3. Collecting Data[1]
3.1. Introduction

3.1.1. Why collect data?

There is a difference between data and information. Essentially, data are the raw numbers or facts which must be processed to give useful information (Figure 3.1). Thus 78, 64, 36, 70 and 52 are data which could be processed to give the information that the average mark of five students in an exam was 60%; data about new business performance could be collected as a large set of numbers, and this could be processed to give the information that two thirds of new companies cease trading within two years of opening; the ten-year government census has individual returns providing data, which are processed to give information about the population as a whole; entries in a company's transaction records give data which are consolidated into accounting information; and so on.

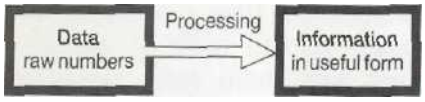

Figure 3.1 Relating data and information.

In this book we emphasize quantitative data, and these appear as sets of numbers. We should recognize, though, that data can be any collection of facts, observations, measurements, opinions, or anything else which gives details about a situation.

In business, data are collected and transformed into the information which allows managers to make their decisions. There are several implications in this statement:

- managers need information before they can make decisions
- they should examine all available information before making decisions
- they should have enough information to allow good decisions
- this information should be reliable
- the information is provided by data collection and analysis

It is clear, then, that data collection is an important, and even vital, function in any organization. Without data collection, managers do not have access to reliable information and cannot make reasoned decisions. The rest of this chapter discusses ways in which data can be collected.

IN SUMMARY

Managers need reliable information to make decisions about the running of their organizations. This information is provided by data collection and processing.

3.1.2. Timing and quantity of data collection

You can see that we are discussing the collection of data before discussing their presentation (which is covered in Chapter 4). This seems a sensible approach, as you have to collect data before presenting them. However, data are collected for a specific purpose and the way they are used should have an effect on the way they are collected. If, for example, we want to decide how many beds a hospital should set aside for road accident victims, we could collect data from local accident statistics; if we want to know how retired people spend their leisure time, we could enclose a questionnaire with information which is routinely posted to pensioners; if we want to find how many people will buy a new product, we could run a market survey; if we want to see how many people

[1] Waters, D., Quantitative Methods for Business, Addison – Wesley, 1993

wear car seat belts, we could stand by a road and observe passing cars. In other words, the way data will be used has an effect on the way they are collected. We should, therefore, design data collection to meet its specific purpose, and not the other way around.

Data collection should be designed after deciding the use of the data.

One problem with data collection is knowing how much to collect. In many circumstances there is an almost limitless amount of data which could be collected and might be useful. We should resist the temptation to collect data simply because they are available, and limit ourselves to those which are relevant and useful. The reason for this is that data collection and processing inevitably costs money and collecting unnecessary data is wasteful.

In principle there is an optimal amount of data which should be collected for any purpose. If we consider the marginal cost of data as the cost of collecting the last 'unit', then the marginal cost increases with the amount of data collected. We could find some general data about, say, British Rail very easily (it operates trains and stations, employs staff, and so on); more detailed data would need a trip to a specialized library (to find exactly how many trains of different types are operated or staff of different grades employed); yet more detailed data would need a search of British Rail's own records (to find the wage bill for each grade of employee in each region); yet more detailed data would need a special survey (to ask what each grade of employee felt about their conditions), and so on.

More detailed data are clearly more difficult and more expensive to collect. Conversely, the marginal benefit of data (which is the benefit of the last 'unit' of data collected) is likely to fall. Using the above illustration, the fact that British Rail runs a rail service is a very useful item of data, but most people would find the views of different grades of employees about their conditions less useful. We could use this observation to suggest the relationship shown in Figure 3.2.

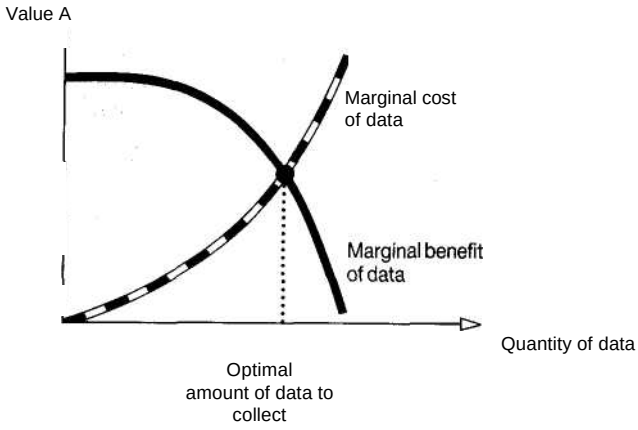

Figure 3.2 Finding the optimal quantity of data to collect

An optimal amount of data collection can be defined by the point at which the marginal cost becomes greater than the marginal benefit. In other words, the cost of collecting another bit of data is greater than the benefit. Collecting more data would be wasteful, but collecting less would lose some potential benefit. In practice, of course, the problem with this analysis lies in the difficulty of defining the costs and benefits of the data collected. This means that the optimal amount of data is not usually calculated, but is suggested in the light of previous experience.

Another factor which is important in data collection is the time available. Some methods of collection, such as reviewing published statistics, can be done very quickly; others, such as running consumer surveys, need a lot of time. The time available can limit both the type of data that can be collected and the amount. If a company decides to launch a new product next year, this automatically sets a limit to the amount of consumer advice that can be collected. Similarly, government-funded research projects are often limited to three years, so any project which aims to follow economic progress over, say, ten years automatically becomes infeasible.

