1 Planning a Statistical Investigation and Collecting Data

1.1 Hypothesis Testing

Statistical investigations test statements and theories that may turn out to be true or false. These are called hypotheses. A hypothesis is a theory that you think might be true. For example:

- **Hypothesis 1**: Successful football teams win most home matches and draw most away matches.
- **Hypothesis 2**: Girls score higher marks on GCSE coursework than boys.
- **Hypothesis 3**: Boys are allowed to stay out later in the evening than girls of the same age.

Identifying the hypotheses to be tested is an important part of the planning process and they need to be stated clearly.

Worked Example 1

Phil needs to throw a six with a dice to start the game of Snakes and Ladders. After 20 throws he has still not thrown a six. State a hypothesis that Phil might investigate.

**Solution**

*Hypothesis*: The dice is biased.

*Note*  This hypothesis could be tested by throwing the dice a large number of times and seeing if the numbers 1, 2, 3, 4, 5, and 6 each come up approximately one sixth of the time.

Worked Example 2

Last year people in a town complained about the high number of burglaries that had been committed. The police then ran a high profile campaign encouraging people to improve their home security. State a hypothesis that should be investigated to test the effectiveness of their campaign.

**Solution**

*Hypothesis*: The campaign was successful.

*Note*  A statistician would investigate this hypothesis by investigating whether the burglary rate fell in the town after the police campaign.
1.2 Data

**Primary and Secondary Data**
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 involves looking at the whole population, not just a sample. A census is usually carried out every 10 years in this country and is compulsory by law to complete.

Data such as census data or, for example, stock exchange data or temperature data (as can be found in most newspapers) is known as secondary data, as you are relying on someone else to collect it. Data which you have collected yourself is called primary data.

**Data Collection**
*Primary data* can be collected in a variety of ways including:
- experiments
- surveys
- data logging
- questionnaires.

Before undertaking the full data collection it is advisable to do a trial run using a small pilot study. This will enable you to check that:
- questions are well phrased,
- the data collection method is effective,
- the data collected will be useful and appropriate,
- the data collected will provide a fair and rigorous test of the hypotheses.

*Secondary data* can be obtained from a variety of sources such as:
- the internet
- databases
- books
- newspapers
- historical records.

When secondary data is used care needs to be taken about its reliability.

**Qualitative and Quantitative Data**
An even more important distinction between types of data is to what extent numbers are involved.

*Qualitative data* is generally descriptive in nature rather than numerical, e.g. the starting letter of someone’s name, the 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 or temperature. This category is further divided into:

(a) *discrete data* – where the data can only be one of a fixed number of numerical values, usually, but not necessarily, whole numbers, e.g. number of accidents on a motorway in a specified month;
(b) **continuous data** – where the data can fall anywhere over a range and the scale is only restricted by the accuracy of measuring, e.g. mass or height of people.

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

### Worked Example 1

A collection of bicycles was displayed in a shop.

(a) State a **qualitative** variable about these bicycles.

(b) State a **continuous quantitative** variable about these bicycles.

*Solution*  
(a) ‘colour’ or ‘crossbar (Yes / No)’  
(b) ‘radius of front wheel’, ‘circumference of back wheel’ or ‘length of chain’, etc.

*Note*  
‘Number of gears’ would be an example of a discrete quantitative variable.

### Worked Example 2

A survey was carried out at a post office.

State whether each of the following variables is **discrete** or **continuous**.

(a) The number of people entering the post office per hour.  
(b) The time it takes to serve each person.  
(c) The number of people buying postage stamps.  
(d) The total amount paid by each customer.

*Solution*  
(a) discrete  
(b) continuous  
(c) discrete  
(d) discrete
Exercises

1. Complete the table by naming the type of data formed by each of the stated measurements. The first one has been completed for you.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height of Rose Trees</td>
<td>Continuous</td>
</tr>
<tr>
<td>Number of brothers</td>
<td></td>
</tr>
<tr>
<td>Length of shoe laces</td>
<td></td>
</tr>
<tr>
<td>Number of pages in library books</td>
<td></td>
</tr>
</tbody>
</table>

2. (a) Write down *one* way of obtaining primary data for a statistical analysis.
   (b) What is the statistical term given to data obtained from published statistics or known databases?
   (c) Give *one* disadvantage of using published statistics.

