

16 DATA ANALYSIS

Objectives

After studying this chapter you should

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

16.0 Introduction

"Statistics is like garbage - you have to know what to do with it before you collect it"

The following page shows the amount spent per pupil in all of the Education Authorities in 1986. The information was presented in this way to enable readers to see how their local Authority fared compared with others. From the point of view of showing changing patterns in spending and regional differences the data is unintelligible and boring.

In this chapter you will see how masses of data such as this can be reduced to a form from which you can gain useful information and present it in a manner which easily and interestingly conveys these ideas to other readers. You may have met some of the basic techniques before, or possibly more advanced techniques, and it is not the purpose of this unit to give a detailed explanation of all the techniques.

What can be done to make this data more understandable and interesting?

Most of your ideas will come under three main headings:

- (a) **Sorting** - putting into order and/or grouping.
- (b) **Summary measures** - finding single values which give important facts about the various parts of the data.
- (c) **Illustration** - using various forms of diagrams to bring out patterns more clearly and present data in a more immediate manner.

	PUPIL TEACHER RATIOS					Costs 1983-4 to 1986-7				
	SECONDARY					UNIT COSTS SECONDARY				
	1983/4	1984/5	1985/6	1986/7	change 1985/6 1986/7	1983/4	1984/5	1985/6	1986/7	change 1985/6 1986/7
LONDON										
ILEA	13.0	12.8	12.4	12.1	-1.47%	£1,588	£1,754	£1,904	£2,011	5.64%
Barking and Dagenham	16.1	16.1	16.1	15.8	-2.17%	£1,027	£1,108	£1,200	£1,266	5.47%
Barnet	14.2	13.9	14.6	14.5	-0.07%	£1,036	£1,134	£1,110	£1,162	4.75%
Bexley	17.0	16.9	16.4	16.1	-1.38%	£987	£1,044	£1,099	£1,147	4.40%
Brent	13.6	12.7	11.4	11.2	-1.25%	£1,465	£1,630	£1,887	£2,063	9.35%
Bromley	15.9	15.7	15.9	15.7	-0.98%	£1,058	£1,126	£1,193	£1,299	8.87%
Croydon	16.7	15.6	15.5	14.7	-5.12%	£1,058	£1,126	£1,324	£1,428	7.82%
Ealing	14.8	14.2	14.3	13.6	-4.86%	£1,310	£1,455	£1,513	£1,673	10.55%
Enfield	15.4	14.8	15.0	14.9	-0.84%	£1,024	£1,070	£1,127	£1,208	7.14%
Haringey	13.0	12.5	12.4	11.9	-4.26%	£1,527	£1,765	£2,010	£2,039	1.46%
Harrow	14.3	13.9	14.4	14.5	0.94%	£1,247	£1,324	£1,392	£1,419	1.93%
Havering	16.1	15.8	15.5	14.6	-5.51%	£1,081	£1,142	£1,235	£1,271	2.89%
Hillingdon	15.8	15.8		16.2		£1,095	£1,176		£1,313	
Hounslow	14.9	15.3	15.7	15.4	-2.20%	£1,014	£1,059	£1,111	£1,175	5.78%
Kingston-upon-Thames	16.4	15.8	14.7	14.3	-2.30%	£946	£1,037	£1,187	£1,296	9.21%
Merton	18.0	18.1	16.6	16.8	1.24%	£876	£911	£1,007		
Newham	13.4	12.7	11.3	12.8	13.00%	£1,303	£1,408	£1,609	£1,628	1.15%
Redbridge	15.9	15.7	15.7	15.3	-2.88%	£1,129	£1,194	£1,267	£1,317	3.96%
Richmond-upon-Thames	16.2	16.7	16.3	16.4	0.60%	£1,116	£1,131	£1,217	£1,175	-3.42%
Sutton	17.1	17.1	16.9	16.7	-1.28%	£963	£1,010	£1,038	£1,127	8.63%
Waltham Forest	13.5	13.4	12.8	12.5	-2.20%	£1,290	£1,429	£1,525	£1,662	8.98%
METROPOLITAN DISTRICTS										
Bolton	16.3	15.9	15.5	15.6	0.43%	£889	£962	£1,052	£1,110	5.56%
Bury	16.0	15.7	15.6	15.4	-1.22%	£1,026	£1,128	£1,191	£1,229	3.12%
Manchester	15.8	14.4	13.8	13.6	-1.44%	£1,114	£1,212	£1,348	£1,443	7.04%
Oldham	16.5	16.0	15.5	15.1	-2.88%	£913	£972	£1,062	£1,131	6.51%
Rochdale	14.7	14.9	14.8	14.4	-2.69%	£1,052	£1,113	£1,191	£1,212	1.75%
Salford	15.8	15.4	15.2	15.3	0.49%	£970	£1,057	£1,158	£1,206	4.14%
Stockport	16.6	16.5	16.1	15.8	-1.90%	£901	£971	£1,070	£1,133	5.93%
Tameside	15.8	15.8	15.4	15.9	3.47%	£966	£1,036	£1,154	£1,159	0.40%
Trafford	16.2	16.1	15.7	16.4	4.16%	£1,110	£1,173	£1,317	£1,346	2.23%
Wigan	15.5	15.5	15.1	15.5	2.47%	£928	£998	£1,111	£1,156	4.11%
Knowsley	15.1	14.1	14.8	14.2	-4.20%	£1,107	£1,227	£1,315	£1,490	13.29%
Liverpool	16.9					£1,081				
St. Helens	15.9	15.6	15.2	14.8	-3.13%	£971	£1,050	£1,112	£1,214	9.22%
Sefton	16.7	16.7	16.8	16.6	-1.54%	£911	£964	£1,019	£1,081	6.06%
Wirral	17.5	16.2	15.8	15.9	0.77%	£931	£1,000	£1,130	£1,179	4.36%
Barnsley	16.7	16.6	16.2	15.3	-5.74%	£935	£1,021	£1,136	£1,190	4.71%
Doncaster	15.8	16.6	16.5	16.9	2.31%	£959	£1,002	£1,125	£1,144	1.64%
Rotherham	16.8	16.5	15.9	16.0	0.61%	£874	£940	£1,029	£1,125	9.30%
Sheffield	15.8	15.4		14.8		£1,044	£1,135		£1,333	
Gateshead	16.4	15.8	15.7	15.3	-2.52%	£989	£1,097	£1,143	£1,202	5.18%
Newcastle upon Tyne	15.1	15.1	15.4	17.2	11.55%	£1,109	£1,201	£1,278	£1,317	3.05%
North Tyneside	15.0	15.9	14.0	13.7	-2.17%	£1,025	£1,036	£1,232	£1,297	5.28%
South Tyneside	15.1	14.8	15.3	15.2	-0.67%	£1,101	£1,082	£1,149	£1,226	6.70%
Sunderland	16.0	15.6	15.4	15.2	-1.01%	£926	£1,009	£1,103	£1,153	4.56%
Birmingham	15.9	15.9	15.4	15.3	-0.50%	£911	£990	£1,083	£1,167	7.79%
Coventry	16.2	15.3	14.8	14.0	-5.35%	£986	£1,069	£1,179	£1,261	6.97%
Dudley	17.1	16.4	15.6	15.4	-1.44%	£867	£937	£1,055	£1,143	8.37%
Sandwell	15.4	15.4	14.4	14.5	0.98%	£1,000	£1,079	£1,207	£1,315	8.93%
Solihull	16.9	16.9	16.3	16.2	-0.76%	£856	£898	£1,008	£1,046	3.82%
Walsall	14.0	14.4	14.0	13.4	-4.30%	£1,029	£1,066	£1,154	£1,265	9.54%
Wolverhampton	15.0	14.8	14.5	15.0	3.53%	£1,027	£1,105	£1,184	£1,260	6.41%
Bradford	17.9	16.8	16.5	16.6	0.24%	£890	£965	£1,055	£1,156	9.59%
Calderdale	17.5	16.8	16.5	16.6	0.24%	£890	£965	£1,055	£1,156	9.59%
Kirklees	17.3	17.1	16.6	16.3	-2.15%	£851	£910	£983	£1,041	5.85%
Leeds	18.7	16.0	16.0	14.2	-11.12%	£887	£976	£1,043	£1,253	20.11%
COUNTIES										
Avon CC	16.3	16.4	16.3	16.2	-1.02%	£987	£1,062	£1,114	£1,174	5.35%
Bedfordshire CC	17.5	16.9	16.9	16.8	-0.37%	£972	£1,035	£1,109	£1,143	3.06%
Berkshire CC	16.0	15.7	15.6	15.6	0.16%	£982	£1,045	£1,129	£1,149	1.77%
Buckinghamshire CC	16.1	15.9	15.8	15.9	0.94%	£1,076	£1,136	£1,200	£1,286	7.13%
Cambridgeshire CC	16.8	16.9	16.5	16.2	-1.92%	£954	£996	£1,071	£1,129	5.35%
Cheshire CC	16.9	16.9	16.6	16.5	-0.48%	£1,004	£1,045	£1,118	£1,195	6.90%
Cleveland CC	16.2	16.1	15.4	15.8	2.84%	£970	£1,030	£1,131	£1,203	6.34%
Cornwall CC	16.8	16.8	16.5	16.4	-0.85%	£911	£975	£1,051	£1,088	3.52%
Cumbria CC	16.1	16.0	15.7	15.4	-1.75%	£968	£1,014	£1,131	£1,225	8.37%
Derbyshire CC	17.5	17.3	16.6	16.0	-3.68%	£932	£995	£1,099	£1,173	6.78%
Devon CC	17.0	16.8	16.7	16.7	0.29%	£919	£976	£1,048	£1,102	5.11%
Dorset CC	17.1	17.0	16.3	16.4	0.51%	£872	£919	£995	£1,066	7.11%
Durham CC	16.7	16.8	17.3	16.5	-4.61%	£899	£964	£1,032	£1,133	9.73%
East Sussex CC	17.0	17.1	17.1	16.7	-2.13%	£939	£994	£1,031	£1,098	6.49%
Essex CC	17.2	17.0	16.8	16.6	-1.49%	£945	£1,013	£1,105	£1,169	5.73%
Gloucestershire CC	17.3	17.2	17.0	16.0	-6.00%	£929	£985	£1,055	£1,126	6.74%
Hampshire CC	16.8	17.1	17.0	16.7	-1.52%	£951	£1,001	£1,053	£1,132	7.57%
Hereford & Worcester CC	18.0	17.2	16.8	16.7	-0.18%	£887	£964	£1,020	£1,065	4.46%
Hertfordshire CC	15.9	15.8	15.6	15.0	-4.03%	£961	£1,016	£1,093	£1,176	7.55%
Humberside CC	16.8	16.2	16.6	16.6	0.08%	£920	£968	£1,023	£1,081	5.58%
Isle of Wight CC	18.3	18.1	18.1	17.1	-5.88%	£911	£960	£1,050	£1,114	6.06%
Kent CC	17.1	16.8	16.7			£904	£996	£1,038	£1,093	5.32%
Lancashire CC	16.9	16.7	16.5	16.0	-3.45%	£951	£1,019	£1,093	£1,174	7.49%
Leicestershire CC	16.2	16.0	15.4	15.2	-1.22%	£968	£1,063	£1,157	£1,242	7.35%
Lincolnshire CC	17.5	17.1	16.8	16.3	-2.66%	£977	£1,058	£1,103	£1,194	8.21%
Norfolk CC	16.8	16.6	16.3	16.1	-1.69%	£958	£1,008	£1,102	£1,162	5.43%
Northamptonshire CC	16.7	16.3	16.0	16.4	2.48%	£923	£980	£1,069	£1,112	4.00%
Northumberland CC	17.5	17.2	16.8	16.4	-2.61%	£888	£943	£1,036	£1,092	5.47%
North Yorkshire CC	16.6	16.4	16.6	16.4	-1.32%	£967	£1,046	£1,100	£1,110	0.87%
Nottinghamshire CC	15.6	15.3	15.0	14.9	-0.51%	£1,007	£1,077	£1,180	£1,258	6.56%
Oxfordshire CC	17.2	17.3	16.8	16.5	-2.14%	£921	£972	£1,037	£1,098	5.88%
Shropshire CC	16.2	16.0	15.9	15.8	-0.94%	£949	£1,013	£1,109	£1,174	5.85%
Somerset CC	17.5	17.7	17.7	16.9	-4.42%	£913	£967	£1,025	£1,102	7.59%
Staffordshire CC	16.4	16.4	16.6	16.5	-0.54%	£956	£1,037	£1,102	£1,155	4.76%
Suffolk CC	17.3	16.8	17.0	16.8	-1.50%	£873	£941	£1,004	£1,063	5.88%
Surrey CC	16.3	16.5	16.0	15.7	-1.97%	£1,029	£1,085	£1,183	£1,258	6.31%
Warwickshire CC	17.2	17.1	16.8	16.6	-1.09%	£930	£981	£1,069	£1,113	4.08%
West Sussex CC	17.2	16.9	16.7	16.6	-0.84%	£879	£936	£989	£1,020	3.08%
Wiltshire CC	17.0	17.3	17.1	16.9	-1.17%	£926	£994	£1,056	£1,126	6.58%