A long period of data collection may also make the results either irrelevant or out of date before they can be properly analysed. It is often said that the 1851 census in America was not properly analysed before the 1861

census was taken.
On a smaller scale it would be pointless to spend so long collecting data about the sales of a product that it was withdrawn from the market before the analysis could be completed.

If there are pressures on the time available for collecting data, or if proper planning is not done, there is a possibility that mistakes might be made. It is a common view that some data, even if they are slightly inaccurate, are better than no data at all. In some circumstances this is certainly a valid opinion. If we are buying a car it is better to ask a salesman for some details, even though we know that the replies may not be entirely accurate. In many circumstances, however, wrong data can be worse than no data at all. A car salesman might mistakenly persuade us that an expensive car is cheap to run, and we might end up being unable to meet the payments. More broadly, inaccurate data may be so misleading that managers make decisions which are not the correct ones and which might actually harm an organization rather than give benefits. The obvious conclusion is that we must ensure data collected are as accurate as possible.

IN SUMMARY

Data collection is expensive, so it is sensible to decide what the data will be used for before they are collected. In principle, there is an optimal amount of data which should be collected. These data should be as accurate as possible.

Self-assessment questions

3.1 What is the difference between data and information?

3.2 Why is data collection important for an organization?

3.3 'It is always best to collect as much data as possible about a situation.' Is this statement true or false?

3.2. Types of data

Data of different types are collected in different ways. The weights of packages coming off an assembly line can be measured directly; the number of customers in a shop can be found by observation; the efficiency of a service can be found by giving customers a questionnaire; the age of a population can be found in statistics published by the government; and so on. We should, then, start by describing different types of data.
Data can be classified in several ways. One classification we have already discussed defines data as either qualitative or quantitative. The collection, presentation and analysis of quantitative data is much easier and more precise so we should, wherever possible, use them in preference to qualitative data. Even data which are essentially qualitative can be given a quantitative form. Everybody has some opinion about a range of political questions, for example, and although it is impossible to **measure** each of these opinions, we can say '70% of people generally agree with this policy' or '60% support the policies of this party'. Sometimes a scale can be added. Doctors, for example, may want to know how bad a patient's pain is. This is impossible to measure, but the patient can be asked to rank it on a scale of 0 to 10, where 0 corresponds to no pain at all and 10 is the worst pain it is possible to imagine. Then if a patient says they have a pain of about 8 a doctor will know that it is serious and needs immediate attention.
Unfortunately, the use of numbers to express qualitative ideas needs considerable care. When we hear 'Eight out of ten dogs prefer' a particular kind of dog food we should think carefully about what this means and compare it with the less positive statement, 'In a limited test eight out of ten owners who expressed an opinion said their dog seemed to prefer this dog food to an alternative.'
Even so, not all data can be transformed into a convincing quantitative form. When we hear a poet ask, 'How do I love thee? Let me count the ways...' we know that this is more for effect than for realism.
An extension of this basic classification of data describes how well they can be measured. This describes data according to:

nominal

ordinal cardinal

3.2.1 Nominal data

This is the kind of data which really cannot be quantified with any meaningful units. They are sometimes referred to as categorical data. The fact that a company is a manufacturer, or a country operates a centrally planned economy, or a cake has cream in it, are examples of nominal data.

A common analysis for nominal data defines a number of different categories and says how many observations fall into each. Thus a survey of companies in a particular area might show that there are 7 manufacturers, 16 service companies and 5 in primary industries. This says nothing about companies' sizes, profits, owners and so on, and it does not matter in

which order the categories are taken. A common example of nominal data comes from political polls, which typically show that 40% of respondents would vote for political party X, 35% for party Y, 20% for party Z, and 5% do not know.

3.2.2 Ordinal data

Ordinal data are one step more quantitative, in that the categories into which observations are divided can be ranked in some order. Sweaters, for example, may be described as extra large, large, medium, small or extra small. Describing a sweater as 'medium' tells us something, but really gives little quantifiable information. In essence we are told, 'A medium sweater is smaller than a large one but larger than a small one'. Similarly, consumer surveys often collect ordinal data by asking questions like, 'Say how strongly you agree with this statement on a scale of 1 to 5 where 1 means strongly agree and 5 means strongly disagree', while sociologists classify people according to A, B, Cl, C2 and D. The essential characteristic is that data can be put into different categories, and that the order of these categories is important. Sometimes, when there are few observations, they can all be ranked individually rather than put into ranked categories. Thus horses are ranked in a race, as are applicants for jobs, students' performance in courses, consumers' preferences between competing products, and so on.

Cardinal data

Cardinal data have some attribute which can be directly measured. Thus we can weigh a bag of chocolates, measure the time to perform a task, find the temperature in an office, and so on. These measures give a precise description of a particular characteristic.

Sometimes it is useful to group observations which are similar, and then a common analysis for cardinal data defines different categories according to direct measurements. Thus a sample of basketball players might have twelve who are between 6 ft and 6 ft 2 in tall, eight who are 6 ft 2 in to 6 ft 4 in tall, and three who are 6 ft 4 in to 6 ft 6 in tall.

Cardinal data are generally the easiest to analyse and are the most relevant to quantitative methods. Cardinal data can be divided into two types depending on whether they are discrete or continuous.

Discrete data

Data are discrete if they can only take integer values. When asking the number of children in families, the answer will be 0, 1, 2 or some other integer number. Similarly, the number of cars owned, machines operated, shops opened and people employed are discrete data which can only come in integer quantities.

Continuous data

Those measures which can take any value and are not restricted to integers are called continuous. Thus the weight of a bag of biscuits is continuous as it can be any value, such as 256.312 grams. Similarly, the time taken to perform a task, the length of metal bars, the area covered by a carpet and the height of flagpoles are continuous data.

