2

Survey Sampling

William Schofield

This chapter is about the methods and problems of designing and undertaking sample surveys. The contents are relevant to other quantified research methods, however, since inferences about population values from sample measurements will be at the heart of all of them. Even at the simple level of a survey conducted on one occasion, possibly by questionnaire, or structured interview, or planned selective observation, inference is involved. What is being inferred is a characteristic, or characteristics, of the population, and this is inferred from the subset of measurements obtained from the sample. Behind this process are mathematical models and theorems which underpin the validity of the inferences so made.

Already in this introduction a number of technical terms have been used, and you will probably be uncertain of their meaning. This is nothing to be concerned about. A specialist topic such as sampling methodology is bound to need a specialized terminology, and the first objective of the sections that follow is to explain this terminology and to give examples. The overall aims of the chapter can be summarized in two sentences:

1 It will introduce methods for obtaining representative samples of appropriate size from a population, and for providing estimates of how accurate statistics calculated from any such sample are likely to be.
2 It will present and discuss problems in applied survey sampling, for example non-response, unreliable or invalid measurement, sample loss, incomplete data, and ways of reducing the effect of these on the final results.

Sampling

A *sample* is a set of elements selected in some way from a population. The aim of sampling is to save time and effort, but also to obtain consistent and unbiased estimates of the population status in terms of whatever is being researched. The important point to note here is the very restricted meaning given to the term *population* in statistics, which is quite different from everyday usage. Thus, a population could be all the children in some group of interest, perhaps all the children in one school, or all the children in a specified age range in a certain district, or city, or in the UK overall. A population consists of individuals, or *elements*, and these could be persons, or events, or cabbages, nuts or bolts, cities, lakes, patients, hospitals or thunderstorms: anything at all of research interest, including observations, judgments, abstract qualities, etc.

Usually, in survey research, we will be interested not just in the characteristics of a sample, but in those of the population from which the sample has been drawn. Descriptive statistics for a population are called *population parameters* to contrast them with *sample statistics.* Usually, the aim of a research project is not exact measurement of population parameters, such as is undertaken in a general census, but the collection of sample data to be used both to calculate sample statistics and to estimate how close these are to the unknown population parameters, i.e. to estimate the extent of *sampling error,* a concept which will be explained fully in this chapter. Thus, matters of interest in applied sampling include:

1. What methods are available and what are the advantages and disadvantages of each of them, theoretically, in practical terms, and in terms of cost?
2. How close will statistics calculated from samples be to the unknown population parameters?
3. How much will sample size influence this?
4. Which will be the most effective methods of drawing representative samples (that is, minimizing sampling error as much as possible) and in which circumstances?
5. Given that a sample has been appropriately drawn, how can the effects of non-response, or sample loss in any form, be estimated?

Researchers and statisticians have developed techniques for dealing with matters such as these, and they have also developed a specialized terminology so that they can be defined and discussed. The objective of the section is to introduce you to the essential basics of this terminology.

Defining the Population to be Sampled

The first step in sampling is to define the population of interest clearly and accurately. Such definition may seem obvious to a novice, but it is where survey design can all too easily be defective. For example, the intended population might be housebound single parents of young children, but if these were found *via* the records of a social or health service agency then a substantial bias might be introduced by the exclusion of parents not using, or not known to, such agencies. A further obvious example is using the telephone to contact respondents; this limits representativeness to persons meeting selection criteria, but only if they also are available by telephone. Such individuals might differ in very relevant ways from the intended population of interest. Problems such as these can be avoided by defining a population as the total collection of elements actually available for sampling, rather than in some more general way. The words ‘group’ and ‘aggregate’ get close to what statisticians mean by a population (Kendall, 1952). A useful discipline for the researcher, therefore, is to bear firmly in mind precisely which elements were available in the intended population and which were not, and to use this information to limit the extent of the claims he or she makes about the generalizability of the results.

Sampling Units

For the purposes of sampling, populations can be thought of as consisting of *sampling units.* These are collections of elements which do not overlap and which exhaust the entire population. For example, if the elements were fingers, and the population all the fingers in the UK, then the sampling units could be geographical regions, provided they covered the whole of the UK and did not overlap. Or the sampling units could be families, or individual persons, or hands. If the elements were persons over 60 who lived alone but who were currently receiving nursing care in hospital immediately following major surgery, and the population were all such individuals in the UK, then the sampling units could be geographical regions, or hospitals, but not cities, because these might not exhaust the population of interest. Sampling cities might, for example, exclude individuals living in rural areas.