3. State whether each of the following variables is qualitative, discrete or continuous.
   (a) the number of goals scored in Premier league soccer matches on a Saturday
   (b) the colour of children's eyes in a class
   (c) the circumference of apples collected from a tree
   (d) the type of vehicle seen on a road at rush hour.
1.3 Questionnaires and Surveys

Note

When designing a questionnaire to use as part of a survey, bear in mind the following guidelines.

- The questionnaire should try to find out the information you need.
- You should know how you are going to collect your responses.
- Questions should be clear and concise with no ambiguity.
- Do not ask for information you do not need.
- Allow for all possible responses.
- Questionnaires should be fair and not biased in any way.
- The people asked to complete the questionnaire should be from a variety of backgrounds.

It should also be noted that if you need to design a questionnaire for your project work, it is strongly recommended that you first try it out with a small number of people. This is called a pilot survey.

The purpose of the pilot survey is:

(a) to detect errors in the wording of the questions,
(b) to check for ambiguity in any of the questions,
(c) to detect unexpected responses to any of the questions (e.g. 'Other' as a response should never be the largest data group!),
(d) to find any other difficulties that may cause problems collecting the data responses required.

It is essential that the outcome of the pilot survey and any subsequent changes to the questionnaire/data collection system should be fully documented and included in any handling data coursework.

Questions can be designed in two distinct ways, namely those that require a specified response to be chosen from a number of options or by giving a number, and those that allow more detailed responses. The first group are often referred to as closed questions; the second as open questions. Here are some examples.

<table>
<thead>
<tr>
<th>Closed</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Did you watch the football match on TV last night?&quot;</td>
</tr>
<tr>
<td>&quot;How many hours of TV did you watch last night?&quot;</td>
</tr>
<tr>
<td>Choose from '0 – 1, 1 – 2, 3 – 4, 4 – 5, more than 5'</td>
</tr>
</tbody>
</table>

The second example of a closed question shows an example with multiple responses – by which we mean more than 2. So 'YES / NO' is not regarded as a multiple response.

There is though a problem with these responses as there is potential overlap, i.e. in which category do you place 1 hour or 2 hours, etc.? It should be made clear, i.e. 0 – 1 includes everything up to and including 1, etc.
The second example is good for a face to face discussion, as it allows a wide range of answers. The questioner will then be able to react to these answers, and continue the discussion. This example is bad for a questionnaire for the same reason! Any answer is possible but the writer of the questionnaire does not have the chance to follow up answers. It could though be modified and become a closed question with multiple responses: for example,

"Please tick a box to show where you went on holiday last year."

- [ ] UK
- [ ] USA
- [ ] France
- [ ] Italy
- [ ] Greece
- [ ] Other (please say where)

**Worked Example 1**

Consider this questionnaire:

Are you concerned about the environment?

Are you concerned about the level of pollution caused by cars?

Do you think the health of young children is at risk due to exhaust fumes from cars?

Is there too much congestion in the city centre?

Is public transport under-used?

Do you think cars should be banned from the city centre?

Comment on the questions given here.

**Solution**

This questionnaire is very biased and has been designed to lead people towards answering 'yes' to the last question. There are no questions about the advantages of cars, the problems of using public transport or other ways of reducing pollution.

**Worked Example 2**

What is wrong with this survey, used to find the favourite sports of the students in one school?

<table>
<thead>
<tr>
<th>Which is your favourite sport?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis</td>
</tr>
</tbody>
</table>

- [ ] Tennis
- [ ] Rugby
- [ ] Netball
- [ ] Basketball
- [ ] Swimming
- [ ] Snooker
- [ ] Other
Solution

The question used in this survey does not cover all the possible responses well. Many students might have, for example, football as their favourite sport and so would tick the box marked ‘Other’. This would give unhelpful results. Also, students are asked to make only one choice although they may well like more than one sport.

The following wording would produce more useful responses.

"Please tick one box in each row to show your preference."

<table>
<thead>
<tr>
<th></th>
<th>I really enjoy this sport</th>
<th>I like this sport</th>
<th>I've no great preference</th>
<th>I dislike this sport</th>
<th>I hate this sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rugby</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netball</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basketball</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swimming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snooker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Please put any other sport that you like here</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Worked Example 3

A survey is carried out to investigate the quality of new houses.