Sometimes there is a mismatch in data types. The circumferences of men's necks, for example, are continuous data, but shirt collars use a discrete measure; feet come in any size, but shoes come in a range of

discrete sizes which are good enough for most needs; heights are continuous but most people describe their height to the nearest inch or centimetre. If the units of measurement are small, the distinction between discrete and continuous data begins to vanish. Salaries, for example, are really discrete as they must be multiples of a penny. Normally, however, they can be considered continuous because the units are small in relation to the values measured.

Finally, there is one more classification of data which is directly related to the method of collection. If an organization wants to use some data for a particular purpose, it may use either primary or secondary data:

- **primary data** are collected by the organization itself for the particular purpose
- **secondary data** are collected by other organizations or for other purposes

Any data which are not collected by the organization for the specified purpose are secondary data. These may be published by other organizations, available from research studies, published by the government, and so on. They may also be collected by the organization itself for another purpose.

The benefit of primary data is that they fit the needs exactly, are up to date and are reliable. Secondary data have the advantages of being much cheaper and faster to collect. They also have the benefit of using sources which are not generally available: companies will, for example, respond to a survey by the government, the Confederation of British Industry, or a group of students, but they would not answer questions from another company.

If adequate secondary data are available they should be used. Unfortunately, there is often not enough appropriate, up-to-date secondary data for a particular purpose. Then a balance must be drawn between the benefits of primary data and the cost of obtaining them. If a company is about to launch a new product it will run a market survey to collect primary data and gauge customer reactions; if it wants to evaluate general economic activity in an area it will use secondary data prepared by the government. Sometimes a combination of primary and secondary data is used, perhaps using secondary data to give the overall picture, and then adding details from primary data. In any case, it is sensible to survey secondary data first, and then consider primary data for extensions and clarifications.

IN SUMMARY

Data of different types can be collected in different ways. **There are several** classifications of data, including quantitative/qualitative, **nominal/ordinal/** cardinal, discrete/continuous and primary/secondary.

Self-assessment questions

3.4 Why is it useful to classify data?
3.5 How might data be classified?
3.6 What is the difference between discrete and continuous data?
3.7 Give examples of nominal, ordinal and cardinal data.

3.3. Sampling methods

3.3.1 Purpose of sampling

If appropriate secondary data are not available, an organization must use primary data. Then there are several ways of collecting them. Most are based on the assumption that data will be collected by sampling. In other words, data are collected from a representative sample of items or people, and these are used to infer characteristics about all items or people. Suppose, for example, a company is about to launch a new product and wants some data about likely sales. There are two ways of finding this:

- it could ask every person in the country who might buy the product whether they actually will buy it, and if so how much they would buy
- it could take a sample of people, ask them how much of the product they will buy, and then estimate the likely demand from the population as a whole

The first of these approaches (which is called a **census)** has the obvious disadvantage of being time-consuming and expensive. The second approach (which uses **sampling)** has a number of advantages, the most obvious being the reduced cost and time. Another consideration is the practicality of collecting data from an entire population. How, for example, could you find the views of everyone living in southeast England when at any time some people are sick, others are on holiday or travelling, some will refuse to answer questions, and so on?

The purpose of sampling is to get reliable results using only a sample of the whole population. Notice that we are using **population** in its statistical sense of a set of items which share some common characteristics. For data collection the population is the set of all items or people which could supply data. Suppose the Post Office want to find how long it takes to deliver first-class letters; then the population is all letters which are posted first-class. Similarly, a toy manufacturer getting reactions to a particular game might define the population of potential customers as all girls between the ages of 10 and 14; a consumer organization wanting to test the quality of a product would define the population as all units of the product which have been made; a bus company testing the reliability of a bus service would define the population as all journeys that their buses make.

When collecting data it is important to identify the proper population which could supply the data. This is not always as easy as it seems. The toy manufacturer above, for example, may find its population should also include boys aged 10 to 14. A survey of student opinion about a particular government policy would have a population which is clearly students, but does this mean full-time students only, or does it include part-time, day-release and distance-learning students? What about students doing block-release courses during their period of work, school students and those studying but not enrolled in courses? Care must be taken in identifying the correct population, because a mistake at this stage will make the remaining analysis useless.

Even when a population can be identified in principle there may be difficulties in practice. If a population is identified as all people who bought a foreign car within the last five years, or all people who use a particular supermarket, how could a list of such people actually be found? In some cases this is relatively straightforward. If the population is houses with telephones, they are easy to identify from entries in telephone directories. Such lists of the population are called **sampling frames,** and are often given by electoral registers, association membership lists (such as the Automobile Association), credit rating agencies, or specialized companies who prepare lists of people with specified characteristics. Unfortunately, sampling frames are often not available and then some other method of identifying a sample must be used.

We have said that the purpose of sampling is to take a sample of units from the population, collect data about the desired property and use this to estimate data for the population as a whole. Then the population is all items or people which **could** give data, while the sample is those items or people which **actually** give data. What we need to discuss now are:

- a means of determining a suitable sample size (large enough to be representative of the population but small enough to be practical and cost effective)
- a method of selecting this sample

The problem of determining a sample size is considered in detail in Chapter 15, but the next section describes some methods of selecting the sample

IN SUMMARY

Data collection often uses sampling, where data from a sample are used to estimate data for the population. This is done when data collection from the entire population would be too expensive, time-consuming or impractical.

