2 DATA COLLECTION

Objectives

After studying this chapter you should

- understand what is meant by qualitative and quantitative data, discrete and continuous variables;
- understand what is meant by primary and secondary data;
- be able to use random number tables to find samples;
- be able to find random, systematic, stratified, quota and cluster samples.

2.0 Introduction

The current 'life expectancy' in the UK is about 71 years for men and 77 years for women. Apart from the obvious interest to individuals, figures such as these are of great concern to others: insurance companies, health organisations, social services, government departments such as the Treasury, leisure companies, etc. This kind of information is therefore collected by the government by means of the census and other surveys. A census is usually carried out every 10 years in this country and is compulsory by law to complete. Before modern technology was available it took several years to analyse the results, by which time much of the information was out of date anyway. In this chapter you will meet some of the techniques which might be used in such an analysis.

Consider the following two questions:

If you were told that your blood pressure was 140/90 would this be normal?

What is the normal weight of a seventeen-year-old in kilograms?

These are typical of the types of questions to be answered. You will need about 30 people for the first activity so you may have to involve other groups. You may not be able to carry out all the tests suggested in the following Activity but do try to obtain some of the equipment to do the more interesting and unusual ones - most of it is probably available in your school or institution if you ask. Do check that you know how to use the equipment properly.

Activity 1

In all tests your subjects should be allowed to test themselves. Keep all results confidential. Record, however, whether each participant was male or female. This Activity involves gathering data and you will be expected to analyse the data later in this chapter.

Measure the following:

(a) The **heights** in cm and **weights** in kg of everyone. Two metre rules taped to the wall and a book on the head works best for height. Weight is most easily measured by bathroom scales.

(b) **Eye** and **hair colour**. Make sure hair colour is natural! Decide on the categories before you start and stick to these.

(c) The number of occasions in the last month that individuals have undertaken hard **physical exercise** lasting 20 minutes or more, e.g. hockey, swimming, cycling to school.

(d) **Blood pressure**. Cheap digital blood pressure meters are available on the market and many Biology/P.E. Departments have these. Blood pressure is measured in two ways:

(i) **Systolic -** taken when the heart is beating and exerting maximum pressure.

(ii) **Diastolic -** taken when the heart is at rest and pressure is at minimum.

These are usually written together, e.g. 120/60. Take both these readings.

(e) **Pulse**. Digital blood pressure machines usually give this as well. If not, rather than use the traditional pulse point on the wrists, it is often easier to measure it with two fingers on the side of the throat. Count the beats in half a minute and double the result.

(f) **Breath power**. Blowmeters are commonly held by medical centres as they are useful in assessing asthmatics. Your Biology or PE Department may have one. By blowing into them the lung capacity can be measured.

(g) **Reaction times**. Reaction rulers are commercially available which can be used to measure your reactions. Alternatively, take a ruler marked in centimetres and hold it above the subject's slightly opened thumb and forefinger so that these are level with the zero on the ruler. When the ruler is dropped, the subject catches it. Measure the distance (in centimetres) the ruler drops before it is caught.

2.1 What sort of data?

The data on the next page give information on share prices on the London Stock Exchange. Data which you have collected yourself are called **primary** data, but data such as the Stock Market publish, where you are relying on someone else's measurements, are a **secondary** source.

Activity 2 Primary and secondary sources

Working in small groups discuss the following questions:

In each of these cases what possible sources of secondary data might be available? How might a survey be carried out? What are the advantages and disadvantages of using primary or secondary sources?

(a) The Health Education Council wants to know if a new campaign to stop young people starting smoking has been effective.

(b) A school canteen wants to see if there is a demand for healthier foods.

(c) A scientist wants to measure if a low fat diet improves athletic performance.

An even more important distinction between types of data is to what extent numbers are involved.

Qualitative data is where the actual measurements have no meaningful value, e.g. starting letter of Stock name, colour of a company logo. Be careful, as sometimes when recording data codes are used, e.g. 0 for male, 1 for female.

Quantitative data is where the data has a valid numerical value, e.g. share price. This category is further subdivided into

(a) **discrete -** where the data can only be one of a fixed number of numerical values, usually, but not necessarily, whole numbers, e.g. change.

(b) **continuous -** where the data can fall anywhere over a range and the scale is only restricted by the accuracy of measuring, e.g. yield (these are rounded to 1 d.p.).

Sometimes the division between discrete and continuous is a little indistinct. For example, share prices are strictly speaking discrete since they can only be to the nearest $\frac{1}{2}$p but because of the wide range of values it would be far more convenient to regard them as continuous.