The Sampling Frame

When a survey is being set up, the sampling units are organized by the researcher into a *sampling frame.* A sampling frame is whatever is being used to identify the elements in each sampling unit. Remember that each sampling unit could contain many elements, in the case of geographical regions, or just one, in the case of simple random sampling from the voting register. Whatever the circumstances, the sampling frame provides access to the individual elements of the population under study, either via sampling units, or directly when these and the population elements are identical (for example, where we are sampling people from a finite population and we have a complete list of the names of the population).

The sampling frame could be anything at all provided that it exhausts the total population. For example, it could be company employment records, or school class lists, or hospital files, or the voting register. Such lists and records will always contain mistakes, but they may be the only method of finding the sample elements so that the population can be surveyed. The survey results, when they are available, will give some information on the extent of inaccuracy of this sort, for example by providing a count of voters no longer resident at the address given in the register. It will then be possible to see whether or not these inaccuracies are fairly evenly spread across the sampling frame. It is possible that greater housing mobility will be more typical of certain sample elements than others, leading to bias in the survey results. (Incidentally, the term *bias* has a precise meaning in statistics. In this chapter it refers to an effect on the sample data from anything that moves the value of a statistic calculated from that sample (such as a mean) further from the true population value than would have been the case if that effect were not present).

Another and more invidious source of bias in sampling is faulty selection of the sampling frame itself. In the real world of survey research, a sample is not really a random set of elements drawn from those which define the population being researched; researchers can only strive to make it as close to this as possible. In practice, a sample can only be a collection of elements from sampling units drawn from a sampling frame, and if that *sampling frame* is not fully representative of the population to be described, then the sample will also be unrepresentative. For this reason, great care should be taken in deciding just what sources will provide the sampling frame for a survey before the frame is set up and the sample drawn.

It is important to understand that if a sampling frame is a biased representation of the population to be studied, increasing sample size will not help – the bias will remain. Even an up-to-date electoral register might not provide an accurate frame for selecting a sample from the population of voters in an approaching election. This is because it will include people who did not, in the event, vote, although they may have intended to do so when surveyed, and these individuals might differ in relevant ways from those who did vote. It will also include those who have moved away, or died, and will not include those who have actively avoided registration for some reason; for example, to avoid jury service or a local tax.

These points have been stressed because, until one is faced with the task of accounting for an unexpected or even improbable result in survey research, locating the elements of a population might seem to involve only the practical issues of gaining access to records or lists. Clearly, there is much more to it than this.

Selecting a Sample

Having identified the population to be researched, and arranged access to it via an accurate sampling frame, the next step will be to decide how the sample itself is to be selected. The objective will be to obtain estimates of population parameters, and some methods will do this more accurately than others. The choice of method will be a question of balancing accuracy against cost and feasibility. The methods available fall into two main categories: probabilistic sampling and non-probabilistic sampling. *Probabilistic sampling* includes simple random sampling, stratified random sampling and, if selection is at least in part random, cluster sampling. The most widely used method of *non-probabilistic sampling* is quota sampling.

Sampling will often be the only feasible method of obtaining data, quite apart from questions of time and cost. But do not assume that extending a sample to include all elements in a population (i.e. conducting a census) would necessarily give better information. In some circumstances, a sample will be more accurate than a census, as well as cheaper, quicker and less invasive of the community. Some sources of discrepancy between the estimated (measured) and true population value, which will hereafter be referred to as *error,* are more likely in a large-scale census than in a small and tightly managed sampling survey.

Activity 2.1 (10 minutes)

Write a brief account (100–120 words) of what you think might be greater problems for a census than for a sampling survey, assuming a fairly large, and dispersed, population (as in the national census).

Answers to activities are given at the end of the chapter.

Error is another word with a specialized meaning in sampling theory. It is not synonymous with 'mistake', and does not mean 'wrong', although a mistake by an interviewer or a wrong answer to a question would each contribute to error in a survey, whether a sample survey or a census. In addition to this, for many things measured there will be variation from many sources, including individual variation, and looking at this from the perspective of a summary statistic such as the mean, this will also be error.

In your answer to Activity 2.1 you probably mentioned some factors such as field-work problems, interviewer-induced bias, the nature or insensitivity of the measuring instrument or clerical problems in managing large amounts of data. Bias from sources such as these will be present irrespective of whether a sample is drawn or a census taken. For that reason it is known as *non-sampling* error. It will be present in sample survey results, but will not be attributable to the sampling method itself. This is an important distinction.

Error which *is* attributable to sampling, and which therefore is not present in census-gathered information, is called *sampling error.* Since a sample has both kinds of error, whereas a census only has the former, you might conclude that the advantage really does rest with the census. The point from Activity 2.1 was, however, that the scale of census-taking makes it difficult to reduce the risk of non-sampling error, and that this can be easier to do in a well-planned sample survey. Also, as you will see later, sampling error can be controlled (or at least the extent of it can be estimated, which amounts to the same thing). Thus there are occasions when a survey could produce less error overall than a full census.

