for boys and girls in each age group and comment on any differences.

(b) By combining the boys' and girls' standard deviations and means, assuming an equal number of each took part in the survey, find overall figures for each age group.

7. In order to monitor whether large firms are taking over from smaller ones the government carries out a survey on company size at regular intervals. The results of such a survey are shown below.

(a) Draw a relative frequency histogram of the data.

(b) Calculate the mean and standard deviation of the size of companies.

(c) Find the median and quartiles of the data and use these to draw a box and whisker plot.

(d) Comment on the suitability of the measures in (b) and (c) and any inaccuracies in the calculation techniques.

Size bands according to numbers of employees	Census units numbers	%
1-10	847 537	73.6
11-24	169 800	14.7
25-49	70 671	6.1
50-99	32 888	2.9
100-199	17 236	1.5
200-499	9 352	0.8
500-999	2 605	0.2
1000+	1 476	0.1
Total	1 151 565	100.0

(Source: Department of Employment, Statistics Division, 1988)

8. 38 children solved a simple problem and the time taken by each was noted.

Time (seconds)	5-	10-	20-	25-	40-	45-
Frequency	2	12	7	15	2	0

Draw a histogram to illustrate this information.

	1st year(11+)		3rd year(13+)		5th year(15+)	
	Boys	Girls	Boys	Girls	Boys	Girls
None	5.3	6.6	4.9	6.0	6.9	8.1
Less than 1hr	13.6	16.9	12.7	16.5	14.4	19.2
1-2hr	20.4	23.4	18.8	21.7	20.8	22.7
2-3hr	19.4	18.4	21.7	18.4	21.0	20.0
3-4hr	14.6	15.0	18.1	16.7	16.1	14.9
4-5hr	11.3	9.3	9.7	9.8	10.3	7.5
5hrs or longer	15.4	10.4	14.1	10.8	10.3	7.6

9. The number of passengers on a certain regular weekday train service on each of 50 occasions was:

165	141	163	153	130	158	119	187	185	209
177	147	166	154	159	178	187	139	180	143
160	185	153	168	189	173	127	179	163	182
171	146	174	149	126	156	155	174	154	150
210	162	138	117	198	164	125	142	182	218

Choose suitable class intervals and reduce these data to a grouped frequency table.

Plot the corresponding frequency polygon on squared paper using suitable scales. (AEB)

10. The percentage marks of 100 candidates in a test are given in the following tables:

No. of marks	0-19	20-29	30-39	40-49
No. of candidates	5	6	13	22

No. of marks	50-59	60-69	70-79	80-89
No. of candidates	24	16	8	6

Draw a cumulative frequency curve.

Hence estimate

(i) the median mark,

(ii) the lower quartile,

(iii) the upper quartile. (AEB)

11. The number of passengers on a certain regular weekday bus was counted on each of 60 occasions. For each journey, the number of passengers in excess of 20 was recorded, with the following results.

15	6	13	8	9	12	8	11	5	12
7	11	7	11	10	10	7	9	14	10
6	7	9	12	13	9	8	8	12	14
9	10	11	13	8	8	8	11	8	13
12	14	13	7	8	6	11	10	15	10
8	13	7	12	9	10	9	8	11	9

(a) Construct a frequency table for these data.

(b) Illustrate graphically the distribution of the number of passengers per bus.

(c) For this distribution state the value of

(i) the mode,

(ii) the range. (AEB)

12. The breaking strengths of 200 cables, manufactured by a specific company, are shown in the table below.

Plot the cumulative frequency curve on squared paper.

Hence estimate

(a) the median breaking strength,

(b) the semi inter-quartile range,

(c) the percentage of cables with a breaking strength greater than 2300 kg.

Breaking strength (in 100s of kg)	Frequency
0-	4
5-	48
10-	60
15-	48
20-	24
25-30	16

13. The gross registered tonnages of 500 ships entering a small port are given in the following table.

Gross registered tonnage (tonnes)	No. of ships
0-	25
400-	31
800-	44
1200-	57
1600-	74
2000-	158
3000-	55
4000-	26
5000-	18
6000- 8000	12

Plot the percentage cumulative frequency curve on squared paper.