London: The FT-SE 100

Stock	Price	Change	Yield
Abbey National	274	-3	4.6
Allied - Lyons	554	+5	4.5
Anglian Water	286	-8	6.8
Argyll Group	305	0	3.9
Arjo Wiggins Teape	252	+1	4.4
Asda Group	105	-8	6.1
Ass Brit Foods	534	+8	3.0
BAA	436	+7	3.5
Bank of Scotland	104	0	6.5
Barclays Bank	432	0	6.5
Bass	967	-7	4.5
BAT Inds	732	+11	5.7
BET	181	+14	9.6
BICC	440	+2	5.8
Blue Circle Inds	243	+2	6.2
BOC Group	562	+14	4.8
Boots	397	+14	4.0
British Aerospace	587	-11	5.7
British Airways	172	+3.5	6.9
British Gas	250	-1	6.9
BP	334	-2	6.6
British Steel	135	+0.5	8.1
British Telecom	381	+3	4.8
BTR	395	+3	5.3
Cable & Wireless	547	+42	2.9
Cadbury Schweppes	352	-13	4.4
Commercial Union	491	+17	6.2
Courtaulds	402	+12	4.0
Enterprise Oil	513	-13	3.9
Eurotunnel Units	470	+7	–
Fisons	494	+7	2.0
Forte	271	+3	4.9
General Accident	528	+8	6.8
GEC	192.5	-1	6.4
Glaxo Holdings	1280	+42	2.3
Grand Metropolitan	771	+12	3.6
Gt Universal Stores	1228	+32	3.7
GRE	199	+4	8.0
Guinness	985	+25	2.5
Hammerson 'A'	608	+2	4.5
Hanson	216.5	-5.5	6.5
Harrisons & Cros	148	+4	8.1
Hawker Siddeley	581	+9	5.7
Hillsdown Holdings	228	-4	4.7
ICI	291	-11	5.7
Kingfisher	499	+4	3.3
Ladbroke	268	+9	5.3
Land Securities	503	0	5.2
Lasmo	327	-12	3.5
Legal & General	433	+18	5.5
Lloyds Bank	338	+5	6.0
Lonrho	243	0	8.8
Lucas Inds	154	+3	6.1
Marks & Spencer	253	+4	3.5
Maxwell Comm	207.5	0	10.0
MEPC	474	-2	5.3
Midland Bank	211	-3	5.3
Nat Power	141	0	5.1
NatWest	313	+4	7.5
NW Water	288	+3	6.9
Pearson	730	+5	4.2
P&O dfd	572	+2	7.1
Pilkington	178	+5	8.2
Powergen	147.5	0	5.0
Prudential Corp	237	+3	5.8
Racal Electronics	221	-20	2.3
Rank Org	685	-4	6.0
RHM	270	-5	6.3
Reckitt & Coleman	1580	+2	2.9
Redland	561	-5	5.9
Reed International	432	+29	4.7
Reuters	824	+5	2.4
RMC Group	657	-16	3.9
Rolls - Royce	155	-7	6.2
Rothmans	914	+21	2.2
Royal Bank of Scotland	180	-1	6.2
Royal Insurance	436	+14	8.0
RTZ	550	-5	4.7
Sainsbury	374	+5	2.6
Scottish & Newcastle	393	+4	4.4
Sears	78	-4	9.2
Severn Trent	254	-4	6.1
Shell Transport	514	+1.5	5.2
Smith Kline Beecham	781	-6	2.4
Smith & Nephew	134.5	-0.5	4.3
Sun Alliance	370	+11	5.0
Tarmac	224	-9	6.7
Tate & Lyle	390	+35	3.4
Tesco	278	-1	2.5
Thames Water	292	-6	6.6
Thorne EMI	739	+3	5.7
Trafalgar House	256	+6	9.6
TSB	147	+1.5	5.8
Ultramar	287	-6	4.9
Unilever	745	-10	3.3
United Biscuits	361	-4	5.3
Wellcome	643	+14	1.3
Whitbread 'A'	500	-7	4.3
Williams Hldgs	308	+11	5.2
Willis Corroon	302	+12	5.8

Activity 3

Make a list of all the information you measured in Activity 1 and classify it under the three types of data.

2.2 Sources of data

The UK Government produces vast quantities of statistical information in its many departments. These are mainly coordinated by the

Office for National Statistics (formerly the Central Statistical Office and the Office of Population Censuses and Surveys) - largely responsible for producing and checking all information and data produced by individual Government Departments and also responsible for data collection based on the general public.

One essential publication to have is:

Government Statistics -A brief guide to sources. This is obtainable from the Press Office of the Office for National Statistics. It contains a list of all the important publications produced by the Government and details of how to obtain them. The most useful of these are shown below and may be available from your library or from Stationery Office Books (formerly HMSO) suppliers.

General digests

Monthly Digest of Statistics
Collection of main series from all Government departments.
Monthly.

Annual Abstract of Statistics
Contains many more series than the *Monthly Digest* and provides a longer run of years.
Annual.

Key Data
Contains over 130 tables, maps and coloured charts and covers a wide range of social and economic data. Each table and chart is accompanied by a reference to sources.
Annual.

Social Trends
Brings together key social and demographic series in colour charts and tables.
Annual.

Regional Trends
A selection of the main statistics that are available on a regional basis.
Annual.

The Annual Abstract and Social Trends are a mine of information in many fields and are kept by all good reference libraries.

In addition to the periodical data collections used in the above, various one-off reports are commissioned by the Government. Examples are:

Skateboarding Accidents in the UK - a report on accidents involving people using skateboards giving information on the nature of the accidents and injuries sustained.

Smoking/drinking amongst schoolchildren. Several studies have been carried out in these areas.

Heights and weights of people. Broken down into different age groups, for example you can find the distribution of heights and weights for 16-19 year olds in the country as a whole.

As well as the UK Government sources there are a number of other international bodies that produce statistical information. Catalogues of available publications can be obtained from your local Stationery Office Books supplier, free of charge. Some useful sources of information are:

European Community - produces much Annual Abstract/Social Trends-type data for countries in Europe. In particular, *Europe in Figures* is an inexpensive book produced annually. In addition there are a great number of reports on different issues such as employment, women's rights and the environment.

UNESCO (United Nations Educational Scientific & Cultural Organisation) **-** produces many publications in its field, not all statistical.

WHO (World Health Organisation) **-** much of it fairly technical but some interesting reports on smoking/alcoholism.

Other UK institutions providing data include:

Association of British Insurers - produces statistical information on all aspects of insurance.

Building Societies Association - in particular produces regular 'bulletins' with information on regional house prices.

High Street banks - produce regular reviews in addition to various economic and business data for their customers.

Market Research Society - in particular the *MRS Yearbook* contains useful tables on 'Market penetration of durable goods' on a regional basis. Also it has useful information on how to carry out surveys.