3.3.2 Types of sample

Samples can be selected in several ways and we can classify these according to:

census
random sample systematic sample
quota sample stratified sample
multi-stage sample cluster sample

Census

If the population is small and the results are important it may be worth doing a census, where data are collected from every member of the population. Then the sample is the same as the population. The UK Government carries out a population census of this kind every ten years.

The benefit of a census is that very accurate data can be obtained. Although the data may not be completely accurate (as there will still be errors and omissions), they are as accurate as possible. Unfortunately, the cost and time needed for a census are prohibitive in all but a few investigations and this means that a smaller sample is usually used.

Random sample

If a census is not taken we have to find a sample which accurately represents the population as a whole. The easiest way of arranging this is to take a **random sample.** The essential characteristic here is that every member of the population has exactly the same chance of being selected for data collection. We should emphasize that if a sample is random it does not mean that it is disorganized or haphazard. If we were collecting data about the contents of tinned soup, we could simply go to a supermarket and buy the first dozen tins of soup that we saw. This would be haphazard, but it would certainly not be random.

There are several ways of selecting random members of a population. A small club could ask each member to write their name on a piece of paper and put this in a hat. Then picking one piece of paper from the hat would give a random selection. Most random sampling is more complicated than this and needs a more formal approach, but this must still ensure that every member of the population has an equal chance of being picked. A common way of organizing this is to use random numbers.

Random numbers are simply a string of random digits, as shown in Appendix D. Traditionally these have been prepared in tables but now they are almost invariably generated by computer. Most computer languages have a function such as RND or RAND, which automatically generates random digits.

Suppose we want to collect data from a random sample of people visiting an office. It might be too disruptive and impractical to take a census of visitors, so we could take a sample which is selected using random numbers. If we generate a series of random digits, 5 4 6 1 5 3 1, we could stand by the office door and interview the fifth person to pass, then the fourth person after that, then the sixth after that, then the first after that, and so on.

Random numbers can give totally random samples. This has a major benefit when using statistical analyses, most of which are only valid if the sample is genuinely random. Unfortunately, some samples which appear to be random are not. Suppose we decide to save time and simply write down a series of numbers which looks random. The series will probably not really be random, as most people have preferences - perhaps for even numbers, or for sequences which are easy to type on numerical keyboards. Similarly, if interviewers are asked to select people at random they will inevitably give a biased sample; they are more likely to approach people they find attractive, and to avoid people they find unattractive, or

very tall people, people in an obvious hurry, or people in groups.
A well-organized random sample will ensure that, in the long run, the sample is representative of the population as a whole. If a sample does not exactly reflect the population, it is said to be **biased** in favour of one section. Unfortunately, random samples must be fairly large, as small samples can contain atypical results and show bias. Exactly how large a random sample should be is discussed in Chapter 15. Even so, a well-organized and relatively large sample could, by chance, give atypical data. This can be avoided by using some form of structured or non-random sample. These try to find results of equivalent accuracy but with a smaller sample

WORKED EXAMPLE 3.1

A company receives 10 000 invoices in a financial year. An auditor does not have time to examine each of these, so takes a random sample of 200. How might this sample be organized?

Solution

The first thing to do is to form the sampling frame by listing the invoices and numbering them 0000 to 9999. Then we generate a set of 200 random numbers, each with four digits. One set is:

4271 6845 2246 9715 4415 0330 8837 and so on

Then we select invoices numbered 4271, 6845, 2246, and so on, as a completely random sample.

Systematic sample

Perhaps the easiest way to organize a non-random sample is to collect data at regular intervals. Then every tenth unit from a production line might be weighed, or every twentieth person using a service. The essence of a random sample is that every member of the population has the same chance of being chosen. If, say, every tenth member is chosen, this means that members 11, 12, 13 and so on have no chance of being selected and the sample is not random. In practice, a systematic sample is almost always acceptable as being random, or at least 'pseudo-random'. There are, however, occasions when the regularity introduces bias. Checking the contents of every twentieth bottle filled in a bottling plant may be invalid if every twentieth bottle is filled by the same head on the filling machine; collecting data from every thirtieth person leaving a bus station may introduce bias as buses hold an average of about 30 people, so we may always be interviewing the older and slower people, who get off a bus last.

WORKED EXAMPLE 3.2

A production line produces 5000 units a day. Quality control checks are needed on *2%* of these. How could a systematic sample identify these?

Solution

The number of samples a day is 2% of 5000, which is 100. A systematic sample would check every hundredth unit.

Quota samples

An alternative way of applying some structure to samples is to ensure that the overall sample has the same characteristics as the population. Suppose, for example, we want to find how people would vote in an election. We could take a large random sample, and this would certainly reflect the views of the population. Unfortunately, the sample would have to be very large to ensure the right mix of people. An alternative would be to look at population figures (where the population is those people who are eligible to vote) and see what proportions have various characteristics. Then the sample is chosen so that it contains the same proportions with these characteristics. If the population consists of 47% men and 32% who are over 50 years old, then the

sample will also have 47% men and 32% over 50 years old. Political opinion polls are generally based on samples of around 1200 people, so 564 of the sample would be men and 384 would be over 50 years old.

This approach is known as quota sampling. Each interviewer is given a quota of people with different characteristics to interview: perhaps 12 women who are single, between 20 and 30 years old, have full-time professional jobs, and so on. Although each interviewer is given a quota of each type to fill, the actual choice of people is left to their discretion, so there is still a significant random element. However, the process is not truly random, as an interviewer who has already filled the quota of one category of people does not interview any others in the category, so they have no chance of selection.