Hence estimate

(a) the median tonnage,

(b) the semi inter-quartile range,

(c) the percentage of ships with a gross registered tonnage exceeding 2500 tonnes.

(AEB)

14. The following table refers to all marriages that ended in divorce in Scotland during 1977. It shows the age of the wife at marriage.

Age of wife (years)	16-20	21-24	25-29	30/over
Frequency	4966	2364	706	524

(Source: Annual Abstract of Statistics, 1990)

(a) Draw a cumulative frequency curve for these data.

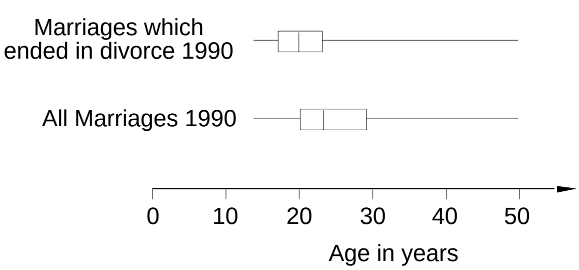
(b) Estimate the median and the inter-quartile range.

The corresponding data for 1990 revealed a median of 21.2 years and an inter-quartile range of 6.2 years.

(c) Compare these values with those you obtained for 1977. Give a reason for using the median and inter-quartile range, rather than the mean and standard deviation for making this comparison.

The box-and-whisker plots below also refer to Scotland and show the age of the wife at marriage. One is for all marriages in 1990 and the other is for all marriages that ended in divorce in 1990. (The small number of marriages in which the wife was aged over 50 have been ignored.)

Age of wife at marriage, Scotland



(d) Compare and comment on the two distributions. (AEB)

15. Give one advantage and one disadvantage of grouping data into a frequency table.

The table shows the trunk diameters, in centimetres, of a random sample of 200 larch trees.

Diameter (cm)	15-	20-	25-	30-	35-	40-50
Frequency	22	42	70	38	16	12

Plot the cumulative frequency curve of these data.

By use of this curve, or otherwise, estimate the median and the inter-quartile range of the trunk diameters of larch trees.

A random sample of 200 spruce trees yield the following information concerning their trunk diameters, in centimetres.

Min	Lower quartile	Median	Upper quartile	Max
13	27	32	35	42

Use this data summary to draw a second cumulative frequency curve on your graph.

Comment on any similarities or differences between the trunk diameters of larch and spruce trees. (AEB)

16. Over a period of four years a bank keeps a weekly record of the number of cheques with errors that are presented for payment. The results for the 200 accounting weeks are as follows.

Number of cheques with errors (x)	Number of weeks (f)
0	5
1	22
2	46
3	38
4	31
5	23
6	16
7	11
8	6
9	2

$$\left(\sum f x = 706 \quad \sum f x^2 = 3280\right)$$

Construct a suitable pictorial representation of these data.

State the modal value and calculate the median, mean and standard deviation of the number of cheques with errors in a week.

Some textbooks measure the **skewness** (or asymmetry) of a distribution by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

and others measure it by

$$\frac{(\text{mean} - \text{mode})}{\text{standard deviation}}.$$

Calculate and compare the values of these two measures of skewness for the above data.

State how this skewness is reflected in the shape of your graph.

(AEB)

17. Each member in a group of 100 children was asked to do a simple jigsaw puzzle. The times, to the nearest five seconds, for the children to complete the jigsaw are as follows:

Time (seconds)	60-85	90-105	110-125	130-145	150-165	170-185	190-215
No. of children	7	13	25	28	20	5	2

(a) Illustrate the data with a cumulative frequency curve.

(b) Estimate the median and the inter-quartile range.

(c) Each member of a similar group of children completed a jigsaw in a median time of 158 seconds with an inter-quartile range of 204 seconds. Comment briefly on the relative difficulty of the two jigsaws.

In addition to the 100 children who completed the first jigsaw, a further 16 children attempted the jigsaw but gave up, having failed to complete it after 220 seconds.

(d) Estimate the median time taken by the whole group of 116 children.

Comment on the use of the median instead of the arithmetic mean in these circumstances.

(AEB)