In recent years new techniques have been developed under the umbrella of 'Exploratory Data Analysis' (EDA). These are attempts to simplify some of the more traditional sorting procedures and diagrams to improve on the information they portray. The name 'exploratory' is used as such methods are particularly useful when dealing with data that you are looking at for the first time. They will be used in this chapter alongside the more traditional methods.

Few people nowadays undertake statistical research of any length with pencil and paper. Many computer packages have been produced to carry out analysis, you may have already met these in other subject areas. Many packages are limited to particular techniques and can be restrictive in some kinds of research. The most useful are the spreadsheet style packages. These not only allow all the normal spreadsheet functions of manipulating rows and columns but allow a wide range of analysis to be performed.

Find out from your computer department what they have available and try to follow through this work with a computer. Most of the diagrams and tables in this chapter could be produced using such packages.

16.1 Stem and leaf diagrams

When you are faced with a large amount of data such as the figures on school spending it is difficult to answer questions like 'What are the highest and lowest spending authorities?' 'Do they all spend pretty much the same or are there some particularly high or low spenders?'

With small amounts of data it is usually sufficient to simply write the values in order, but with more than just a handful of data items some kind of grouping is required. The **frequency table** simply shows the number of items in each category or group. For example the pupil teacher ratios for London Authorities in 1986/7 would be listed as shown opposite. Note that the class limits are given to the same degree of accuracy as the original data, this avoids any overlap or confusion. Open ended groups, e.g. '22.0 and over', are sometimes used where there are a lot of scattered items at one end. Open ended groups can cause inaccuracies and difficulties in later work, as do tables where groups are of uneven size, and these are best avoided. Tables are the backbone of official statistics publications where a number of variations can be shown at one time.

A modern alternative to the frequency table is the **stem and leaf diagram**. The stem is chosen in the same manner as the groups in a frequency table, so for the above data these would be the whole numbers. The stem and leaf of the above data would then take the form opposite.

Pupil Teacher Ratios	Frequency
11.0 - 11.9	2
12.0 - 12.9	3
13.0 - 13.9	1
14.0 - 14.9	6
15.0 - 15.9	4
16.0 - 16.9	5
17.0 - 17.9	0

Stem	leaf
11	29
12	158
13	6
14	355679
15	3478
16	12478
17	

The advantages of these over conventional tables are:

- (i) the actual values are recorded in the table and can be used in later work;
- (ii) the length of row gives a good visual indication of the spread of results.

With figures where the data range covers more than one significant figure some rounding needs to be used. The 'Unit Costs' for London in 1986/7 would be as shown opposite.

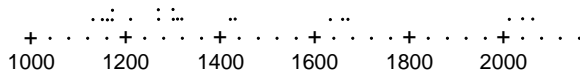
11	35677
12	177
13	0012
14	23
15	
16	367
17	
18	
19	
20	146

When you wish to compare two sets of data, e.g. London and Counties, you can draw a 'back-to-back' stem and leaf as shown opposite.

Pupil Teacher Ratio 1986/7

London		Counties
92	11	
158	12	
6	13	
976553	14	9
8743	15	02467889
87421	16	0001223444445555666677778899
	17	1

Another useful EDA technique is the **dotplot** where each item is stored as a dot on a continuous scale. The data for London Unit Costs would be shown as below.



Whilst they usefully show the spread of data it is difficult to read off exact values.

Exercise 16A

In Questions 1 and 2 use a computer if possible.

1. Draw frequency tables for the Pupil Teacher Ratios in the Metropolitan Districts and Counties in 1986/7. Comment on any differences in the pattern of pupil/teacher ratios in the three types of area.
2. Draw stem and leaf diagrams for the Unit Costs in 1986/7 for the Metropolitan Districts and Counties. Use these along with the one in the text for London to comment on the pattern of spending in the three types of area.
3. Use the data from the table below 'Notifiable offences...' to answer the following.
 - (a) Add together the figures for the three areas in each of the three years. Express these as a percentage of the total crimes in that year and put the results in a table.
 - (b) What does your table tell you about the pattern of crime over that period?
 - (c) Calculate the % of "violence against the person" crimes for Northern Ireland only in 1985. Does this support the view that Northern Ireland is a more violent place to live?