All the people in the country who have recently bought a new house are to be surveyed.

(a) Give one reason why a pilot survey might be carried out first.

(b) Give one reason why the survey should be carried out by post.

(c) Give one possible disadvantage of a postal survey.

Solution

(a) To see if there are any difficulties with any of the questions in the survey.

(b) It is far cheaper than sending people to each new house in the country.

(c) Not everyone will respond to a postal survey so the survey might not be truly representative.
Exercises

1. Consider the following two sets of questions.

<table>
<thead>
<tr>
<th>Is meat a good source of protein?</th>
<th>Is all meat obtained from dead animals?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is protein an important part of your diet?</td>
<td>Do you like killing animals?</td>
</tr>
<tr>
<td>Do you like eating meat?</td>
<td>Do you like eating meat?</td>
</tr>
</tbody>
</table>

(a) Comment on the way that the last question might be answered in each case.
(b) Write a better set of questions to find out if people like eating meat.

2. Design a questionnaire that would encourage students at a school to say:
   (a) that school uniform is a good policy,
   (b) that school uniform should be abandoned.

3. Wendy asks her class the following question.

   Which of these breakfast cereals do you like best?
   - Cornflakes
   - Porridge
   - Bran Flakes

   She then says that her class's favourite breakfast cereal is Cornflakes.

   (a) Is her conclusion valid?
   (b) Criticise her question.
   (c) Write a better question for her to use.

4. Design a questionnaire which you could use to find out:
   (a) if people think they have a good public transport service,
   (b) whether the music tastes of girls and boys are different,
   (c) whether younger people are more likely than older people to be vegetarian,
   (d) who might win the next General Election.

   Use your questionnaire to collect some data and present your results.

5. You have been asked to estimate what percentage of all cars on the road have each registration number code (51, 02, 52, ...).

   (a) Describe how you would collect the data for a survey to answer this problem.
   (b) Describe any problems that might arise as you collect data.
6. Akiko did a survey to find out which colours of car are popular. She found that red was the most popular colour. She did not find any cars that were purple or pink.

Think about a survey which you could do. It must not be about the colour of cars.

(a) What is your survey about?
(b) Write down two things which you might expect to find out.

7. In a survey of community life on a new housing estate the following question is suggested.

(A) "What do you most like about living here?"

An alternative is proposed.

(B) "Tick the box which describes why you most like living here."

- Design of houses
- Friendliness of neighbours
- More open space

Give one advantage of each form of question.

8. A survey is done to find out in which sports pupils take part.

The results of the survey are shown below.

<table>
<thead>
<tr>
<th>Sport</th>
<th>% Boys</th>
<th>% Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athletics</td>
<td>37.1</td>
<td>33.3</td>
</tr>
<tr>
<td>Basketball</td>
<td>12.7</td>
<td>4.5</td>
</tr>
<tr>
<td>Cycling</td>
<td>71.7</td>
<td>60.5</td>
</tr>
<tr>
<td>Dancing</td>
<td>4.9</td>
<td>31.5</td>
</tr>
<tr>
<td>Football</td>
<td>72.0</td>
<td>14.5</td>
</tr>
<tr>
<td>Gymnastics</td>
<td>5.0</td>
<td>16.3</td>
</tr>
<tr>
<td>Hockey</td>
<td>6.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Judo</td>
<td>16.0</td>
<td>6.7</td>
</tr>
<tr>
<td>Netball</td>
<td>3.3</td>
<td>35.0</td>
</tr>
<tr>
<td>Roller-skating</td>
<td>7.3</td>
<td>26.0</td>
</tr>
<tr>
<td>Rugby</td>
<td>11.8</td>
<td>2.5</td>
</tr>
<tr>
<td>Swimming</td>
<td>53.1</td>
<td>55.0</td>
</tr>
</tbody>
</table>

(a) Which two sports are the most popular with both boys and girls?
(b) The question asked in the survey was:

Which sports do you take part in?

You want to find out more about pupils' involvement in sport.