As mentioned above, there are two basic methods of sampling: *probability sampling* and *non-probability sampling.* The former includes simple random sampling, stratified random sampling and some forms of cluster sampling. The latter, sometimes called *purposive,* includes (at its *most* sophisticated) quota sampling and (at its *least* sophisticated) what is sometimes called 'opportunity' sampling: the simple expedient of using as a sample whoever is available and willing (e.g. a 'captive' school class). A practical approach to each of these will be given in the following paragraphs.

Probability samples have considerable advantages over all other forms of sampling. All samples will differ to some extent from the population parameters, i.e. they will be subject to sampling error. Thus, suppose we sampled 100 children, and the average height of the 100 children was 1.2 metres. If the average for all the children in the population from which the sample was drawn was 1.1 metres, then the error attributable to drawing the sample rather than measuring the whole population would be 0.1 metres. For probability samples, very accurate estimates can be given of the likely range of this error, even though the population value will obviously not be known. This involves a fundamental statistical process, the randomization of error variation. Because randomization is missing from non-probabilistic methods, they have no such advantage. Just what this means will be explained in the next section (on simple random sampling).

Simple Random Sampling

The fundamental method of probability sampling is *simple random sampling.* Random sampling means that every element in the population of interest has an *equal* *and independent* chance of being chosen. Here the word 'independent' means that the selection of any one element in no way influences the selection of any other. 'Simple' does not mean that random sampling is easier to carry out than other methods, but that steps are taken to ensure that nothing influences selection each time a choice is made, other than chance. In theory, this requires selection with replacement – any element sampled should be replaced in the sampling frame so that it has a chance of being chosen again – or else the probability of being selected would change each time an element was removed from the sampling frame and placed in the sample. In practice, however, samples in survey research will generally be comparatively small in contrast with the number of elements potentially available for sampling, and the effect of non-replacement will be trivial and need not concern us further.

Beginner researchers sometimes think that if they do nothing, but simply take what comes along, then this will somehow amount to 'chance' or random' selection. However, *setting the probability* of selecting elements from a population cannot be left to chance. 'Random', in sampling, does not mean 'haphazard' or following no thought-out plan in obtaining sampling elements. Possibly even worse than the beginner who does nothing is the experienced scientist who thinks he or she can select randomly, near enough. There is much literature showing that this is not the case. Even for something as simple as selecting wheat plants from a field for the study of growth characteristics, experienced researchers trying to choose randomly have been shown to introduce strong biases. For samples taken when the plants are young there tends to be over-selection of those that are tallest, whereas a month later, when the ears have formed, there is a strong bias towards middle-sized, sturdier plants (Yates, 1935).

Simple random sampling might not be at all simple to achieve, depending on circumstances. For example, it might be very difficult to achieve a random sample of serving seamen in the merchant navy, even if an accurate sampling frame could be compiled. Random sampling does not just mean stopping individuals in the street. Which individuals? Which street? Obviously, there would be a risk of stopping only individuals who looked as if they would be helpful. Or of choosing a well-lit, safe street. The flow of passers-by might be influenced by some biasing event: a store sale, an office or bureau somewhere in the not too immediate vicinity, etc.

Random sampling is similar to tossing a coin, throwing a dice or drawing names from a hat, and in some circumstances procedures such as these might be adequate, but usually random number tables or computerized random number generators will be used. The first step is to number in order the individual elements in the population, as listed in the sampling frame. Sometimes this numbering will already be present, or will be implied. If tables are to be used, the next step is to enter them at some random place; for example, by dropping a small item on to the page and selecting the number nearest to it. This then provides the first random number. From this start the required set of random numbers is achieved by stepping forwards or backwards or sideways through the tables in any systematic way. Until recently, statistical texts usually contained tables of random numbers, but nowadays most researchers use readily available computer programs.