WORKED EXAMPLE 3.3

56 300 people are eligible to vote in an electoral constituency. Census records suggest the following mix of characteristics:

Age	18 to 25	16%
	26 to 35	27%
	36 to 45	22%
	46 to 55	18%
	56 to 65	12%
	66 and over	5%
Sex	Female	53%
	Male	47%
Social	A	13%
class	B	27%
	Cl	22%
	C2	15%
	D	23%

A poll of 1200 people is to be taken to assess their probable voting behaviour. How many people should be in each category?

Solution

The sample should contain exactly the same proportion in each category as the population. 16%, or 192 people, should be aged 18 to 25. Of these 192 people, 53% or 102, should be women. Of these 102 women, 13% or 13 should be in social class A. Similarly, 5%, or 60 people, should be over 66 years old, 47%, or 28 of these should be male, and 23% of these, or 6 people, should be in social class D. Repeating these calculations for all other combinations gives the quotas shown in Table 3.1.

Table 3.1

Age		18 to 25	26 to 35	36 to 45	46 to 55	56 to 65	66 and over
Female	A	13	22	18	15	10	4
	B	27	46	38	31	21	9
	Cl	22	38	31	25	17	7
	C2	15	26	21	17	11	5
	D	23	40	32	26	18	7
Male	A	12	20	16	13	9	4
	B	24	41	34	27	18	8
	Cl	20	34	27	22	15	6
	C2	14	23	19	15	10	4
	D	21	35	29	23	16	6

The only problem with such calculations is that rounding to integers may sometimes cause small errors in the quotas. Provided the sample size is fairly large, these errors are small enough to be ignored.

Stratified samples

An extension to quota sampling can be used when there are distinct groups or strata in the population. Then it may be desirable to have some representatives from each stratum in the sample. Before any samples are taken, the population is divided into strata and a random sample is selected from each stratum. This randomness is the main difference from quota sampling. If, for example, we wanted to find the views of various companies, we might want views from manufacturers, transport operators, retailers, wholesalers, and so on. In a particular area there might not be many transport operators, but it would still be important to get their views. Then a stratified sample would specify that certain

numbers of each type of company be approached, even if this meant that small groups were over-represented. Any conclusions drawn from the results would, of course, have to bear this in mind.

Multistage samples

Suppose an organization wants to take a sample of people who share certain characteristics - perhaps the fact that they subscribe to a certain magazine. The organization could simply take a random sample of the population (that is, the people who share this characteristic). Unfortunately, they would then incur a lot of expense in travelling to meet these people and collecting their views. A cheaper solution would be to use multistage sampling. In this, the country is divided into a number of geographical regions (independent television regions, for example). Some of these regions are chosen at random, and then subdivisions are considered, perhaps parliamentary constituencies or local authority areas. Some of these are again selected at random and then divided into smaller areas (perhaps towns or parliamentary wards). This process is continued until, say, streets are identified and then appropriate individuals in these streets are identified as the sample.

The benefit of this multistage approach is that samples are found which are concentrated in a few geographical areas. This dramatically reduces the amount of travelling needed by interviewers and reduces the costs.

Cluster sampling

This chooses the items in a sample not individually, but in clusters. If, for example, we wanted views from people living in a town it would be more convenient to visit a sample which was clustered in a single area than to visit a sample spread over the whole town. Thus the population is divided into a number of groups or clusters, and a number of these clusters are chosen at random to be the sample. Then one cluster might be everybody who lives in a particular road.

Cluster sampling has the benefits of reducing costs and being convenient to organize. It is especially useful when surveying people working in a particular industry. Then individual companies can form the clusters. In other words, companies are selected at random and the sample is made up of all people who work in these random companies. This method works best if the clusters are somewhat dissimilar so that a representative sample can be found

IN SUMMARY

There are several ways of sampling. These can be classified according to census, random, systematic, quota, stratified, multistage and cluster samples.

Self-assessment questions

3.8 Why is sampling used to collect data?

3.9 Why is it important to identify the correct population for a survey?

3.10 What types of sampling might be used?

3.11 What is the key feature of random sampling?

3.12 What is the difference between quota sampling and stratified sampling?

3.13 Where could you find data about UK exports and imports?

3.4. Ways of collecting data

3.4.1 Types of survey

When an appropriate sample has been selected (and for simplicity we shall assume that this is a sample of people), the next stage is to approach them and actually collect data. In many cases this involves observation (including measurement, counting, recording, and so on). In other cases data are collected by asking people relevant questions, in which case we often use a series of related questions presented in a questionnaire.

The Gallup organization suggests five possible objectives for a survey of this type:

- to find if a respondent is aware of an issue ('Do you know of any plans to develop...')

- to get general feelings about an issue ('Do you think this development is beneficial...')
- to get views about specific points in an issue ('Do you think this development will affect...')
- to get reasons for a respondent's views ('Are you against this development because...')
- to find out how strongly these views are held ('On a scale of 1 to 5 how strong are your feelings about this development...')

There are other ways of collecting data, and the best one depends on a combination of use and type of sample. One classification of methods is as follows:

- observation
- personal interview
- telephone interview
- postal survey
- panel survey
- longitudinal survey

Observation

If the population to be sampled consists of machines, animals, files or other inanimate objects the only feasible way of collecting data is direct observation. Even when the population is people, there are many circumstances in which the most reliable results come from direct observation. This is because people often give the answer they feel they ought to give or the answer the interviewer wants, rather than the true answer. Studies have shown, for example, that more people say they use their car seat belts than are shown by direct observation. Similarly, more people say they wash their hands after going to the toilet than is found from observation.