Write down another question that you could ask.
9. The school governors are worried about road safety outside the school gates. They think there ought to be a Pelican Crossing. To decide on this, they need to find out how much traffic comes along the road at different times of day and on different days of the school week. Design an observation sheet to gather this information. (NEAB)

10. Mee Ling thinks that pupils who come to school by bus are more likely to be late than those who do not travel by bus. In order to test whether or not this is true, she carries out a survey on 100 pupils, from years 7 and 8, for 5 consecutive Tuesdays. The results are shown in the following table.

<table>
<thead>
<tr>
<th>METHOD OF TRAVEL</th>
<th>NUMBER OF STUDENT-DAYS</th>
<th>NUMBER OF LATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>150</td>
<td>40</td>
</tr>
<tr>
<td>Cycle</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Car</td>
<td>100</td>
<td>22</td>
</tr>
<tr>
<td>Walk</td>
<td>200</td>
<td>25</td>
</tr>
<tr>
<td>TOTAL</td>
<td>500</td>
<td>97</td>
</tr>
</tbody>
</table>

(a) Do the results show that Mee Ling is correct? Show the working on which you base your answer.
(b) Suggest 3 ways in which Mee Ling could have improved her survey.
(c) A pupil is selected at random from Year 7. Mee Ling stated:

"The table shows that the probability that this pupil walks to school on Tuesdays is \( \frac{2}{5} \)."

Would you describe this statement as correct, or about right, or wrong? Choose one of the three alternatives and give reasons for your choice. (MEG)

11. (a) A headline in a newspaper this year stated:

Students skip Breakfast

Our survey shows that few students are eating cereals, fruit, or bread for breakfast.

In fact they are eating nothing at all!

You are asked to conduct a survey to find out what students eat for breakfast.
Design an observation sheet to collect the data you need.

Invent the first 20 entries on your data sheet.

(b) The newspaper made the following statement about the eating habits of teenagers.

Only one in a hundred teenagers eat fruit and vegetables each day. Over half eat no vegetables other than chips.

You are asked to find out whether this statement is true in your area.

Give three questions you could ask teenagers to see if what the article says is true in your area.

(NEAB)

12. A bus company attempted to estimate the number of people who travel on local buses in a certain town. They telephoned 100 people in the town one evening and asked 'Have you travelled by bus in the last week?'

Nineteen people said 'Yes'. The bus company concluded that 19% of the town's population travel on local buses.

Give 3 criticisms of this method of estimation.

(MEG)
1.4 Experimental Design

**Paired Comparisons**
The simplest experimental design is the use of *paired comparisons*. For example, to compare the weight loss due to two different slimming diets, an ideal design would be to get the cooperation of several pairs of identical twins. One twin of each pair would follow one diet and the other twin the other diet. Thus experimental error due to differences in the people on the different diets would be minimised.

**Control Groups**
Sometimes there are not two or more treatments, but only one. For example, we may wish to find the effect of a particular medical treatment on arthritis or the effect of a coaching course on a student's tennis playing skills. The effect of these treatments cannot be judged in isolation. An arthritis sufferer may improve (or deteriorate) with no treatment. Similarly, a tennis player may improve without attending a coaching course.

It is necessary to have a *control group* and an *experimental group*. These two groups should be matched as closely as possible. That is, the people in one group should be as similar as possible to the people in the other group as far as characteristics relevant to the investigation are concerned. This does not mean that all the people in a particular group must be similar to each other, but that the group as a whole must be similar to the other group.

For example, in the case of the arthritis sufferers the two groups should contain people of similar age, sex, general health and severity of arthritis. In the case of the tennis players, the groups should contain students of similar age, sex, fitness and tennis playing ability.

The groups should be selected and then one group should be chosen at random to receive the experimental treatment (or tennis coaching) and the other group will be the control group. The control group will receive no treatment (or coaching) or will continue with the standard treatment. The effect on the two groups can then be compared.

**Placebos and Blind Trials**
In the case of medical treatment, it is sometimes thought that patients will improve or recover without treatment and that in some cases this improvement will be greater or quicker if they are told they are having treatment, even when they are not. Thus it is standard practice in drugs tests to give the control group a *placebo*. This is a harmless substance which looks like the real medication but, in fact, does not contain any drug. Many patients will improve after taking placebos. To show a drug to be effective, significantly more patients who took the drug must show improvement than those who took the placebo. (There are, of course, other issues such as possible side effects to consider as well.)