Meteorological Office - produces summary statistics on weather.

Various directories of business information exist giving details of companies' activities and important financial information. Company reports/share prospectuses give information in the notes to the accounts.

It should also be noted that the quality newspapers make frequent use of statistics in articles, as well as regularly publishing statistics, particularly financial. Other periodicals in fields such as economics, sociology, etc. have regular features that use statistics.

Activity 4

Take **one** of the following topics as an investigation. Collect as much information as you can using the above sources or any others you can find. Try to find at least three different sources. Write a short report using the information as reference. Outline what primary information you might collect locally for further investigation.

(a) Does the legal age of drinking/smoking need to be lowered in view of the fact that many under-age youngsters already partake?

(b) Has current Government economic policy enabled small businesses to survive more easily?

(c) Has the AIDS publicity in the early 1990s promoted a more responsible attitude towards sex in young people?

(d) Has the greater awareness of environmental issues in recent years led to any noticeable improvements in the way we look after the environment?

2.3 Sampling: factors and bias

You will have seen that secondary data can be extremely useful in investigations and will probably be collected on a much grander scale than can be done at your level. However, frequently you will be working in a new area or wish to collect your own data locally.

Every 10 years (since 1801) the Office for National Statistics (formerly the Office of Population Census and Surveys) carries out a census for the Government. The word **census** means to include everybody.

The article on the following page shows the scale of such a piece of work.

ON Sunday all householders in England, Scotland and Wales will have to fill out a form giving details of everyone who lives at their address as part of the 19th full British census.

A census is a national survey to count the population and collect information which government departments will use to plan policies. The census will attempt to give a picture of Britain at midnight on April 21. People who use the figures will be able to compare the results with statistics collected in previous censuses to find out how Britain's population and society are changing.

A 12-page form is being delivered to, and will be collected from the country's 23 million households by people known as "numerators". There are about 115,000 of these specially recruited temporary staff. Each is responsible for about 200 households.

A further 1, 800 temporary staff will key the census information into a massive government computer in Titchfield, Hampshire. The whole process of collecting and processing the data costs the Government about £135 million.

In this year's census new questions will be asked about people's ethnic origin and any long-term illness they might have. For the first time, an attempt will be made to count the number of homeless people in Britain.

The census is held every ten years on a Sunday, the day most people are at home. It is organised by the Office of Population Censuses and Surveys (OPCS) in England and Wales, and by the Registrar General in Scotland. Separate censuses will also be held on April 21 in Northern Ireland and the Irish Republic.

Most countries count their populations. The United States, for instance, has held a census every 10 years since 1790. Early this year a census in India showed that it has a total population of 844 million people. Australia's latest census, by contrast, showed it has just 17 million people spread across a land area twice as large as India's.

In 1975 the Government wanted census information before the 1981 full census, so the OPCS carried out a ten per cent census using 1 in 10 of the population. This is known as a **survey**. Data are obtained by asking people to fill in forms which are then given to collectors trained to sort out any queries.

In a research project looking at the disappearance of vegetation on mountain moorland, a scientist chose three specific sites to investigate. Fifty samples were selected at each site using a device called a quadrat (a 10 cm wire square) thrown at random into the undergrowth. The number of species of each type and the sizes were noted by students who were able to identify the plants.

Both these examples illustrate the same principle. When deciding how to carry out a data collection there are several decisions to be made:

(a) What size of sample can you reasonably expect to take, given limited time, money and resources?
(b) How are the items to be used in the sample to be chosen to avoid introducing bias?
(c) How is the data to be collected to avoid any bias?

The answer to question (a) clearly depends on the individual circumstances. It should be obvious, however, that the larger the sample the more sensitive the result.

In questions (b) and (c) the key element is to eliminate possible bias. In order to understand **bias** the idea of **factors** in an experiment is important. You are usually interested in one or more factors and their effect. However, there will always be other factors which might affect the result. For example, a horticulturist wishes to test the effect of a new fertilizer on different varieties of wheat. Some possible factors affecting the experiment could be listed as:

Relevant Factors	Bias Factors
Whether fertilizer used	Type of soil
Strength of fertilizer	Weather conditions
Variety of wheat used	Quality of seeds
	Care of plants
	Measurement of crop
	Position in field

The strength of fertilizer is really a sub factor of whether a fertilizer is used or not. You could list the strength as litres per square metre including zero. These are called the **levels of a factor**.

Activity 5

Make a list of relevant and bias factors for these experiments :

(a) Testing a new fuel additive to improve mileage in different cars.

(b) Testing whether a new language laboratory improves student performance in modern and classical languages.

(c) Examining the effect of alcohol on men's and women's reaction time.

(d) Asking people's opinions of current unemployment.

Where there are levels of a factor, indicate possible values the levels could take.

If a firework manufacturer wanted to test whether his product worked he could not possibly test every item as he would have nothing left to sell. He would try to take a 'representative' sample of all the fireworks he produced. By **representative** we mean that the sample has approximately the same properties as the total 'population'. This is illustrated in the following case study.