Notifiable offences recorded by the police: by type of offence

England & Wales, Scotland, and Northern Ireland

Thousands

	England & Wales			Scotland			Northern Ireland		
	1971	1984	1985	1971	1984	1985	1971	1984	1985
Notifiable offences recorded									
Violence against the person	47.0	114.2	121.7	5.0	9.2	10.7	1.4	3.4	3.5
Sexual offences	23.6	20.2	21.5	2.6	2.4	2.6	0.2	0.5	0.7
Burglary	451.5	897.5	871.3	59.2	112.1	100.7	10.6	22.4	20.2
Robbery	7.5	24.9	27.5	2.3	4.5	4.4	0.6	2.0	1.8
Theft and handling stolen goods	1003.7	1808.0	1884.1	104.6	215.8	208.9	8.6	28.7	29.5
Fraud and forgery	99.8	126.1	134.8	9.4	29.1	30.6	1.5	3.1	3.7
Criminal damage	27.0	497.8	539.0	22.0	79.0	79.5	7.4	4.4	3.2
Other notifiable offences	5.6	10.4	12.2	5.9	22.7	24.6	0.5	2.2	2.0
Total notifiable offences	1665.7	3449.1	3612.1	211.0	474.8	462.0	30.8	66.7	64.6

Activity 1

The worksheet on 'World Data' shown here gives various facts about countries from four key world regions. Use this as a piece of ongoing research through this chapter. If you have computer facilities you will be able to work individually otherwise split into groups to do different regions. By drawing frequency charts/stem and leaf diagrams compare life expectancy, GNP, literacy, etc. within the regions. How do the underdeveloped regions compare with Western Europe?

World Data	Populn. - Population in 1984 in millions	Life - average lifespan (life expectancy)
	Area - in thousands of km ²	Doct - the ratio of patients to doctor
	GNP - Gross National Product, roughly speaking the annual wealth (US\$) produced by a country divided by the number of inhabitants.	Read - adult literacy, i.e. the % of adults able to read.
	* Denotes data not available.	

African States

Row	State	Populn.	Area	GNP	Life	Doct	Read
1	Egypt	45.9	1001	720	60	800	44
2	Nigeria	96.5	924	730	50	10540	34
3	Zimbabwe	8.1	391	760	57	6650	69
4	Camaroon	9.9	475	800	54	*	*
5	Botswana	1.0	600	960	58	9250	*
6	Tunisia	7.0	164	1270	62	3620	62
7	Libya	3.5	1760	8520	59	660	*
8	Ethiopia	42.2	1222	110	44	88120	15
9	Mali	7.3	1240	140	46	25380	10
10	Zaire	29.7	2345	140	51	*	55
11	Malawi	6.8	118	180	45	52960	25
12	Niger	6.2	1267	190	43	*	10
13	Tanzania	21.5	945	210	52	*	79
14	Uganda	15.0	236	230	51	22180	52
15	C.A.R.	2.5	623	260	49	23090	33
16	Somalia	5.2	638	260	46	15630	60
17	Benin	3.9	113	270	49	16980	28
18	Rwanda	5.8	26	280	47	10260	50
19	Kenya	19.6	583	310	54	7540	47
20	Sierra L.	3.7	72	310	38	17670	15
21	Guinea	5.9	246	330	38	*	20
22	Ghana	12.3	239	350	53	6760	*
23	Sudan	21.3	2506	360	48	9070	32
24	Senegal	6.4	196	380	46	13060	10
25	Chad	4.9	1284	*	44	*	15
26	Mozambique	13.4	802	*	46	33340	33
27	Liberia	2.1	111	470	50	8550	25
28	Zambia	6.4	753	470	52	7110	44
29	Lesotho	1.5	30	530	54	*	52
30	Ivory Co.	9.9	322	610	52	*	35

South & Central American States (cont'd)

Row	State	Populn.	Area	GNP	Life	Doct	Read
11	Chile	11.8	757	1700	70	950	*
12	Brazil	132.6	8512	1720	64	1200	76
13	Panama	2.1	77	1980	71	1010	85
14	Uruguay	3.0	176	1980	73	510	94
15	Mexico	76.8	1973	2040	66	1140	83
16	Argentina	30.1	2767	2230	70	*	93
17	Venezuela	16.8	912	3410	69	930	82
18	Trinidad	1.2	5	7150	69	1390	95
19	Haiti	5.4	28	320	55	*	23
20	Bolivia	6.2	1099	540	53	1950	63
21	Honduras	4.2	112	700	61	*	60
22	El Salvador	5.4	21	710	65	3220	62

West European States

Row	State	Populn.	Area	GNP	Life	Doct	Read
1	Greece	9.9	132	3770	75	390	*
2	Portugal	10.2	92	1970	74	450	78
3	Spain	38.7	505	4440	77	360	*
4	Ireland	3.5	70	4970	73	780	98
5	Italy	57.0	301	6420	77	750	98
6	UK	56.4	245	8570	74	680	99
7	Belgium	9.9	31	8610	75	380	99
8	Austria	7.6	84	9140	73	580	99
9	Holland	14.4	41	9520	77	480	99
10	France	54.9	547	9760	77	460	99
11	Finland	4.9	337	10770	75	460	100
12	Germany	61.2	249	11130	75	420	99
13	Denmark	5.1	43	11170	75	420	99
14	Sweden	8.3	450	11860	77	410	99
15	Norway	4.1	324	13940	77	460	99
16	Switzerland	6.4	41	16330	77	390	99

South & Central American States

Row	State	Populn.	Area	GNP	Life	Doct	Read
1	Nicaragua	3.2	130	860	60	2290	90
2	Dominican	6.1	49	970	64	1390	67
3	Peru	18.2	1285	1000	59	*	80
4	Ecuador	9.1	284	1150	65	*	81
5	Guatamala	7.7	109	1160	60	*	*
6	Costa Rica	2.5	51	1190	73	*	90
7	Jamaica	2.2	11	1150	73	*	90
8	Paraguay	3.3	407	1240	66	1310	84
9	Colombia	28.4	1139	1390	65	*	81
10	Cuba	9.9	115	*	75	600	95

East European States

Row	State	Populn.	Area	GNP	Life	Doct	Read
1	Yugoslavia	23.0	256	2120	69	670	85
2	Hungary	10.7	93	2100	70	320	99
3	Poland	36.9	313	2100	71	550	98
4	Albania	2.9	29	*	70	*	*
5	Bulgaria	9.0	111	*	71	400	*
6	Czechos.	15.5	128	*	70	350	*
7	Romania	22.7	238	*	71	650	98

16.2 Typical data

There are two basic questions that you need to ask about data:

- What is a typical value for a particular group of data?
- How closely does all the data conform to this typical value?

There are many different values that can be taken as typical, usually referred to as **measures of central tendency**. Each of these has an associated **measure of spread** which indicates how widespread the data is. The most useful are:

- mode and range;
- median and percentiles;
- mean and standard deviation.

Mode and range

These are the simplest and subsequently crudest measures available. The **mode** is defined as the most frequently occurring item or group of items.

The data opposite is part of an analysis of accidents on which insurance claims were submitted. The three sets of data illustrate the different types of data that occur.

- Qualitative** - involves no meaningful numbers, e.g. 'Use of Vehicle'.
- Discrete** - numerical data with only a fixed number of options e.g. 'Casualties'.
- Continuous** - numerical data with infinite options, the only restriction being accuracy of measurement, e.g. 'Cubic Capacity'.

What is the modal value in each case?

With cubic capacity the modal group is '1501 -2000'. One problem with this is that the value could depend on the particular groupings. Because of this, attempts to find a precise mode are pointless.

With 'Use of a Vehicle' you can say that the most common form of use was 'Pleasure'. The mode is in fact the only value you can use with qualitative data.

With 'Casualties' the mode is clearly 0. This measure however ignores the rest of the data.

Another problem with the mode occurs when two or more categories have equal highest frequencies. (See London Pupil/Teacher Ratios in last unit.)

Use of a Vehicle

Pleasure	4550
Journey to/from work	1217
Business	1244
Motor Trade	40
Hire or reward	92
<hr/>	
Total	7143
(Unknown	825)

Casualties

0	7383
1	448
2	103
3	23
4	10
5	0
6	1
<hr/>	
Total	7968

Cubic Capacity

Up to 1000	706
1001-1500	2558
1501-2000	3011
2001-2500	269
2501-3000	162
Over 3001	97
<hr/>	
Total	6803
(Unknown	1165)

The **range**, which is the **difference** between the highest and lowest value, is difficult to calculate from the data as presented (it has no meaning in the qualitative case) but could simply be found from the raw data. In neither of the two quantitative cases is it particularly informative. If you look at the 'Life Expectancy' in the data met earlier however, the highest and lowest values are of great interest. Find out what these values are.

The main difficulty with the range however is that it only needs one extreme value at each end to totally distort the data.

Median and percentiles

The title 'measures of central tendency' suggests that you are looking for a value in the middle. This is exactly what the median does. If 5 people are put in order of height, then the median height would be that of the middle person.

Note that the middle item of five is the third. In general if there are n items of data then the median is found at the $(n + 1)/2$ th value. Clearly with an even number of items this will fall exactly half way between two values.