If the patients who took the placebos knew that they were taking placebos the effect would be lost. It is essential that the patients should not know whether or not they are taking placebos and this is known as a *blind trial*.

Even more subtle effects can be at work. It has been found that, even if the patients do not know whether or not they are taking placebos, the doctor may expect those patients taking the drug to fare better than those taking placebos. This expectation may somehow transmit itself to the patient whose condition may improve as a result. It is therefore, necessary that the doctor does not know which patients are receiving placebos and which are receiving the drugs. Of course, someone must know who is receiving the drugs otherwise it would be impossible to analyse the results. However, it should be someone
who has no direct contact with the patient. Trials where neither the patient nor the doctor know who is receiving the drugs are known as double blind trials.

It has been suggested that the person carrying out the statistical analysis should also not know which patients took the drug to prevent this influencing the analysis. This would be described as a triple blind trial.

**Worked Example 1**

A new brand of ‘slim-fast’ milk has been introduced for sale into a store. It is claimed that users will achieve significant weight loss after using the product for a period of seven consecutive days.

A statistical experiment is to be set up to test this claim.

The experiment will involve using 50 members each in both experimental and control groups. These participants are to be selected from the first 200 shoppers entering the store on a given day.

Explain briefly

(a) why a control group should be used in this case,

(b) how members of the control and experimental groups should be selected if paired comparisons are to be made,

(c) what procedures should be followed to ensure valid conclusion are reached from this experiment.

*(NEAB)*

**Solution**

(a) If there was no control group, you could not be certain that any significant weight loss was actually due to the new brand of milk.

(b) For paired comparisons, the pairs must be selected to be as alike as possible in all ways, e.g. weight, age, family circumstances, income, etc; ideally they should all be identical twins!

(c) The experimenter must ensure, as far as possible, that the control group members do *not* use the new brand of milk or undertake any other special weight loss programme during the week. Similarly, the experimental group must use the same brand of milk but also not undertake any other special weight loss programme beyond what they would normally do daily. The experimenter would also need to ensure that the weights of each group member were accurately measured, under the same conditions each time.
Exercises

1. **DAILY RAG**

   WINE MAKES YOU TALLER – IT’S OFFICIAL

   Yesterday we chose 100 women at random and asked them if they drank wine. Fifty said they did and fifty said they didn't. The average height of the 50 wine drinkers was 2 cm more that the average height of the non-drinkers. Positive proof that wine makes you taller!

   Explain why this is *not* proof that drinking wine makes you taller.

2. A teacher has a jar containing only red jelly babies. She wants to find out approximately how many red jelly babies there are.
   She puts 20 black jelly babies in with the red ones and mixes them thoroughly.
   She then pulls out 12 jelly babies without looking. 1 is black and 11 are red.
   (a) Estimate the number of red jelly babies in the jar.
   (b) She puts the 12 jelly babies back in the jar and mixes them in.
       She then pulls out 40 jelly babies without looking. 4 of them are black and the other 36 are red.
       Estimate the number of red jelly babies in the jar.
   (c) Which estimate do you think is more reliable? Give a reason.
   (d) Explain how you could estimate the number of fish in Lake Windermere using a similar method.

3. (a) A gardening magazine claims that woodlice prefer damp conditions.
   How would you test this claim using a simple statistical experiment?
   (b) The magazine also claims that woodlice prefer dark and damp conditions.
   How would you test this claim?
1.5 Sampling

Census and Population

Every 10 years (since 1801) the Statistics department (website www.statistics.gov.uk/), previously known as the Office of Population Census and Surveys, carries out a census for the government which involves the whole population of the country. (*Census* is a Latin word meaning 'registration of citizens'.)

When conducting a survey it is often impossible to ask every individual who might be concerned or involved. For example, for a political opinion poll it is only necessary to ask a sample of the population how they would vote.

The term *population* can be any group about which information is required. For example, the following could be populations:

- Manchester United supporters,
- kettles produced in a factory,
- adult voters in the UK,
- all pupils in your school.

Samples and Bias

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. This is an example of a sample.

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 items to be used in the sample to be chosen to avoid introducing bias?