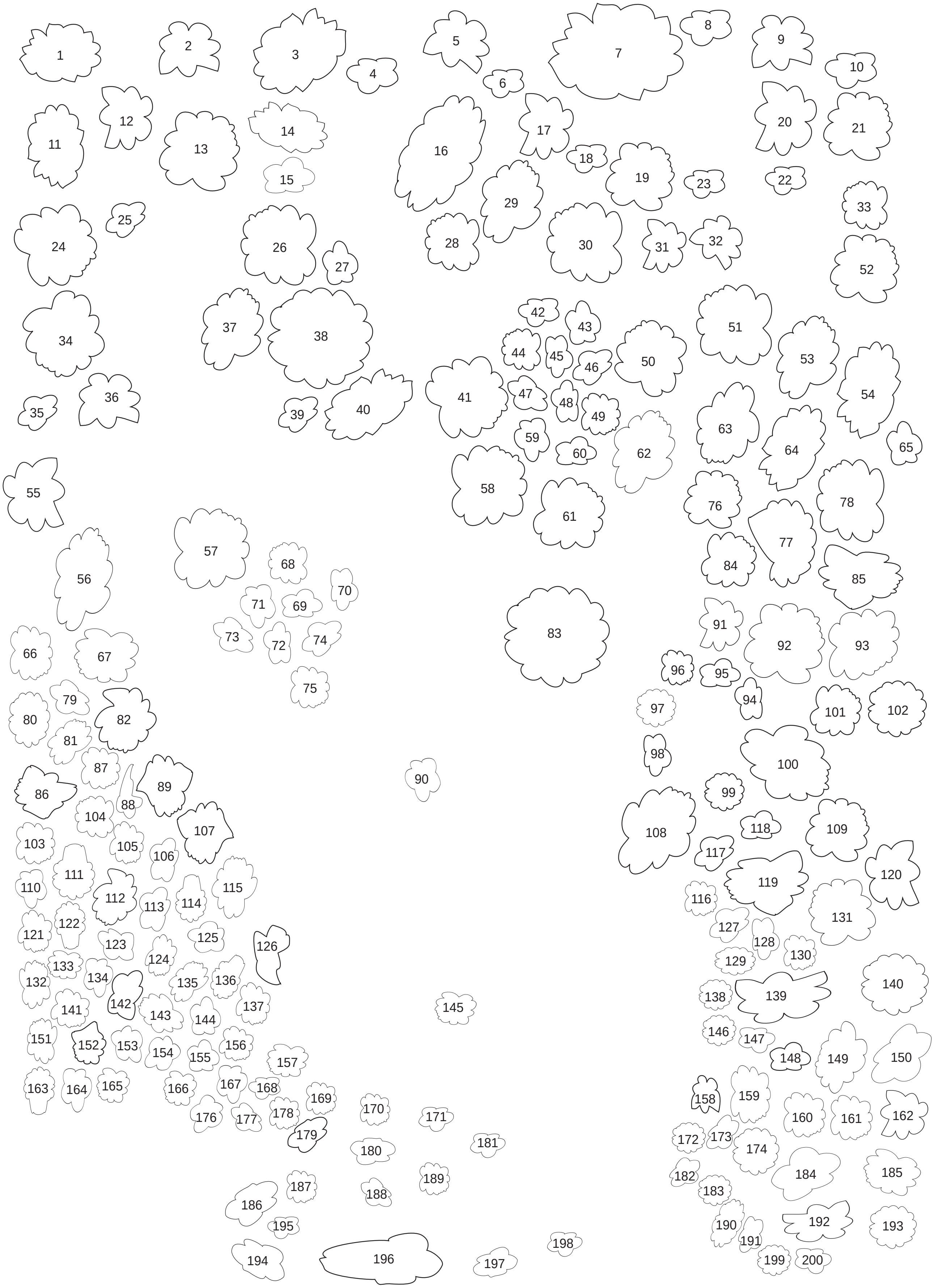
A landowner has decided to sell a mature piece of deciduous woodland of 200 trees. He has asked a surveyor to come and assess the quality of the woods, but in the time available she can only carefully examine 50 trees. The landowner has a map of the woods (shown on the following page) on which he has numbered all the trees and indicated the variety. The surveyor says that the following details will be needed for each of 50 trees:

(a) the girth ;

(b) the age;

(c) whether it suffers from a major disease;

(d) the approximate height.