Medians can easily be found from stem and leaf diagrams. For example using the Life Expectancy figures for African States, you obtain the plot opposite, from which, since $n=30$, you can readily read off the median value.

3	88	
4		
4	3	
4	445	
4	66667	
4	899	
5	(00)11	Median = 50
5	2223	
5	444	
5	7	
5	89	
6	0	
6	2	

With data in groups you can only obtain an estimate (see why stem and leaf are better!).

Using the Literacy figures for Africa you get a frequency table which is shown opposite.

Literacy (%)	Freq.
10 - 19	6
20 - 29	4
30 - 39	5
40 - 49	3
50 - 59	4
60 - 70	3
70+	1
Total	26

The median is the $(26+1)/2 = 13.5$ th item of data.

This must be $3.5/5 = 0.7$ of the way up the 30-39 group.

The median therefore is $29.5 + (0.7 \times 10) = 36.5$

Note that 29.5 is the lowest rounded value which could be in the 30-39 group and is called the **lower class boundary**.

This method is called **linear interpolation**, which you may have met elsewhere, and assumes that the frequency goes up in a straight line. The result is therefore only an estimate.

Activity 2

- (a) Work out the median life expectancies for Central & South America, Western Europe and Eastern Europe using stem and leaf diagrams. Comment on your results.
- (b) The information opposite gives details of property deals carried out in a particular year. Use the number of deals (No. 000s) column to estimate the median price of property deals that year.

**Residential Property Deals
England and Wales 1989**

Price	No. 000s	%	Value (£m)	%
Under £10,000	124	8.4	472	0.6
10-20,000	175	11.9	2,622	3.2
20-30,000	226	15.4	5,706	6.9
30-50,000	310	21.2	12,361	15.0
50-100,000	470	32.0	32,628	39.6
100-250,000	142	9.7	19,860	24.1
over £250,000	21	1.4	8,718	10.6
Total	1,468	100	82,367	100

The median divides the data into two halves, i.e. it has 50% of the data above it and 50% below. Using the same idea you could in fact divide the data into any fraction you like. One of the most useful is to divide the data into quarters. The data for Life Expectancy in Africa is shown opposite.

The first quarter is given by the $(30 + 1) / 4 = 7.75$ th item. Since the seventh and eighth item are both 46 this is 46. This is called the **lower quartile**.

The third quarter or **upper quartile** is given by $3 \times 7.75 = 23.25$ th item, again with repeated values is 54.

In the table opposite, the running total or cumulative frequencies from each end have been shown to help you find these values.

A commonly used measure is the **interquartile range**, the difference between the two quartiles i.e. $54 - 46 = 8$. The **semi-interquartile range** is half this value.

The interquartile range gives the range within which the middle half of the data lie. In conjunction with the median you can see that the data is fairly evenly spread out about the middle.

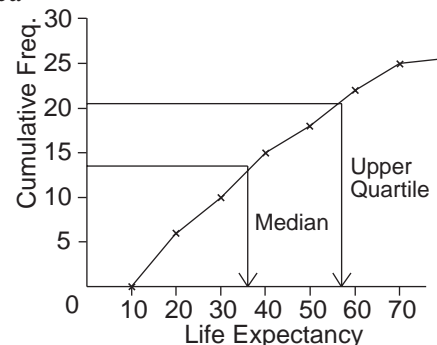
To find these values from a frequency table you could use a linear interpolation method, but this can be difficult and the assumption of linearity between values is not always a good one. A commonly used alternative is to plot a **cumulative frequency curve**. You first need to find the cumulative frequency or running total. Look at the Literacy data for Africa shown opposite.

Note that the cumulative frequency values are related to the **upper class boundaries**, e.g. 10 in the second row tells you that 10 states had a literacy rate of 29.5% or less. You therefore plot a graph of cumulative frequencies against upper class boundaries.

The median and quartiles can be read off at the 13.5th, 6.75th and 20.25th values respectively. Although the results depend on graphical accuracy the fact that a curve is used rather than a straight line gives a more realistic result.

Frequency	Stem	leaf
2	3	88
2	4	
3	4	3
6	4	445
11	4	66667
14	4	899
(4)	5	0011
12	5	2223
8	5	444
5	5	7
4	5	89
2	6	0
1	6	2

Literacy (%)	Freq.	Cumulative Freq.
10 - 19	6	6
20 - 29	4	10
30 - 39	5	15
40 - 49	3	18
50 - 59	4	22
60 - 70	3	25
70+	1	26



This same idea can be used for any other percentiles but the only other ones in use are the 5% and 95% values.

The median and quartiles are commonly used measures. The one criticism made of them is that they do not consider all the data and ignore the tail ends. This can however be an advantage in that they are not prone to distortion by extremes.

Activity 3

In this activity use a method appropriate to the data.

- (a) Using the World Data worksheet find the median and interquartile range of the patient/doctor ratios in each of the four areas. Comment on your results.
- (b) Find the median and quartiles for
 - (i) Cubic Capacity
 - (ii) Casualties
 from the 'Analysis of Insurance Claims' data earlier in this unit.
- (c) The data below shows the population of the UK in millions over this century and predicted values for the future. Find the median and interquartile range for:
 - (i) 1901 (ii)1941 (iii)1991 (iv)2015.

Comment on your results in particular with reference to what the possible implications on future Government planning might be.

Age and sex of the population

United Kingdom

Millions

	0-4	5-14	15-29	30-44	45-59	60-64	65-74	75-84	85+	All ages
Census enumerated										
1901	4.4	8.0	10.8	7.5	4.6	1.1	1.3		0.5	38.2
1911	4.5	8.4	11.2	8.9	5.6	1.2	1.6		0.6	42.1
1921	3.9	8.4	11.2	9.3	7.0	1.5	1.9		0.7	44.0
1931	3.5	7.6	11.8	9.7	8.0	1.9	2.5		1.0	46.1
Mid-year estimates										
1941	3.4	6.8	9.2	10.3	8.5	2.3	3.2		1.3	44.9
1951	4.3	7.0	10.2	11.2	9.6	2.4	3.7		1.8	50.3
1961	4.3	8.1	10.3	10.5	10.6	2.8	4.0	1.9	0.3	52.8
1971	4.5	8.9	11.8	9.8	10.2	3.2	4.8	2.2	0.5	55.9
1976	3.7	9.2	12.4	10.0	9.8	3.1	5.1	2.3	0.5	56.2
1981	3.5	8.1	12.8	11.0	9.5	2.9	5.2	2.7	0.6	56.4
1983	3.6	7.6	13.1	11.1	9.4	3.2	5.0	2.8	0.6	56.3
1984	3.6	7.4	13.3	11.2	9.3	3.3	4.8	2.9	0.7	56.5
1985										
Males	1.9	3.7	6.8	5.7	4.6	1.5	2.2	1.0	0.2	27.6
Females	1.8	3.5	6.6	5.6	4.7	1.7	2.8	1.9	0.5	29.0
Total	3.6	7.3	13.4	11.3	9.3	3.1	4.9	2.9	0.7	56.6
Projections										
1987		10.7	13.5	11.6	9.2		8.0		3.8	56.9
1991		11.0	12.9	12.1	9.5		7.9		4.0	57.5
1996		11.7	11.6	12.6	10.5		7.7		4.2	58.3
2001		12.0	10.8	13.2	11.0		7.6		4.4	59.0
2006		11.7	11.0	12.6	11.6		7.9		4.5	59.3
2011		11.1	11.7	11.3	12.1		8.8		4.5	59.4
2015		10.8	12.0	10.6	12.6		9.1		4.5	59.6

Mean and standard deviation

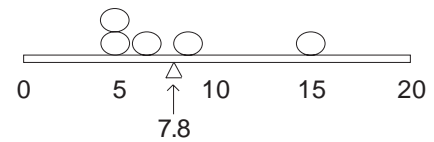
So far in this text the word 'average' has been avoided. In fact all measures of central tendency are averages, though if asked to find the average most people would calculate the mean. The mean is found by adding together all the values in the data and dividing by the number of items. In notation form you write this:

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

So the mean of 5, 5, 6, 8 and 15 is $\frac{(5+5+6+8+15)}{5} = 7.8$

But what does the mean mean?

If you imagine the numbers above being balanced on a 20 cm strip, to balance the strip you would need to put the pivot at a point 7.8 cm from the end.



The mean is mathematically equivalent to the centre of gravity. The mean then 'balances' all of the data available. Although it does use all the data it is prone, as in this case, to distortion by odd extreme values.

Data presented in the form of a stem and leaf diagram can easily be used to calculate the mean as the original values are still recorded. When data is stored in a frequency table a quicker method can be used. The Insurance data for the number of casualties is shown in the table opposite.