The reliability of observation depends largely on the observer and the circumstances, so it is best for counting, but less good for data which require some judgement. This is particularly true when there is personal involvement. Asking people who are leaving a restaurant what they thought of the meal would give replies based on the whole experience (including who they were with, how they felt, or what the weather was like) rather than a valid opinion of the food. Asking motorway police to give data on accidents would give biased results, because their involvement with the consequences of accidents would lead to emotional rather than factual replies.

Personal interview

Personal interviews are the most reliable way of getting accurate information from people. They have the benefit of ensuring a high response rate, with only 10% of people generally refusing to answer questions. They also allow interviewers to help with questions which are unclear. In some situations, such as quota sampling, some assessment of people is needed before they are questioned, so personal interviews are the only feasible method.

In principle, collecting data by personal interviews is easy; it needs someone to pose questions and listen to the answers. The reality is more complicated, and interviewers need training to ensure that they get reliable replies. Without training, some interviewers might, for example, explain the questions to people, or help those having trouble with an answer (hence introducing the interviewer's

bias to the answer). Similarly, interviewers should be careful not to direct respondents to a particular answer by their expression, tone of voice or additional comments. If an interviewer listens to an answer and then says, 'How strange - not many people give that answer!' the respondent is likely to reconsider the answer and change it.

One of the main drawbacks of personal interviews is the cost. Each interviewer must be transported to the right place and given meals, accommodation, and so on. Typically, 40% of an interviewer's time is spent in travel, while only 35% is available for asking questions (the rest is spent on preparation and

administration).

Telephone interview

About 90% of houses have a telephone, so this provides a popular way of organizing surveys. It has the advantages of being cheap and easy to organize, it involves no travel for interviewers and gets a high response rate. Conversely, it has the disadvantages of introducing bias (as only those with telephones can be contacted), allowing no personal observation of respondents, and annoying people who object to the intrusion of their homes.

A common procedure for telephone interviews is for a computer to select a telephone number at random from a directory listing. Then an interviewer asks the questions presented on a computer screen and record answers directly into the computer. This prevents any errors from being introduced during the transfer of answers from paper forms to the analysing computer.

Postal survey

Sending a printed questionnaire through the post has the advantage of being very cheap and easy to organize, so that very large samples can be used. Postal surveys work best when a series of short questions asks for factual (preferably numerical) data. Major drawbacks are the lack of opportunity to observe respondents and clarify points which respondents do not understand. Perhaps the main disadvantage of postal surveys is the lack of response. Generally, a survey can be expected to generate replies from about 20% of questionnaires. This response can be increased by ensuring the questionnaire is short and easy to complete and is sent to the correct, named individual, by enclosing a pre-paid return envelope, by promising anonymity of replies, by using a follow-up letter or telephone call if replies are slow, by promising a summary of results, or by offering some reward for completion. Unfortunately, a reward for completion (which is typically a small gift or discount on a future purchase) introduces bias, since respondents feel more kindly disposed towards the questionnaire.
One common problem with postal questionnaires is bias. When people are asked for their views on, say, a holiday more people who have had bad experiences will write to complain, than those who have had good experiences. This is an extension of the principle of book reviews, which are always written by 'critics' rather than by 'supporters'

Panel survey

Panel surveys are generally concerned with monitoring changes over time. A panel of respondents are selected, and they are asked a series of questions on different occasions. Thus the political views of a panel can be monitored during the lifetime of a government, or awareness of a product can be monitored during an advertising campaign. Panel surveys are expensive and difficult to administer, so they must rely on small samples.
One interesting problem with panel surveys is that respondents often become so involved in the issues raised that they change their views and behaviour. A panel which is looking at the effects of an anti-smoking advertising campaign might be encouraged to look more deeply into the question of smoking and change their own habits. Another problem is that panel members inevitably leave for some reason and the remainder of the panel become less representative of the population.

Longitudinal survey

This is an extension of a panel survey that involves the monitoring of a group of respondents over a long period. One television company has, for example, been monitoring the progress of a group of children for the past 35 years. The obvious problem with this approach is that considerable resources are needed to sustain an extended survey, and even then a small initial sample must be used. These small samples become vulnerable when some members leave during a long investigation. Longitudinal surveys are generally limited to studies of sociological, health and physical changes.

IN SUMMARY

When a sample has been identified, data can be collected by several means, including observation, personal interview, telephone interview, postal survey, panel survey or longitudinal survey.

3.4.2 Design of questionnaires

Most data collection uses a questionnaire. Even observers are generally asked to record their observations on a sheet of questions. It is important, therefore, that questionnaires are designed carefully and after a great deal of thought. There are many examples of surveys which have failed because they asked the wrong questions, or asked the right questions in the wrong way.
Although it may seem easy, the design of a good questionnaire is difficult. An enormous amount of work has been done on questionnaire design and this allows us to give some guidelines for good practice. The following comments are by no means complete, and although most of them are common sense, they are often overlooked.