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200

Tree	Type	Girth	Age	Disease	Height	Value
1	Oak	2.1	80	0	7	120
2	Elm	1.8	65	0	6	90
3	Oak	3.5	115	0	8	200
4	Birch	0.8	20	0	3	0
5	Elm	1.9	65	0	6	95
6	Birch	0.6	18	0	3	0
7	Oak	4.6	150	0	8	300
8	Birch	0.7	19	1	3	0
9	Elm	1.7	60	1	5	0
10	Birch	0.8	21	0	2	0
11	Elm	1.7	70	0	6	80
12	Elm	1.7	72	0	6	80
13	Oak	2.1	90	0	7	120
14	Yew	2.3	130	0	7	300
15	Birch	0.7	20	1	3	0
16	Oak	4.5	145	0	7	240
17	Elm	2.1	75	0	6	90
18	Birch	0.7	18	0	3	0
19	Oak	3.2	108	0	8	180
20	Elm	1.7	67	1	6	0
21	Elm	1.6	65	1	6	20
22	Birch	0.7	18	1	3	0
23	Birch	0.6	15	0	2	0
24	Oak	2.9	102	0	7	115
25	Birch	0.6	21	0	3	0
26	Oak	3.1	110	0	8	175
27	Birch	0.9	23	1	3	0
28	Elm	1.8	74	0	6	90
29	Oak	3.2	110	0	8	170
30	Oak	3.8	120	0	9	195
31	Elm	2.0	75	1	6	0
32	Elm	2.3	75	1	7	30
33	Elm	2.2	75	1	7	0
34	Elm	2.6	80	0	7	90
35	Birch	0.6	20	0	3	0
36	Elm	2.5	78	0	7	95
37	Elm	2.8	85	0	7	100
38	Oak	3.7	116	1	8	80
39	Birch	0.7	23	1	3	0
40	Elm	2.8	80	0	7	95
41	Elm	3.3	95	0	7	110
42	Birch	0.6	21	0	3	0
43	Birch	0.6	20	0	3	0
44	Birch	0.5	17	0	2	0
45	Birch	0.6	22	0	3	0
46	Birch	0.6	21	1	3	0
47	Birch	0.6	20	0	3	0
48	Birch	0.6	21	1	3	0
49	Birch	0.5	18	0	3	0
50	Elm	3.5	98	0	7	120
51	Oak	4.1	120	0	8	180
52	Oak	3.9	115	0	7	165
53	Oak	3.1	85	0	7	135
54	Oak	4.1	118	0	8	170
55	Elm	2.8	80	0	7	95
56	Oak	4.0	118	0	8	170
57	Yew	4.7	120	0	9	280
58	Elm	3.3	90	0	6	100
59	Birch	0.6	21	0	3	0
60	Birch	0.6	20	0	3	0
61	Elm	3.2	85	0	6	80
62	Elm	3.2	88	0	6	80
63	Oak	3.5	108	0	8	150
64	Oak	3.4	105	0	8	145
65	Elm	2.1	45	0	6	60
66	Beech	2.5	55	0	5	70
67	Oak	3.0	90	0	7	130
68	Birch	0.7	23	0	3	0
69	Birch	0.6	22	0	3	0
70	Birch	0.6	22	1	3	0
71	Birch	0.6	21	0	3	0
72	Birch	0.6	20	1	3	0
73	Birch	0.7	22	0	3	0
74	Birch	0.6	21	0	3	0
75	Birch	0.6	21	0	3	0
76	Elm	2.9	81	0	7	90
77	Oak	4.3	125	0	8	190
78	Oak	4.4	127	0	8	195
79	Beech	2.4	55	0	5	70
80	Beech	2.6	55	0	5	75
81	Beech	2.4	55	0	5	70
82	Oak	3.5	98	0	7	150
83	Yew	5.0	150	0	9	300
84	Elm	2.8	78	0	7	85
85	Oak	4.3	125	0	8	185
86	Beech	2.6	55	0	6	80
87	Beech	2.5	55	0	5	75
88	Beech	2.5	55	0	5	75
89	Oak	3.6	100	0	7	145
90	Beech	2.9	80	0	7	90
91	Elm	2.8	81	0	7	85
92	Oak	3.4	102	0	7	150
93	Oak	3.6	102	0	7	150
94	Birch	0.6	21	0	3	0
95	Birch	0.6	20	0	3	0
96	Birch	0.5	18	0	3	0
97	Birch	0.6	20	0	3	0
98	Birch	0.6	21	1	3	0
99	Birch	0.6	20	1	3	0
100	Elm	2.9	80	1	7	20
101	Elm	2.8	83	1	7	0
102	Elm	2.7	80	1	7	15
103	Beech	2.6	55	0	7	75
104	Beech	2.5	55	0	7	70
105	Beech	2.4	55	0	7	60
106	Beech	2.4	55	0	7	60
107	Oak	4.2	102	1	8	30
108	Elm	4.3	98	0	8	110
109	Elm	3.1	84	1	8	15
110	Beech	2.5	55	0	7	75
111	Beech	2.4	55	1	7	10
112	Beech	2.5	55	0	7	70
113	Beech	2.5	55	0	7	75
114	Beech	2.4	55	0	7	70
115	Oak	3.9	95	0	8	130
116	Birch	0.6	20	0	3	0
117	Birch	0.6	19	0	3	0
118	Birch	0.7	22	0	3	0
119	Yew	4.1	110	0	8	200
120	Elm	3.3	85	0	8	120
121	Beech	2.6	55	0	7	75
122	Beech	2.5	55	0	7	70
123	Beech	2.5	55	0	7	70
124	Beech	2.5	55	0	7	70
125	Beech	2.6	55	0	7	75
126	Oak	3.7	90	0	8	125
127	Birch	0.6	20	0	3	0
128	Birch	0.7	21	0	3	0
129	Birch	0.6	20	0	3	0
130	Birch	0.6	20	0	3	0
131	Elm	3.5	90	0	8	130
132	Beech	2.5	55	0	7	75
133	Beech	2.4	55	0	7	70
134	Beech	2.5	55	0	7	75
135	Beech	2.3	55	0	6	60
136	Beech	2.5	55	0	7	75
137	Beech	2.5	55	0	7	75
138	Birch	0.6	20	0	3	0
139	Beech	2.2	48	0	6	60
140	Elm	3.7	87	1	7	10
141	Beech	2.5	55	1	7	20
142	Beech	2.6	55	0	7	80
143	Beech	2.5	55	0	7	75
144	Beech	2.5	55	0	7	75
145	Beech	2.3	47	0	6	60
146	Birch	0.6	20	0	3	0
147	Birch	0.7	22	1	3	0
148	Oak	3.8	85	0	7	140
149	Oak	3.6	85	0	7	130
150	Oak	4.1	85	0	8	150
151	Beech	2.5	55	0	7	75
152	Beech	2.5	55	0	7	75
153	Beech	2.4	55	0	7	70
154	Beech	2.5	55	0	7	75
155	Beech	2.5	55	1	7	15
156	Beech	2.4	55	0	7	70
157	Elm	3.9	85	0	8	80
158	Birch	0.6	20	0	3	0
159	Oak	4.3	85	0	7	160
160	Oak	3.9	85	0	7	150
161	Oak	3.8	85	0	7	150
162	Oak	3.8	85	0	7	150
163	Beech	2.4	55	0	7	70
164	Beech	2.5	55	0	7	75
165	Beech	2.4	55	0	7	70
166	Beech	2.5	55	0	7	75
167	Beech	2.5	55	0	7	75
168	Birch	0.6	19	0	3	0
169	Birch	0.6	20	0	3	0
170	Birch	0.6	19	0	3	0
171	Birch	0.6	20	1	3	0
172	Birch	0.5	17	1	3	0
173	Birch	0.6	18	0	3	0
174	Elm	3.2	76	1	7	10
175	Beech	2.5	55	0	7	75
176	Beech	2.7	55	0	7	80
177	Birch	0.7	21	0	3	0
178	Birch	0.6	19	0	3	0
179	Beech	1.4	22	0	4	15
180	Birch	0.6	19	0	3	0
181	Birch	0.6	18	0	3	0
182	Birch	0.6	19	0	3	0
183	Birch	0.6	19	0	3	0
184	Elm	3.5	83	0	7	85
185	Elm	2.9	72	0	7	75
186	Beech	2.5	55	0	7	75
187	Birch	0.6	20	0	3	0
188	Birch	0.5	15	0	2	0
189	Beech	1.7	31	0	5	30
190	Beech	1.6	28	0	4	20
191	Birch	0.6	17	0	3	0
192	Elm	2.7	54	0	5	30
193	Elm	2.9	51	0	5	30
194	Elm	2.9	48	0	5	30
195	Birch	0.6	15	1	3	0
196	Yew	4.2	124	0	8	200
197	Beech	1.9	38	0	6	35
198	Birch	0.6	19	0	3	0
199	Birch	0.6	18	0	3	0
200	Birch	0.6	21	0	3	0