Casualties	Freq.
0	7383
1	448
2	103
3	23
4	10
5	0
6	1
Total	7968

This table tells us that there were 7383 accidents with no casualties, 448 with one, etc. Therefore the total of all casualties can be found by:

$$\begin{aligned} \Sigma x &= (0 \times 7383) + (1 \times 448) + (2 \times 103) + (3 \times 23) \\ &\quad + (4 \times 10) + (5 \times 0) + (6 \times 1) \\ &= 769 \end{aligned}$$

So $\bar{x} = \frac{769}{7968} = 0.097$

As with the median when data is given in a grouped table you can only estimate the mean. Here you must repeat the above procedure but use the group mid-mark as the x value. Using 'Cubic Capacity' data from the Insurance data shown opposite:

$$\begin{aligned} \Sigma x &= (750.5 \times 706) + (1250.5 \times 2558) + \dots \\ &= 10\,365\,651.5 \end{aligned}$$

So $\bar{x} = \frac{10\,365\,651.5}{6803} = 1524$

Cubic Capacity	Mid value	Freq.
Up to 1000	750.5	706
1001 - 1500	1250.5	2558
1501 - 2000	1750.5	3011
2001 - 2500	2250.5	269
2501 - 3000	2750.5	162
Over 3000	3250.5	97
Total		6803

Note that the final answer has been rounded to a whole number in recognition of accuracy.

What are the assumptions that have been made and how true are they likely to be?

What other possible sources of error are there? (Hint: How were the mid-marks of the open ended groups chosen? Did it matter?)

Activity 4

- (a) Find the mean life expectancy in the four areas in the World Data information, using the stem and leaf diagrams drawn earlier.
- (b) From the 'Age and Sex of the Population' table in the last activity find the mean age of the population in the years
(i) 1901 (ii) 1941 (iii) 1991 (iv) 2015

Compare these values with the medians calculated in the last activity. Under what circumstances would you expect these two values to be approximately the same?

To accompany the mean it seems appropriate to have a measure of spread which incorporates all of the data. These are the life expectancies of the Middle Eastern states:

Yemen Arab Rep.	45	Yemen PDR	47
Jordan	64	Syria	63
Israel	75	Iran	61
Iraq	60	Oman	53
Libya	59	Saudi Arabia	62
Kuwait	72	U.A.E.	72

The mean is 61.1. To look at the way these differ from the mean you need to look at the differences between each value and the mean, i.e.

-16.1	-14.1
+2.9	+1.9
+13.9	-0.1
-1.1	-8.1
-2.1	+0.9
+10.9	+10.9

Add these differences up and you should not be too surprised to find this gives zero. In order to avoid this the differences are squared and to give an overall measure you sum these,

$$\Sigma(x - \bar{x})^2 = 16.1^2 + 14.1^2 + \dots + 10.9^2 = 972.92$$

As it stands the figure will increase with more data so it needs to be averaged out. You also need to compensate for the

squaring so you square root the final answer. This measure is the **standard deviation**. In this case

$$s = \sqrt{\left(\frac{972.92}{12}\right)} = 9.00$$

In general, the standard deviation is given by:

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

But what does the standard deviation stand for?

One of the difficulties with the standard deviation is that it is difficult to explain to a layperson what it represents. It is at its most useful in comparing different sets of figures, i.e. is one set of data more spread out than another? A rule of thumb often used is that when the data follows a bell shaped pattern (i.e. most data is close to the mean) 95% of data lies within **two** standard deviations from the mean.

The formula above is a rather clumsy method for actually calculating the standard deviation. Another formula more commonly used is:

$$s = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

Also note that the **variance** is defined as the square of the standard deviation, s^2 .

This has the advantage that Σx^2 and Σx can be found cumulatively, i.e.

Value	Σx	Σx^2
45	45	2025
47	92	4234
64	156	8330
63	219	1229
75	294	17924
.		
.		
.		
72	733	45747

$$\bar{x} = \frac{733}{12} = 61.1$$

$$s = \sqrt{\frac{45747}{12} - (61.1)^2} = 9.00$$

Activity 5

Noting that $\bar{x} = \frac{\sum x}{n}$, prove that $s^2 = \frac{\sum x^2}{n} - \bar{x}^2$.

If it was not for this alternative formula calculators would find this very difficult to calculate. As it is most modern calculators have the function to calculate this. Find out how to calculate the mean and standard deviation on your calculator.

Activity 6 Do personal stereos help you concentrate?

A simple mathematical puzzle is shown on the worksheet on the next page. The purpose of this experiment is to see whether people can complete the puzzle more quickly if they are using a personal stereo at the time. As a group design an experiment to test this hypothesis. Calculate the mean and standard deviation of the times subjects took to complete the test with and without headphones and use these to report your findings.

Exercise 16B

- Find the standard deviation of the Life Expectancies for the four regions in the World Data. What does this tell you about regional differences in these main areas?
- The Unemployment Rates in three regions in the UK in 1989 were listed as shown in the table opposite.
Find the mean and standard deviation for each region and comment on the regional differences in unemployment.
- Calculating the standard deviation from data in a frequency table works in the same way as the mean. $\sum x$ and $\sum x^2$ are found by multiplying the value (or mid mark in the case of grouped tables) by the frequencies. Some calculators do this automatically. Find the standard deviations of
 - Casualties;
 - Cubic Capacity as given in Insurance data earlier.
- The OPCS monitors the ages of mothers giving birth to children. The data for 1941, 1951, 1971 and 1989 are shown opposite.
Find the mean and standard deviation for each year and comment on the changes in pattern of the age at which women give birth.

East Anglia	%	North-West	%
Cambridge	5.8	Accrington	13.4
Gt. Yarmouth	9.7	Ashton U. Lyme	14.0
Ipswich	8.1	Birkenhead	18.6
Lowestoft	11.7	Blackburn	13.9
Norwich	9.1	Blackpool	11.2
Peterborough	11.8	Bolton	15.4
		Burnley	11.8
		Bury	13.2
		Chester	12.2
		Crewe	10.0
		Lancaster	11.8
		Leigh	15.3
		Liverpool	18.1
		Manchester	12.3
		Nelson	14.2
		Northwich	15.7
		Oldham	14.1
		Preston	11.9
		Rochdale	16.9
		Southport	15.7

Age of Mother	Thousands of Births			
	1941	1951	1971	1989
15 - 19	28.8	33.0	92.2	61.6
20 - 24	169.9	212.0	317.6	202.0
25 - 29	208.1	247.3	273.2	265.9
30 - 34	147.8	159.0	122.7	158.5
35 - 39	85.0	88.3	50.5	53.3
40 - 44	28.1	26.1	13.1	9.8
45 - 49	2.0	1.5	0.9	0.8

16.3 Illustrating results

So far all the techniques have been about simplifying the data with the main intention being your own understanding. When presenting your findings to others you will need to make them easy to understand and attractively presented. In this section look at various forms of diagrams which can help in this task.

Bar charts

These are the most commonly used form of diagram. Strictly speaking there are 3 types of bar chart, one for each type of data. The illustration below show these using the Insurance data.

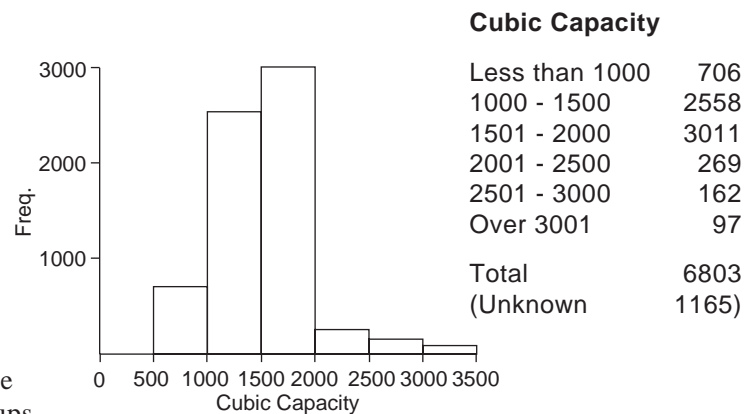
In practice frequency diagrams are not often used as they are not particularly eye-catching and it is more common to see a bar chart used where the numbers are treated as qualitative.

Histogram

Used only for quantitative continuous data.

A proper continuous mathematical scale must be used along the bottom. There are no gaps between bars as they are drawn up to class boundaries.

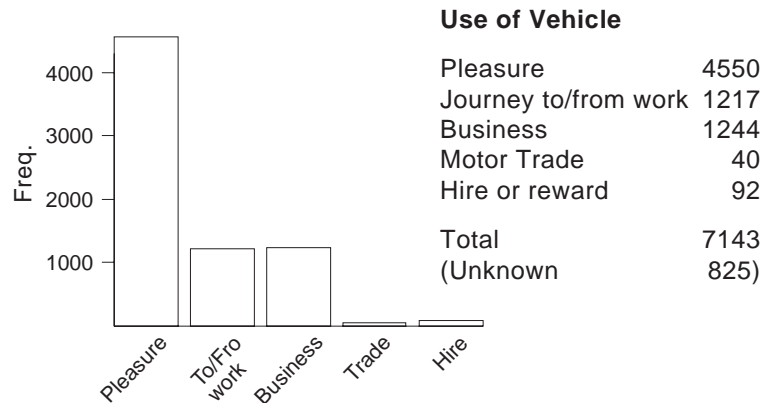
Problems arise when groups have uneven sizes so open-ended groups are treated as the same size.