(c) How is the data to be collected to avoid 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. For example, if you are a political pollster wanting to know the state of the parties, then a sample taken only in an affluent, expensive out-of-town suburb will be biased just as a sample taken in a run-down inner city area will be. So the sample must not be biased – but the same also applies to the questions. Questions such as

"Should we increase spending on education?"

are biased as they do not represent the complete picture. If we spend more on education, this means that we must spend less on something else. A less biased question would be,

"Should we spend more on education and less on defence?"

Representative Samples

When we sample we are looking for methods that lead to a representative sample, by which we mean that the sample has similar properties to the total population. As we said before, here we use the word 'population' not just for human or animal populations but to represent the complete group under study, e.g. all schools in the UK, all cars in England, etc.
Sampling Methods

Often, conclusions are reached by looking at a sample taken from a population. The main methods used for sampling in practice are described below.

Random Sample

The sample is formed by selecting members of the population at random. It is important to make sure that each member of the population is equally likely to be selected. Tables of random numbers can be used to help this process but more mundane methods, such as choosing a number from a hat of numbers, can be used. Scientific calculators also provide you with random numbers.

To be truly random each individual must have an equal chance of being chosen. 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). However, random sampling does not necessarily ensure a representative sample.

Systematic Sample

This type of sample is formed by taking members of the population at regular intervals. For example, by selecting every 5th or every 10th or every 12th member of the population or every 4th tree when sampling in a forest. 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.

Stratified Sample

The population is split into a number of groups. Random samples are then taken from each group so that the ratio of the sizes in the sample is the same as the ratio of the number of members of the groups in the population. For example, if a population contains 1000 women and 500 men, a stratified sample of total size 75 would contain 50 women and 25 men. A stratified sample is used to try to ensure that the sample is representative by ensuring that it has the same proportions as the population. To do this you first need to divide the whole of the population into appropriate categories and then, inside each category, choose a random sample of appropriate size. This can be difficult in practice if the sample has to be stratified in several ways, e.g. age, gender, ethnicity, income, etc.

Other sampling methods include quota sampling and cluster sampling.

Worked Example 1

The diagram on the following page shows 57 small fish. To estimate the average length of a fish (that is, the length in mm from the tip of its 'nose' to the middle of its tail, as shown in the diagram on the right) take random samples of size

(a) 5  
(b) 10.

Which will give you the more accurate estimate?

Solution

Here there are no obvious different categories so we will take a random sample. A table of random numbers is given separately.
Diagram for Worked Example 1
Choosing a random start point, in this case row 10, combining two digits together gives numbers from 00 to 99. We use only numbers in the range 01 to 57 (and ignore any repeats), so that each fish has an equal chance of being picked.

Thus our sample of size 5 consists of the
{57th, 21st, 12th, 40th, 25th} fish.

The lengths of these fish are measured, to the nearest mm, as
2.0, 2.1, 1.8, 3.0, 2.0

giving an estimate of
\[ \frac{2.0 + 2.1 + 1.8 + 3.0 + 2.0}{5} = 2.18 \text{ mm} \]

for the average length of all 57 fish.

For a sample of size 10, we continue in the same way with the random number table, choosing as our sample, the
{57th, 21st, 12th, 40th, 25th, 39th, 22nd, 49th, 43rd, 54th } fish

and these have an average length of
\[ \frac{2.0 + 2.1 + 1.8 + 3.0 + 2.0 + 3.7 + 1.7 + 1.8 + 3.1 + 2.7}{10} = 2.39 \text{ mm} \]

We would expect this second estimate of 2.39 mm to be more accurate than the first one since it is obtained from a larger size sample.

\textbf{Note 1}

Using a different set of random numbers will almost certainly produce a different sample. That is why two statisticians investigating the same problems can sometimes reach different conclusions.

\textbf{Note 2}

Random sampling can also be done using the random number generator on a calculator.

\textbf{Worked Example 2}

A headteacher wishes to select a stratified sample of 50 pupils from Years 10, 11, 12 and 13. The table shows how many students are in each year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of pupils</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>320</td>
</tr>
<tr>
<td>11</td>
<td>300</td>
</tr>
<tr>
<td>12</td>
<td>180</td>
</tr>
<tr>
<td>13</td>
<td>150</td>
</tr>
</tbody>
</table>
Solution

First find the total number of pupils in the population.

\[ \text{Total} = 320 + 300 + 180 + 150 = 950 \]

The fraction of the population in Year 10 is \( \frac{320}{950} \).