From this information it should be possible to estimate the value of the trees as timber.

The surveyor and landowner discuss various methods which might be used to pick the 50 trees. They come up with the following ideas:

(a) Drop a pin on the map and take the tree nearest to the point of the pin.

(b) Use a random number generator on a calculator to give 50 numbers between 1 and 200 and select these trees.

(c) Take every 4th tree using the numbers in order.

(d) Divide the area into squares and take the same number of trees in each square.

(e) Count the total number of oaks and divide by 4. Choose that number of oaks at random. Similarly with each of the other varieties.

Activity 6

As a group get everyone to try one of the methods (a) to (e) or one of their own choice. Shade on a copy of the map of the woods the trees you would sample.

Now using the information on the data worksheet on the previous page, find for each method:

(a) the proportion of oaks in your sample.

(b) the average girth of trees.

(c) the average age of the trees.

(d) the proportion of diseased trees.

(e) the tallest tree.

(f) the total value of the woods.

The data columns on the data worksheet show

girth in metres
age in years
disease: 0 - clear, 1 - diseased
height (approx) in metres
value in £.

Using all 200 trees the values are:

(a)	18%	(b)	2.15 m	(c)	58 years
(d)	16%	(e)	9 m	(f)	£12 925

Compare the results of each method with the overall results. What problems occurred in using the various methods in practice?

The main methods used for sampling in practice are as follows:

(a) **Random -** to be truly random each individual must have an equal chance of being chosen. Dropping a pin on the map is not truly random in this case as it is more likely to select the larger trees. This method is often used for selecting people from Electoral Registers. If the researcher is calling at people's houses the system must be rigidly adhered to (i.e. call back if people are out). It does not necessarily ensure a representative sample.

(b) **Systematic -** taking items at regular intervals e.g. every 4th tree. Although this does not necessarily ensure a representative sample it should be better than random sampling. Again the system must be rigidly adhered to. This method is often used when sampling goods on a production line.

(c) **Stratified -** this is used to ensure that the sample is representative and that it has the same proportions as the population, e.g. ensuring that the sample of trees has the right proportion of each variety. To do this you would need first of all to divide the whole of the population into appropriate categories. This can be very difficult in practice. What is commonly used in street surveys is a **quota** sampling method where interviewers are simply asked to interview a certain proportion of each type, e.g. age, and these can be chosen at random. A common division used is social class. This is defined by the type of job done. The table opposite gives the approximate divisions of social class currently in use.

Composition of Social Classes

Social Class	Main Occupations %
I Professional	Men: engineers and scientists (47.6), accountants (9.2), surveyors (8.5), doctors (5.0), architects (4.7) Women: company secretaries (23.6), doctors (12.7), engineers and scientists (9.5), pharmacists (5.4), clergy and members of religious orders (5.2)
II Intermediate	Men: managers (28.9), proprietors and managers, sales (17.8), teachers (10.6), technicians (9.6), farmers (9.2). Women: teachers (26.7), nurses (24.5), proprietors and managers, sales (16.6), technicians (4.7), managers (4.7).
IIIN Skilled non-manual	Men: clerks, cashiers (51.3), salesmen (20.5), shop assistants (10.6), draughtsmen (8.0), policemen (6.3). Women: clerks, cashiers (46.3), shop assistants (23.7), typists (23.7), office machine operators (4.4).
IIIM Skilled manual	Men: lorry drivers (10.5), lifters (10.5), carpenters (7.2), electricians (5.0), bricklayers (4.9). Women: hairdressers (15.1), cooks (14.1), skilled textile workers (11.7), dressmakers (10.7), printing workers (7.4).
IV Partly skilled	Men: warehousemen (14.4), construction workers (8.8), agricultural workers (8.4), machine tool operators (8.7), metal makers (6.1). Women: maids (18.4), canteen assistants (12.7), partly skilled textile workers (12.6), packers (9.3), telephone operators (4.2).
V Unskilled	Men: labourers (82.6), office cleaners (5.8). Women: office cleaners (64.2), labourers (19.6), kitchen hands (15.1).

(d) **Purposive -** in some cases a deliberately biased sample is taken for a particular purpose. If, for example, you wished to test the popularity of a new teenage magazine you would not ask senior citizens. You would, however, ensure the correct proportion of male/female in relation to overall readership.