- A questionnaire should ask a series of related questions. These should be short, simple questions phrased in everyday terms, and should follow a logical sequence.
- Make questions simple and easy to understand; if people do not understand the question they will give any convenient answer rather than the true one.
- Make questions brief, unambiguous, and without too many conditional clauses.
- Be very careful with the phrasing of questions. Even simple changes in phrasing can give different results, so that, for example, a medical treatment which gives a 60% success rate is viewed differently from one which gives a 40% failure rate. Similarly, a phrase such as 'four out of five people' is viewed differently from '16 out of 20 people' or '80% of people'.
- Avoid leading questions such as 'Do you agree with the common view that BBC television programmes are of a higher quality than IBA television programmes?' Such questions will encourage conformity rather than truthful
  answers.
- Use phrases which are as neutral as possible. Then, 'Do you like this cake?' would be rephrased as 'Say how you feel about the taste of this cake on a scale of 1 to 5'.
- Remember that respondents are not always objective, so the question 'Do you think prison sentences should be used to deter speeding drivers?' will get a different response from 'If you were caught speeding do you think you should go to prison?'
- Phrase all personal questions carefully. 'Have you retired from paid work?' might receive better responses and be just as useful as the more sensitive 'How old are you?'
- Do not start questions with warning clauses. A question which starts 'We hope you do not mind answering this question, but will understand if you do not want to...' will discourage everyone from answering.
- Avoid vague questions such as 'Do you usually buy more meat than vegetables?' This raises questions about 'What is usual?', 'What is more?', 'Should frozen meals be counted as meat or vegetables?' and so on.
- Ask positive questions such as 'Did you buy a Sunday newspaper last week?' rather than the less definite 'How has the number of Sunday newspapers you read changed in the past few years?'

Avoid hypothetical questions such as 'How much would you spend on life insurance if you suddenly won £500,000 on the football pools?' This does not give useful data, because the answer is speculative and has probably not been thought out in any detail

Avoid asking two or more questions in one, such as 'Do you think this development should go ahead because it will increase employment in the area and improve facilities?' This will get confused answers from people who think the development should not go ahead, or those who think it will increase employment but not improve facilities, and so on.

Make the questionnaire as short as possible, consistent with its purpose. A poorly presented questionnaire, or a long one, will frequently not be answered.
Do not ask irrelevant questions. There are a lot of data which could be collected and might be useful. As a questionnaire is being used, it is tempting to assume that an extra question or two could be added with little effect. In reality this costs more to collect and analyse and discourages people from completing the questionnaire

---

Open questions (such as 'Have you any other comments to make?') allow general comments, but they favour the articulate and quick-thinking.

Ask questions which allow precoded answers, so that respondents are offered a series of choices and have to select the most appropriate. There are many formats for these, some of which are illustrated in Figure 3.3.

Address postal surveys to a named person (or at least a title), enclose a covering letter to explain the purpose of the survey, benefits to the respondent, guarantee of anonymity, contact to discuss any difficulties, etc, and include a stamped, addressed return envelope.

Be prepared for unexpected effects, such as sensitivity to the colour and format of the questionnaire, or different types of interviewer getting different responses.

Always run a pilot survey before starting the whole survey. This will highlight any poor questions or other difficulties, and allow improvements to be made to the questionnaire design.

IN SUMMARY

Getting a good design for a questionnaire is difficult and needs a lot of thought. A number of guidelines can be given, but a pilot survey is essential to sort out any problems.

3.4.3 Non-responses

We have already mentioned that around 80% of questionnaires sent by post and 10% of personal interviews can expect to generate no response. There are a number of reasons for non-response, including the following:

- people are unable to answer the questions (perhaps because of language difficulties or ill health)
- they were out when the interviewer called (this problem can be reduced by careful timing of calls and making revisits as necessary)
- they were away for some longer period (holiday or business commitments make surveys during summer more difficult)
- they have moved house and are no longer at the given address (it is rarely worth following up a new address)
- they refuse to answer (probably only 10% of people refuse to answer on principle, but nothing can be done about these)

There is an obvious temptation to simply ignore non-responses and assume that the data collected are typical of the sample: in other words, that the respondents properly represent the sample which in turn properly represents the

population. This is not necessarily true. In an extreme case a postal questionnaire might be used to see how fluently people can read and write (in the same way that people who have reading difficulties are told that they can pick up packages of information when visiting their local library, or write to a central address for more information). Biased replies can also be found when, for example, a survey asks companies how they use computers, with an initial question, 'Does your company use a computer?' Those companies which would answer 'No' to this question are unlikely to be interested enough to complete the rest of the questionnaire, so the responses are biased towards companies which actually use computers.

To avoid this kind of bias, there should be follow-up of non-respondents, with perhaps another visit, telephone call or letter. Initially this should encourage non-respondents to reply, and surveys often increase their response rate by over 20% with a well-timed telephone call or letter. This does not always work and then non-respondents should be examined closely to make sure that they do not share some common characteristic which is absent in respondents.

IN SUMMARY

Most surveys can be expected to yield some non-respondents. These should be carefully examined to ensure that they do not introduce bias to the data collected.

Self-assessment questions

3.14 What method of data collection is appropriate for:

(a) asking how much different companies use computers
(b) asking colleagues for their views on a proposed change in working conditions
(c) testing the effect of exercise on heart disease
(d) testing the accuracy of invoices

3.15 What is wrong with the following questions in a survey:
(a) 'Most people want higher retirement pensions. Do you agree with them?'
(b) 'Does watching too much television affect children's school work?'
(c) 'Should the United Kingdom destroy its nuclear arms, reduce spending on conventional arms and increase expenditure on education?'
(d) 'What is the most likely effect of a single European currency on pensions?'