So the number of pupils selected from Year 10 is given by:

\[ \frac{320}{950} \times 50 = 16.84, \quad \text{so select 17 pupils.} \]

The number of pupils selected from Year 11 is given by:

\[ \frac{300}{950} \times 50 = 15.79, \quad \text{so select 16 pupils.} \]

The number of pupils selected from Year 12 is given by:

\[ \frac{180}{950} \times 50 = 9.47, \quad \text{so select 9 pupils.} \]

The number of pupils selected from Year 13 is given by:

\[ \frac{150}{950} \times 50 = 7.89, \quad \text{so select 8 pupils.} \]
Exercises

1. (a) Paul is carrying out a survey to find the most popular colour for a vehicle. He finds the colour of his mother’s car and two neighbours’ cars. Give two reasons why this sample may give a poor result.

(b) Give two advantages of sampling. \((SEG)\)

2. Natasha was asked to find an estimate of the mean height of adult women. To save time she found the mean height of her 10 aunts. Give two reasons why this sample may not give an accurate result. \((NEAB)\)

3. An interviewer stopped people in a local shopping centre to ask their views on abortion. The age and sex of the people asked is shown.

<table>
<thead>
<tr>
<th>Age</th>
<th>under 18</th>
<th>18 – 25</th>
<th>26 – 40</th>
<th>41 – 64</th>
<th>65 and over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>0</td>
<td>5</td>
<td>17</td>
<td>20</td>
<td>8</td>
</tr>
</tbody>
</table>

Give two reasons why the views of this sample may not be representative of the whole population.

4. A television producer wished to find out how popular a new television series was. To find out she sent letters to a random sample of 1000 viewers. 400 viewers replied. 120 said they liked the series, 280 said they disliked it.

(a) How many viewers did not reply to the letter?

(b) What is the greatest possible percentage of the 1000 people sampled who (i) could have liked the series? (ii) could have disliked the series?

(c) The producer talks to her friend who tells her that more reliable results can be obtained by using larger samples.

In this case why would it not be a good idea to send out a larger number of letters?

(d) How would you attempt to get a reliable answer? \((NEAB)\)

5. In a mixed comprehensive school of 500 pupils it was decided to produce a monthly index of leisure expenditure based on a sample of school pupils.

(a) Using a random number table, explain briefly (i) how you would select a simple random sample of 10 pupils from the school without replacement.

(ii) how you would select a systematic sample of size 10 from the school.

(b) (i) Suggest two types of strata into which the pupils could be classified for sampling purposes.

(ii) Explain why it would be preferable to introduce stratification into the sampling procedure. \((NEAB)\)
6. Graham wants to find out if the pupils in his school are in favour of having a school uniform. He considers these methods of choosing a sample of 60 pupils.

*Method 1:* Choose the first 60 pupils arriving at school in the morning.

*Method 2:* Choose 60 names at random from the list of pupils in the school.

*Method 3:* Choose 60 pupils he knows.

(a) Which method is *most* likely to produce an unbiased sample? Give a reason for your answer.

(b) Explain why the other two methods could produce biased samples.

Graham asked the pupils the following question:

"*Don't you think school uniform shouldn't be worn?*"

(c) Explain why this question is *not* suitable.

(d) Write a suitable question. \(\text{(AQA)}\)

7. A market research company wishes to interview a sample of 100 adults in a large town in order to obtain their views on the proposed construction of a by-pass around the town.

The person assigned to select the sample decided to ask the first 100 adults leaving the local railway station after 16.30 hours on Thursday 18th May 1999.

(a) List *three* reasons why this method of sample selection would be unsuitable.

(b) Suggest how a random sample of 100 adults could be drawn from this town.

(c) In each of the following cases sampling is necessary for different reasons. State briefly the reasons in each case.

(i) Testing the lifetime of car batteries.
(ii) Conducting a national opinion poll prior to an election.
(iii) Carrying out periodic checks on engineered items as they are being manufactured. \(\text{(AQA)}\)

8. In a large school, a group of pupils decides to produce a monthly index of expenditure based on the spending habits of pupils within the school.