(e) **Cluster -** sometimes there is a natural sub-grouping of the population - for example, parliamentary constituencies. In this case, you first choose a random sample of clusters and then a sample inside each one. This method can be far less costly than taking a random sample from the whole population.

Use of random digit tables

For method (a), you could use the random digit table given in the Appendix. Starting arbitrarily on row 10, combining three digits together gives numbers from 000 to 999. Only use numbers in the region 001 to 200; the start of the sequence is:

572	178	878	377	127	957	834	066	...	...
	↑			↑			↑		
	accept			accept			accept		

(You normally ignore any repeats if they exist.)

You can attempt to find a random sample more quickly by dividing each three-digit number by 200 and taking the remainder. This would give:

172 178 078 177 127 157 034 066 ...

Would this sample be truly random?

Unfortunately not quite unless 000 is taken as 200, or you take 'the remainder on division by 200 of the three-digit number plus 1'.

Activity 7

Suppose your population is numbered 000 to 299. Use the random digit sheet by taking consecutive three digits. Taking the remainder after division by 300 does **not** give a random sample. Why not?

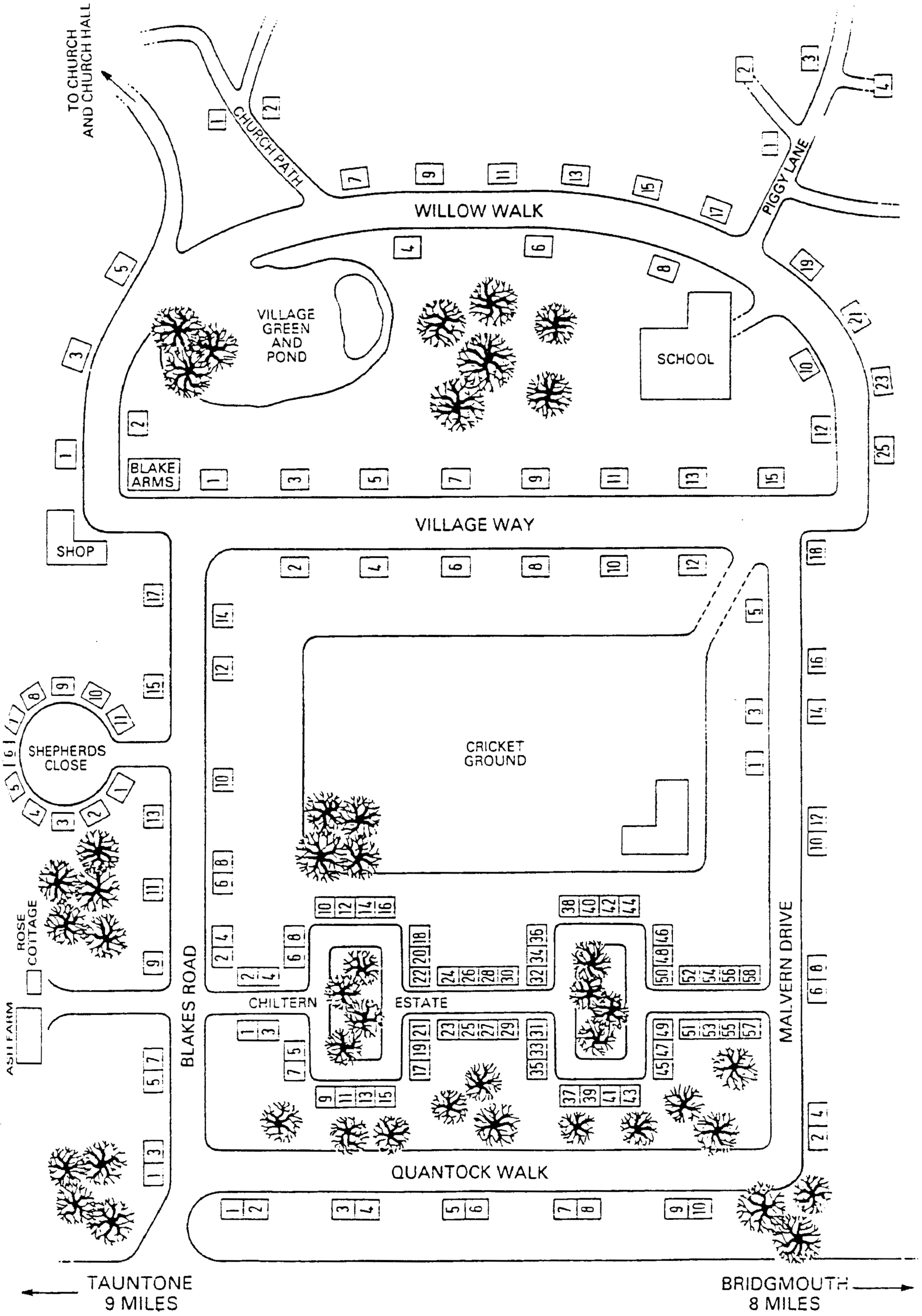
Activity 8

The map opposite shows a small village of 150 houses (including Ash Farm, Rose Cottage, The Blake Arms and the Shop). The village is due to be redeveloped and the Parish Council wishes to know which of three types of development the village would prefer (these are referred to as C - community centre, H - housing estate, L - large supermarket).

You are asked to undertake a survey of views by sampling 20% of the houses. Use

(a) a systematic sample (b) a random sample

to survey opinion. The views of all the householders are given in the table following. Compare your answers from (a) and (b) with the views of the complete population.