Bar chart

Used for qualitative data.

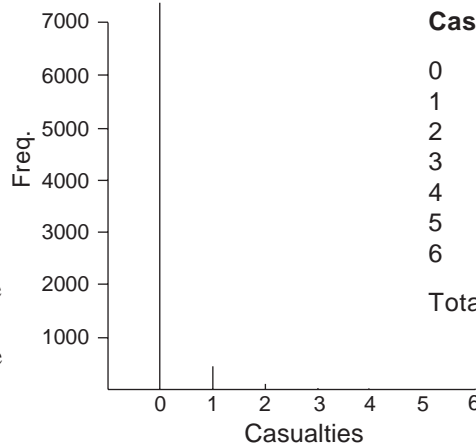
Gaps are left between bars. There is no scale along the horizontal axis.



Frequency diagram

Used for quantitative discrete data.

A proper scale is used along the horizontal axis. Lines are used instead of bars to emphasize the discreteness.



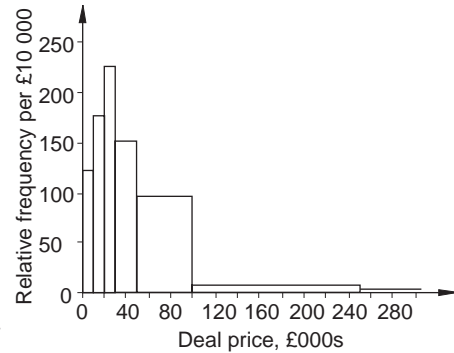
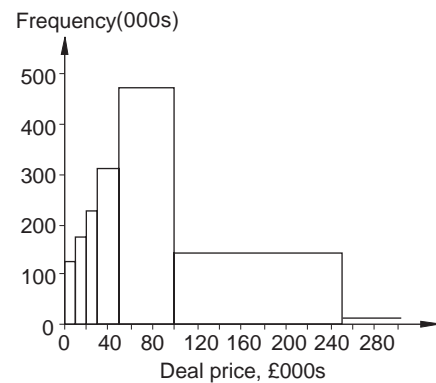
Casualties

0	7383
1	448
2	103
3	23
4	10
5	0
6	1
Total	7968

Care needs to be taken where groups are of uneven size. Consider the data on 'Residential Property Deals' used in Activity 2. The histogram is shown opposite.

Although there were only 1.5 times as many deals in the 50-100 category compared to the 30-50 category, because your eye looks at areas, it appears that there are more than three times as many on the histogram. To compensate for this the frequencies can be adjusted to a standard group size so that the area represents the frequency. These are known as **relative frequencies**. i.e.

Price	Freq. per £10 000
Under 10 000	124
10 - 20 000	175
20 - 30 000	226
30 - 50 000	155
50 - 100 000	94
100 - 250 000	9.5
Over 250 000	1.4



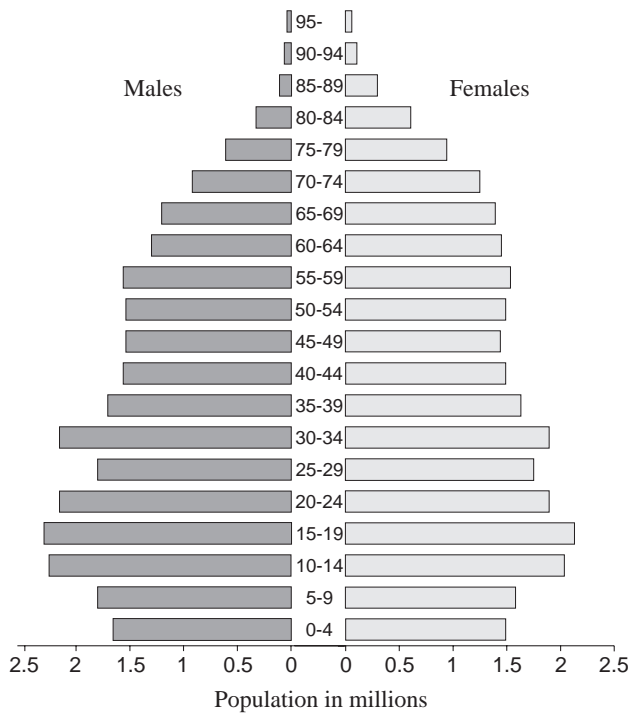
The resulting relative frequency histogram gives a much fairer view of the distribution of deals.

In some cases you may wish to show two sets of comparable data on one bar chart, this is called a **composite bar chart**. In the age distribution figures shown on the next page, three different methods have been used to show the male and female age group distributions.

Activity 7

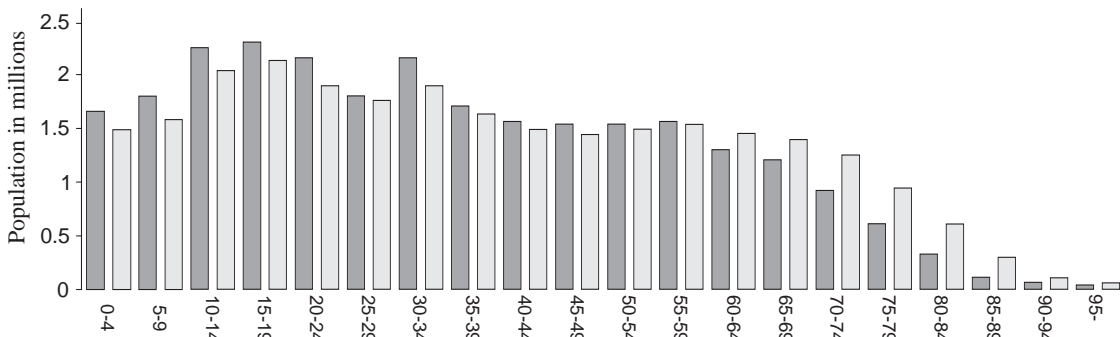
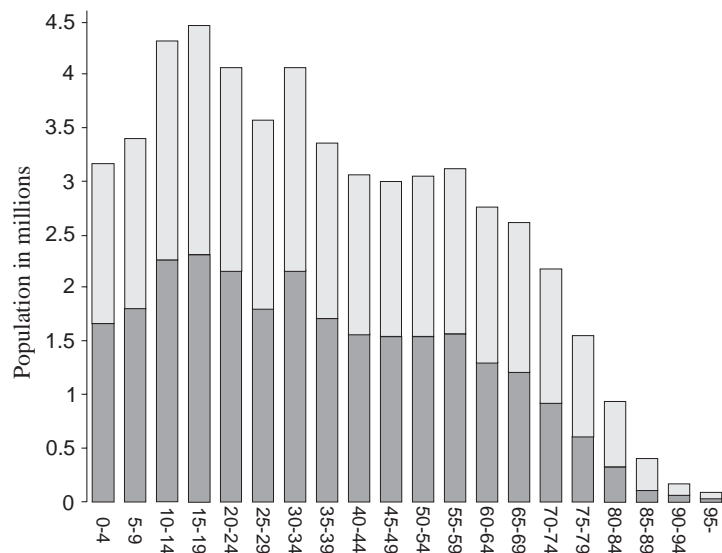
Draw histograms for 1941 and 1989 using the 'Ages at which mothers give birth' data in Question 4, Exercise 16B.

Age group distribution, Great Britain, 1981



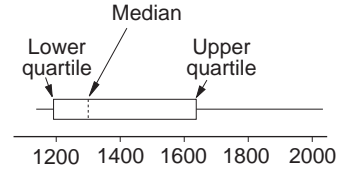
Bar charts are often used to show sets of comparable information side by side, as shown opposite. This is called a **composite bar chart**.

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

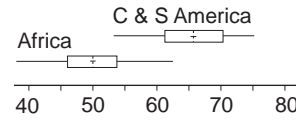


Box and whisker plot

This is an EDA technique that has become increasingly popular in recent years. It is drawn using the quartiles to make a box, with a line indicating the median. Lines or whiskers are then drawn to the extremities. The data for London Unit Costs used earlier would be as shown opposite.



They are particularly useful when you wish to compare more than one set of data on the same scale. For example you can easily compare the Life Expectancies of Africa with Central & South America as shown opposite.



Where there are odd extreme items or 'outliers' these can be shown as crosses outside the normal whisker.

Activity 7

- Draw box and whisker plots for the 'Age and sex of the population' data given earlier for 1901, 1941, 1991 and 2015 using the same scale.
- Draw box and whisker plots for the patient/doctor ratios for Africa and Central & South America on the same axis.