They agree to base the index on data collected each month from the same sample of pupils.

(a) Explain why stratified sampling rather than simple random sampling would be more appropriate in this case.

(b) (i) The students decided to stratify the sample by gender. Why is this sensible?

(ii) Twenty boys are to be included in the sample. Explain how they should be selected. \(\text{(AQA)}\)

(a) Create a systematic sample containing:

(i) 8 machines, (ii) 6 machines, (iii) 4 machines.

(b) Use the random numbers in the table to create random samples containing:

(i) 5 machines, (ii) 10 machines.

(c) Describe another way that a random sample could be made.

<table>
<thead>
<tr>
<th>Random Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>66 06 42 35 06 15</td>
</tr>
<tr>
<td>50 96 24 21 82 53</td>
</tr>
<tr>
<td>55 40 32 46 74 76</td>
</tr>
<tr>
<td>06 57 85 48 20 92</td>
</tr>
<tr>
<td>28 74 44 73 44 25</td>
</tr>
<tr>
<td>65 08 56 68 45 74</td>
</tr>
<tr>
<td>63 77 87 58 00 09</td>
</tr>
<tr>
<td>13 26 76 76 81 60</td>
</tr>
<tr>
<td>80 51 89 59 38 10</td>
</tr>
<tr>
<td>27 22 87 55 80 52</td>
</tr>
</tbody>
</table>

10. The table shows the number of students in each year group of a school. How many students should be selected from each year group to create a stratified sample of 80 students?

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>148</td>
</tr>
<tr>
<td>9</td>
<td>162</td>
</tr>
<tr>
<td>10</td>
<td>154</td>
</tr>
<tr>
<td>11</td>
<td>152</td>
</tr>
<tr>
<td>12</td>
<td>80</td>
</tr>
<tr>
<td>13</td>
<td>62</td>
</tr>
</tbody>
</table>

11. A company wants to form a stratified sample to discuss issues with the staff.

(a) How many of each type of employee should be included in a sample of size 20?

(b) If a stratified sample contains 8 manual staff, what would be the size of the complete sample?

<table>
<thead>
<tr>
<th>Number employed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Managers</td>
</tr>
<tr>
<td>Supervisors</td>
</tr>
<tr>
<td>Administrators</td>
</tr>
<tr>
<td>Manual Staff</td>
</tr>
<tr>
<td>Delivery Staff</td>
</tr>
</tbody>
</table>

12. In order to form a sample a number of pupils are selected from a number of different teaching groups. The size of the samples and the teaching groups are given in the table below.

<table>
<thead>
<tr>
<th>Group</th>
<th>Number in Group</th>
<th>Number in Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>18</td>
<td>4</td>
</tr>
</tbody>
</table>

How should the sample sizes be changed to give a stratified sample?
13. To test a hypothesis in Biology a square metre of ground has been divided into 100 squares as shown.

You are required to create a sample of 10 squares.

(a) Describe two ways of creating a systematic sample, giving the results from each method.

(b) Use the following list of random numbers to create a random sample.

```
11 57 42 70 41 12 88 51
32 38 14 65 49 46 18 62
48 20 47 44 56 65 46 36
42 59 61 41 78 42 72 11
50 42 58 78 71 78 12 57
66 08 21 84 94 61 31 30
14 64 51 05 53 93 45 86
93 49 05 27 54 18 64 57
87 61 55 67 23 26 70 75
03 87 19 48 10 69 35 61
42 52 83 74 35 09 13 36
51 43 76 62 91 39 89 75
```

Create a second random sample, by starting in a different place in the list of random numbers.

How many squares do the two samples have in common?

14. What problems might be encountered if samples are formed in the following ways?

(a) Selecting people at random from a telephone directory.

(b) Selecting every third person entering a shopping arcade.

(c) Selecting people at random at a football match.

(d) Selecting people leaving a golf course in red cars.

15. For each of the samples in question 14, suggest a hypothesis that might be being tested.

16. Describe how you might select a sample if you were asked to conduct a survey to test the hypotheses:

(a) parents of primary school children are happy with their school,

(b) users of mobile phones are happy with the service provided,

(c) the local bowling green is properly maintained,

(d) the local public transport services are adequate.