TO CHURCH AND CHURCH HALL
CHURCH PATH
WILLOW WALK
PIGGY LANE
VILLAGE GREEN AND POND
SCHOOL
BLAKE ARMS
VILLAGE WAY
SHOP
SHEPHERDS CLOSE
ROSE COTTAGE
ASH FARM
CRICKET GROUND
BLAKES ROAD
CHILTERN
ESTATE
MALVERN DRIVE
QUANTOCK WALK
TAUNTONE 9 MILES
BRIDGMOUTH 8 MILES

Road	House name or number	Preference
	Rose Cottage	H
	Ash Farm	C
	Blake Arms	H
	Shop	H
Blakes Rd	1	L
	2	L
	3	H
	4	L
	5	H
	6	L
	7	L
	8	C
	9	H
	10	H
	11	C
	12	C
	13	H
	14	L
	15	H
	17	C
Chiltern Estate	1	L
	2	L
	3	H
	4	H
	5	L
	6	L
	7	L
	8	C
	9	L
	10	L
	11	C
	12	H
	13	L
	14	L
	15	L
	16	H
	17	C
	18	L
	19	L
	20	L
	21	L
	22	H
	23	H
	24	H
	25	L
	26	L
	27	C
	28	C
	29	L
	30	L
	31	H
	32	H
	33	C
	34	H
	35	C
	36	C
	37	C
	38	L
	39	L
	40	H
	41	H
	42	L
	43	L
	44	L
	45	C
	46	C
	47	L
	48	L
	49	L
	50	H
	51	H
	52	H
	53	L
	54	H
	55	C
	56	L
	57	H
	58	H

Road	House name or number	Preference
Church Path	1	C
	2	C
Malvern Drive	1	L
	2	L
	3	L
	4	C
	5	H
	6	L
	8	H
	10	C
	12	H
	14	L
	16	H
	18	L
Piggy Lane	1	L
	2	L
	3	H
	4	H
Quantock Walk	1	L
	2	L
	3	C
	4	C
	5	C
	6	H
	7	L
	8	L
	9	L
	10	H
Shepherds Close	1	H
	2	H
	3	C
	4	C
	5	C
	6	H
	7	H
	8	H
	9	L
	10	H
	11	H
Village Way	1	L
	2	H
	3	L
	4	H
	5	C
	6	C
	7	C
	8	L
	9	H
	10	C
	11	H
	12	H
	13	C
	15	C
Willow Walk	1	C
	2	C
	3	C
	4	C
	5	H
	6	H
	7	C
	8	C
	9	H
	10	H
	11	H
	12	H
	13	L
	15	L
	17	H
	19	L
	21	H
	23	H
	25	C

KEY

L – large supermarket
H – housing estate
C – community centre

2.4 Miscellaneous Exercises

1. Pupils in a statistics class want to choose a sample of 100 from a school where the numbers of pupils in each year are shown below.

Year	1	2	3	4	5	6
No. of pupils	290	285	310	175	92	48

(a) Explain how this sample could be obtained by picking a random sample.

(b) If a stratified random sample is chosen, explain how this could be done and how many pupils from each year group are to be chosen for the sample.

2. A factory has 500 employees, each one having a 'works number'. For the purposes of a survey a sample of 25 is picked from the work-force.

Explain

(a) how a systematic sample of 25 could be chosen;

(b) how a random sample, using random numbers, could be chosen;

(c) how a random sample could be chosen, without the use of random numbers.

3. Following a spell of particularly bad weather, an insurance company received 42 claims for storm damage on the same day. Sufficient staff were available to investigate only six of these claims. The others would be paid in full without investigation. The claims were numbered 00 to 41 and the following suggestions were made as to the method used to select the six. In each case six different claims are required, so any repeats would be ignored.

Method 1	Choose the six largest claims
Method 2	Select two-digit random numbers, ignoring any greater than 41. When six have been obtained choose the corresponding claims.
Method 3	Select two digit random numbers. Divide each one by 42, take the remainder and choose the corresponding claims (eg if 44 is selected claim number 02 would be chosen).
Method 4	As 3, but when selecting the random numbers ignore 84 and over.
Method 5	Select a single digit at random, ignoring 7 and over. Choose this and every seventh claim thereafter (e.g. if 3 is selected, choose claims numbered 03, 10, 17, 24, 31 and 38).

Comment on each of the methods, including an explanation of whether it would yield a random sample or not.

4. In a small village, the population is divided by age groups as shown in the table.

Age (years)	0-4	5-14	15-44	45-64	65+
No. of people	14	41	50	70	14

It is proposed to choose a stratified random sample of 40 from the village. Explain how this should be done and calculate how many people should be picked from each age range.

5. Explain briefly what is meant by a random sample. State an advantage of using random, rather than non-random, sampling methods.

Explain the difference between a stratified random sample and a quota sample, and state one advantage of the latter as compared with the former.

An area health authority decides to undertake a survey, using a questionnaire, to determine the proportion of adults who are in favour of local hospitals becoming self-governing trusts. The survey will also investigate patients' attitudes to the treatment presently provided by the hospitals, and aims to collect information from at least 500 adults.

Three possible methods of obtaining the required information are considered.

Method A Choose 1000 adult patients at random from the area's hospitals' records. Arrange for interviewers to visit these patients and ask for the questionnaire to be completed there and then.

Method B Choose names at random from the area's telephone directories. Contact the individuals so chosen, by telephone, and ask if they are willing to answer the questionnaire over the telephone. Continue until enough individuals have agreed to take part.

Method C Choose 2000 names from the area's electoral registers. Send out the questionnaire, by post, to the selected individuals with prepaid envelopes for the questionnaires' return.

(a) Comment critically on the suitability of each of these three methods.

(b) Outline the method you would advise for collecting the required information.