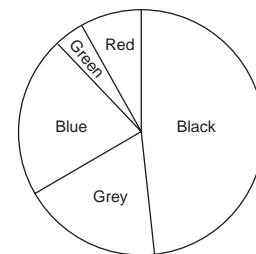
Pie charts

Pie charts are commonly used if you wish to show how a total is split into component parts. A school is to introduce a new school sweatshirt on a trial basis in a single colour. To determine the likely range of colours the school carries out a survey on a representative sample of students. The results are shown opposite.

Preferred colour	Boys	Girls
Black	48	35
Grey	19	4
Blue	21	8
Green	4	1
Red	8	16
Total	100	64

To form a pie chart it is easiest to calculate the percentage in each group first (this is simple for boys since there were 100). Pie chart scales are now available which are broken down in percentages. If however a protractor is used you will simply find percentages of 360° .

e.g. Boys voting Black need 48% of $360^\circ = 172.8^\circ$. The resulting pie chart is shown opposite.

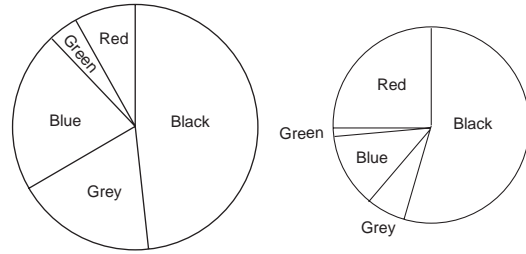


You might want to show the girls on a different pie chart, but since there are less of them a smaller chart should be drawn. Be careful however since, as you have already seen with bar charts, people perceive areas when they look at diagrams. You therefore need to draw the smaller pie chart in proportion to the **square root** of the radius of the larger.

For example if the boys' pie chart radius was 5 cm, area of girls' needs to be $64/100 = 0.64$ times smaller, so the radius needs to be $\sqrt{0.64} = 0.8$ times smaller. Hence

$$\text{radius for girls} = 0.8 \times 5 = 4 \text{ cm.}$$

The two pie charts together are shown opposite.



The following page shows some interesting ways that conventional diagrams have been made more eye-catching.

Other forms of diagram

Two forms of diagram are particularly useful when you have two sets of linked data. Where data is measured on a continuous scale, e.g. measurements taken over a period of time, a line graph is used as shown in the first diagram opposite.

Where the two values are measurements taken from individuals a scattergram can be used. For example suppose a scientist wishes to investigate whether reaction times are linked to pulse rates. The reaction time and pulse of some subjects are measured. Plotting these on a scattergram gives the diagram below.

In order to examine the link between two such variables you can mark the medians of the two variables on the graph. By examining the number of points in each quadrant an idea of the link or **correlation** between the two variables can be found.

In this case the points are evenly distributed so there appears to be little correlation between pulse and reaction time.

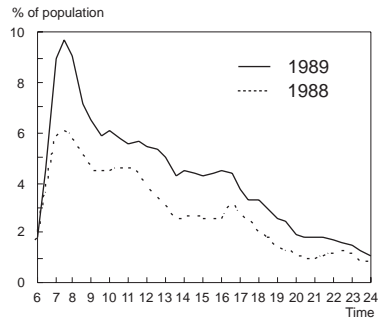
Deceptive diagrams

The page after next shows various diagrams from the media.

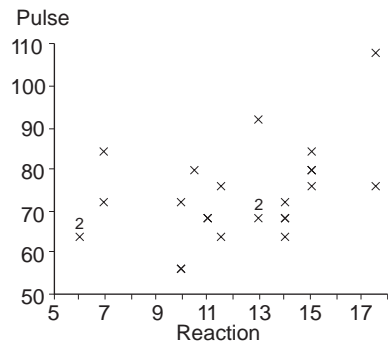
Discuss as a group why these perhaps might be misleading.

What could be done to make these more reasonable?

Average half-hour audience: All adults Monday to Friday averaged
Capital Radio Rating (on 9.7m adult pop)



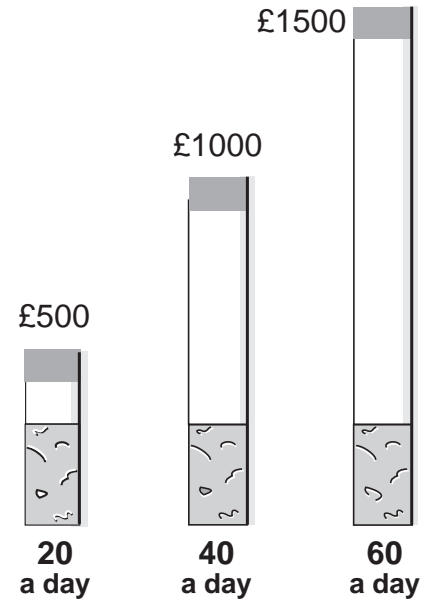
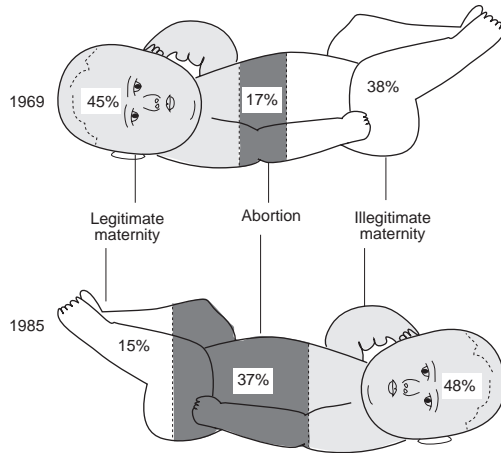
Capital Radio has gained significantly in audience across most of the day



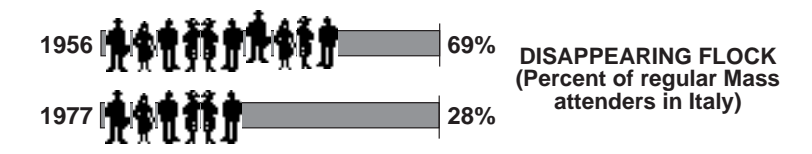
note: 2 indicates that there are two items at this point

Interesting Illustrations

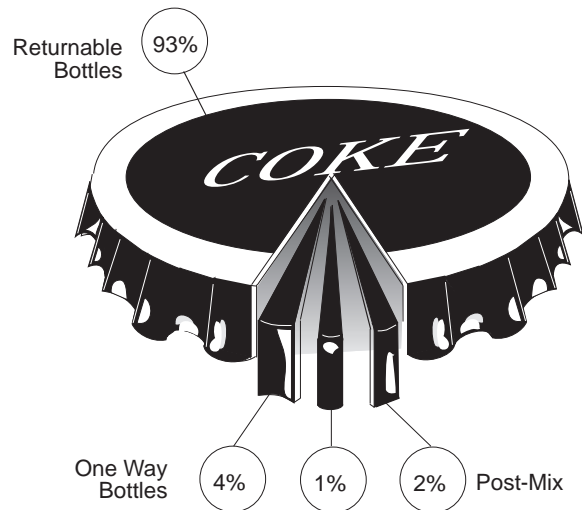
To have and to have not
Outcomes of extra-marital conception, England & Wales