3.16 What should be done about non-responses in a postal survey?
Why are non-responses irrelevant for quota sampling?

CHAPTER REVIEW

This chapter considered the collection of data. In particular it:

- reviewed the need for information and explained how this relied on data collection
- considered the amount and timing of data collection
- classified data according to qualitative/quantitative, nominal/ordinal/cardinal, discrete/continuous and primary/secondary
- described how data collection relies on taking samples from appropriate populations
- discussed sampling methods, including census, random, systematic, quota, stratified, multistage and cluster samples
- classified alternative ways of collecting data from the sample (including observation, personal interview, telephone interview, postal survey, panel survey and longitudinal survey)
- gave some guidelines for questionnaire design
- mentioned non-responses

From these discussions it is clear that the stages in data collection can be summarized as follows:

- set the objectives and type of data needed
- check available secondary data
- define the relevant population to give primary data
- determine the best sampling method and sample size
- identify an appropriate sample
- design a questionnaire or other method of collection
- train any interviewers, observers or experimenters needed
- run a pilot study
- do the main study
- do any necessary follow-up, such as contacting non-respondents
- analyse and present the results

Problems

3.1 How would you describe the following data:

(a) weights of books posted to a bookshop
(b) numbers of pages in books
(c) positions of football teams in the leagues
(d) opinions about a new novel

3.2 Use government statistics to find how the Gross National Product has changed over the past 20 years.

3.3 What is the appropriate population to give data on:
(a) likely sales of a computer game
(b) problems facing small shopkeepers
(c) parking near a new shopping mall
(d) proposals to close a shopping area to all vehicles

3.4 Describe a sampling procedure which would find reliable data about house values around the country.

3.4 An auditor wants to select a sample of 300 invoices from 9000 available. How might this be done?

3.6 The readership of a Sunday newspaper is felt to have the following characteristics:

Age	16 to 25	12%
	26 to 35	22%
	36 to 45	24%
	46 to 55	18%
	56 to 65	12%
	66 to 75	8%
	76 and over	4%
Sex	Female	38%
	Male	62%
Social class	A	24%
	B	36%
	Cl	24%
	C2	12%
	D	4%

What would be the quotas for a sample of size 2000?

3.7 Describe how you would collect data from a sample of shops selling stage stamps in a particular area

3.8 Give five examples of poor questions used in a survey.

3.9 Give five examples where non-respondents could introduce bias to data

3.10 Run a survey to find the opinions of your colleagues on proposed restrictions on smoking in public places.

3.11 Design a questionnaire to collect data on the closure of a shopping area to all vehicles.

3.12 Find a copy of a recent survey by the Consumers' Association (or any equivalent organization). Describe the data collection used.

Computer exercises

3.1 Use a computer to generate a set of random numbers. Now use these numbers to design a sampling scheme for finding the views of passengers using a local bus service.

3.2 Problem 3.6 gives some characteristics of the readers of a Sunday newspaper. Design a spreadsheet to find the quotas in each category automatically for different sample sizes.

3.3 Conduct a survey into the use of computers by companies operating in your area. How would you select a sample of companies for this? Now use a word processor to design a questionnaire to collect information from the companies. The combination of sample and questionnaire should be good enough to give a reliable view of computer use. Design a spreadsheet to record the data collected by your questionnaire. Now analyse the results and use a word processor to write a report on your findings.

Design a questionnaire to find the views of a sample of your colleagues on a topical issue. Now use this questionnaire to collect actual data. Use a statistical package to record views and see how the results can be presented. Discuss the relative advantages of using a statistical package and a spreadsheet to record results. Write a report on your findings

Case study

Natural Biscuits

Natural Biscuits make a range of foods which are sold to health food shops around the country. They divide the UK into 13 geographical regions based around major cities. The populations, number of shops stocking their goods and annual sales in each region last year are shown in Table 3.2.

Table 3.2

Region	*Population (millions)*	*Shops (£'000s)*	*Sales*
Greater	8130	94	240
Birmingham	1205	18	51
Glasgow	870	8	24
Leeds	853	9	18
Sheffield	64.1	7	23
Liverpool	580	12	35
Bradford	556	8	17
Manchester	541	6	8
Edinburgh	526	5	4
Bristol	470	17	66
Coventry	372	8	32
Belfast	365	4	15
Cardiff	336	4	25

Natural Biscuits are about to introduce a Vegan Veggie Bar which is made from a combination of nuts, seeds and dried fruit, and is guaranteed to contain no animal products. The company want to assess likely sales of the bar and are considering a market survey.

Natural Biscuits already sell 300 000 similar bars a year at an average price of 40 pence, and with an average profit of 7.5 pence. An initial survey of 120 customers in three shops earlier this year gave the characteristics of customers for these bars shown in Table 3.3.

Experience suggests that it costs £10 to interview a customer personally, while a postal or telephone survey would cost £5 a response. The analysis of information can be done relatively cheaply by the Management Information Group at Natural Biscuits.

The problem is to help Natural Biscuits to collect information about the potential sales of their Vegan Veggie Bar. They want as much information as possible, but obviously want to limit costs to reasonable levels.

Table 3.3

Sex	*Female*	*64%*
	Male	*36%*
Age	*Less than 20*	*16%*
	20 to 30	*43%*
	30 to 40	*28%*
	40 to 60	*9%*
	More than 60	*4%*
Social class	*A*	*6%*
	B	*48%*
	Cl	*33%*
	C2	*10%*
	D	*3%*
Vegetarian	*Yes*	*36%*
		(5% vegan)
	No	*60%*
	Other response	*4%*
Reason for buying	*Like the taste*	*35%*
	For fibre content	*17%.*
	Never tried before	*11%*
	Help diet	*8%*
	Other response	*29%*
Regular buyer of bar	*Yes*	*32%*
	No	*31%*
	Other response	*37%*

Your problem is to design a data-collection project. Full details should be given of all aspects of the project, including timing and costs. You can use any relevant secondary information and make valid assumptions where appropriate.