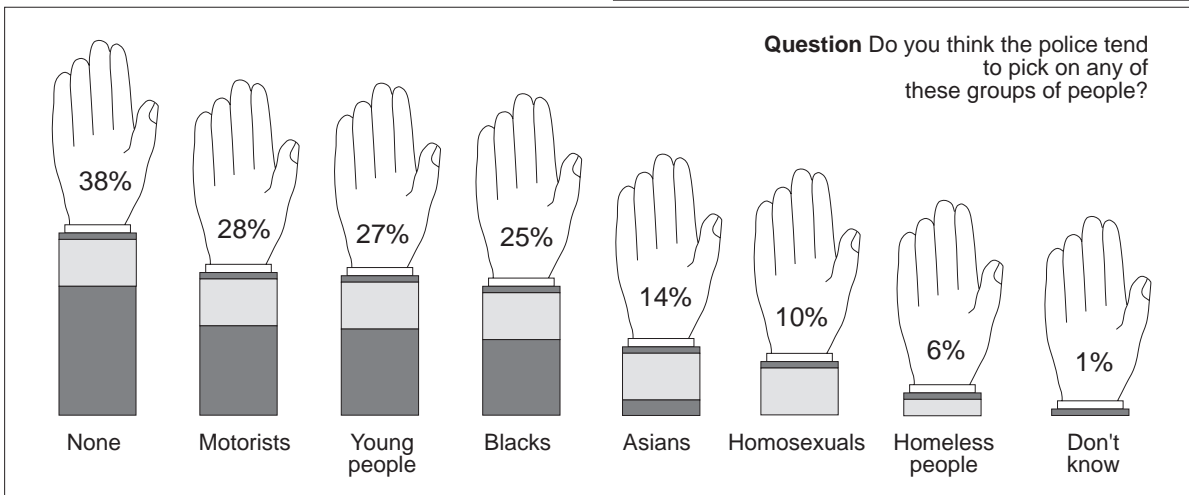
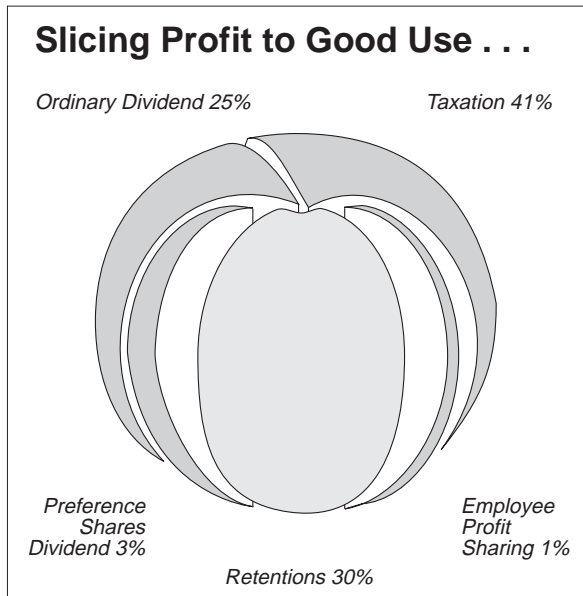
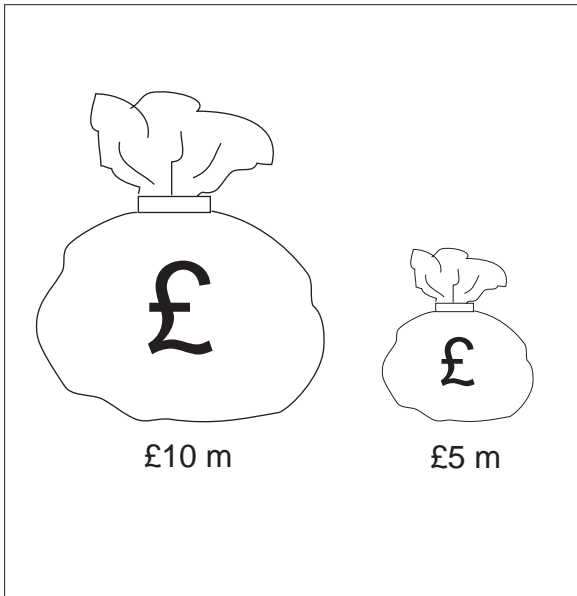
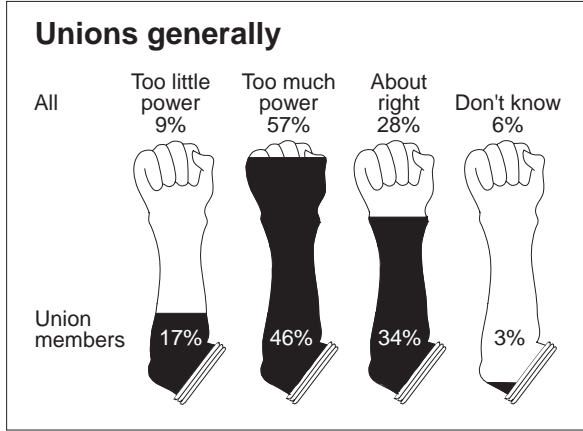
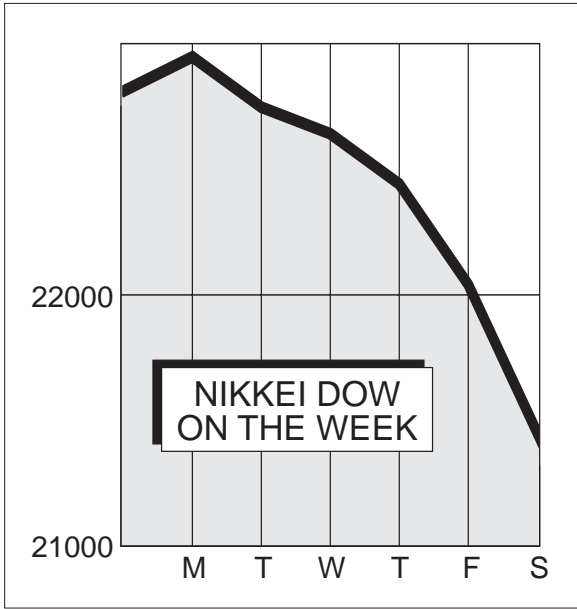
Based on £1.40 a pack



PACKAGES IN TOTAL MARKET - 1986



Deceptive diagrams



Exercise 16C

1. The masses (measured to the nearest g) of washers are recorded in the table. Draw a histogram to illustrate the data.

Mass (g)	0-2	3-5	6-11	12-14	15-17
Frequency	5	6	14	4	3

2. 100 people were asked to record how many television programmes they watched in a week. The results were as follows:

Number of programmes	0-	10-	18-	30-	35-	45-	50-	60-
Number of viewers	3	16	36	12	12	9	3	0

Draw a histogram to illustrate the data.

3. The table shows the sales, in millions of dollars, of a company in two successive years.

Year	Africa	America	Asia	Europe
1972	8.4	12.2	15.6	23.8
1973	5.5	6.7	13.2	19.6

Draw two pie charts which allow the total annual sales to be compared.

4. Five companies form a group. The sales of each company during the year ending 5th April, 1988, are shown in the table below.

Company	A	B	C	D	E
Sales (in £1000's)	55	130	20	35	60

Draw a pie chart of radius 5 cm to illustrate this information.

For the year ending 5th April, 1989, the total sales of the group increased by 20%, and this growth was maintained for the year ending 5th April, 1990.

If pie charts were drawn to compare the total sales for each of these years with the total sales for the year ending 5th April, 1988, what would be the radius of each of these pie charts?

If the sales of company E for the year ending 5th April, 1990, were again £60000, what would be the angle of the sector representing them?

16.4 Miscellaneous Exercises

1. 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 a cumulative frequency curve of these data.

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

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

Minimum	Lower quartile	Median	Upper quartile	Maximum
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.

2. 68 smokers were asked to record their consumption of cigarettes each day for several weeks. The table shown is based on the information obtained.

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

Illustrate these data by means of a bar chart.

3. The table below shows the marks, collected into groups, of 400 candidates in an examination. The maximum mark was 99.

Marks	0-9	10-19	20-29	30-39	40-49
No. of candidates	10	26	42	66	83
Marks	50-59	60-69	70-79	80-89	90-99
No. of Candidates	71	52	30	14	6

Compile a cumulative frequency table and draw the cumulative frequency curve.

Use your curve to estimate (a) the median, and (b) the 20th percentile.

If the minimum mark for Grade A was fixed at 74, estimate from your curve the percentage of candidates obtaining Grade A.

4. The frequency distribution given in the table refers to the heights, in cm, of 50 men corrected to the nearest 10 cm.

Height (cm)	140	150	160	170	180	190
Frequency	1	6	8	21	10	4

- (a) State the least possible height of the one man whose height is recorded in the table as 140 cm.
- (b) Draw on graph paper a histogram to illustrate the data from the table, drawing five columns, with the first column representing the seven shortest men. Label the axes carefully and explain clearly how frequency has been represented on your histogram.
- (c) Draw a cumulative frequency diagram on graph paper for the data given in the table. From your diagram, estimate the upper and lower quartiles, the median height and the interquartile range.

5. In an agricultural experiment the gains in mass, in kilograms, of 100 pigs during a certain period were recorded as follows:

Gains in mass (kilos)	5-9	10-14	15-19	20-24	25-29	30-34
Frequency	2	29	37	16	14	2

Construct a histogram and a cumulative frequency polygon of these data. Obtain (a) the median and the semi-interquartile range, (b) the mean and the standard deviation.

Which of these pairs of statistics do you consider more appropriate in this case, and why?

(AEB)

6. In a certain industry, the numbers of thousands of employees in 1970 were as shown in the tables below, by age groups.

Age last birthday	15-19	20-24	25-29	30-34	35-39
No. of thousands	66	65	56	50	42

Age last birthday	40-44	45-49	50-54	55-59	60-64
No. of thousands	37	35	30	24	22

Calculate the arithmetic mean, median, variance and standard deviation of the ages of employees in the industry.

Estimate the percentage of the employees whose ages lie within one standard deviation of the arithmetic mean. (AEB)

7. Two hundred and fifty Army recruits have the following heights.

Height (cm)	165-	170-	175-	180-	185-	190-195
No. of recruits	18	37	60	65	48	22

Plot the data in the form of a cumulative frequency curve. Use the curve to estimate (a) the median height, (b) the lower quartile height.

The tallest 40% of the recruits are to be formed into a special squad. Estimate (c) the median (d) the upper quartile of the heights of the members of this squad.