Many surveys will not have used true random sampling, but something called *systematic sampling*. If, for example, you have a list of 100 names from which to sample 10, an easy way to obtain a sample is to start from a randomly chosen point on the list and take every tenth item (treating the list as circular and starting again at the beginning when the end is reached). The great advantage of systematic sampling over simple random sampling is that it is easier to perform and thus provides more information per unit cost than does simple random sampling. Also, because it is simpler, fieldworkers are less likely to make selection errors. For example, constructing a simple random sample of shoppers leaving a certain supermarket might be very difficult to achieve in practice, but selecting one at random, and then subsequently stopping every twentieth shopper, would be less so. A systematic sample is more evenly spread across a population than a simple random sample, and in some circumstances this could be advantageous, for example in monitoring items on a production line, or for choosing a sample of accounts for detailed auditing. Mostly, systematic sampling will be adequate as a form of random sampling, but only to the extent to which there is no 'pattern' in the sampling frame and the placing of any item in it really *is* independent of the placing of other items. This is by no means always the case. If we were using the Electoral Register, for example, we would expect the members of each household to be listed together and, as there are seldom as many as 10 people in a household, the selection of a given household member would guarantee the exclusion of all other household members if we were to take every tenth name. Worse, systematic sampling can occasionally introduce a *systematic bias:* for example, if the names in a school class were listed systematically as 'boy, girl, boy, girl …' and we sampled every second name, we should obtain a sample made up of a single gender from a class made up of both genders in equal proportions. When such risks are known, they can be avoided by choosing a suitable sampling interval; or, after a set number of elements has been sampled, a fresh random start can be made.

Stratified Random Sampling

One problem with simple random sampling is that sample size may need to be disproportionately large to ensure that all subgroups (or *strata)* in the population are adequately represented. For example, a researcher who intends surveying the attitudes of school leavers to further training might see age at leaving school as important. A simple random sample would need to be large enough to remove the risk of inadequate representation of the ages of leaving with least frequency in the population. This could be avoided by dividing the population into age strata, and randomly sampling from each of these. The objective would be adequate representation at reduced cost.

To draw a *stratified random sample,* the elements of a population are divided into non-overlapping groups – *strata.* Simple random samples are drawn from each of these, and together they form the total sample. If the proportion of the sample taken from each stratum is the same as in the population, then the procedure is called *proportionate* stratified random sampling, and the total sample will match the population. In some cases, however, this might result in small strata of interest not being represented adequately in the final sample. This can be avoided by increasing sample size in all such strata, but not for other strata, and still with random selection. The result would be *disproportionate* stratified random sampling.

Here the sample will not match the population, but it will differ from it in known ways which can be corrected arithmetically. (An unbiased estimator of the population mean will be a weighted average of the sample means for the strata; that is, the contribution of each subset of data to the population estimates will be in proportion to its size. Estimates of variance and of sampling error can also be weighted).

You may well wonder: why not always stratify, and thus reduce cost? One problem is that stratification is not always a simple matter. In any reasonably sized survey, a number of variables will be possible candidates to govern stratification. Deciding which can be no easy matter. Decisions which seem clear cut when the sample is being selected are sometimes reassessed as unfortunate when the survey results are interpreted.

Further, although the purpose of stratification is to increase precision by reducing sampling error without increasing cost, it can, in some circumstances, lead to less precision than simple random sampling. For example, in a national educational survey, stratification could be made in terms of education authorities, on the grounds that these vary greatly in size and character, and also for administrative convenience, but there could be other units unequally distributed within authorities, such as type of school district or level of institution, with greater variability than between authorities. Unfortunately, this might remain unknown until the survey results are analyzed, when it would become clear that a simple random sample of the same size would have provided better population estimates.

Even so, for many surveys, an advantage is seen for stratification, and it is possibly the most popular procedure in survey research. Money is saved by reduction in sample size for the required degree of statistical precision. Fieldwork costs, such as time, travel, interviewer and administration fees, and the printing and processing of questionnaires, are reduced.

Table 2.1 illustrates proportionate and disproportionate random sampling from a population of 6,000 school leavers in a provincial city. The objective of the survey was to find what proportion of the school leavers were in full-time education or employment 18 months after leaving school. The researchers were asked to provide a breakdown of the findings by sex and any other factor found to be relevant when the sample data were analyzed. For the purpose of the example, it will be assumed that sample size was limited to 400.

Activity 2.2 (allow 30 minutes)

The columns in Table 2.1 for proportionate and disproportionate sample size for the separate strata have been left blank. Calculate and enter on the table the missing figures. Before reading further, make a note of which method of sampling you think ought to have been used on this occasion and, very briefly, why. Check your calculations against the completed table given at the end of this chapter.

Table 2.1 *Proportionate and disproportionate stratified random sampling from 6,000 school leavers*

School leaving age	Population size	% of total in each stratum	Proportionate		Disproportionate	
			Sample size	Sampling fraction	Sample size	Sampling fraction
16	2,730	45.5	182	1/15	134	1/20
17	1,950	32.5	130	1/15	134	1/15
18+	1,320	22.0	88	1/15	134	1/10
Total	6,000	100.0	400	1/15	402	1/15

[a]The denominators for the disproportionate sampling fractions have been rounded to give whole numbers.

The decision as to whether proportionate or disproportionate stratified random sampling should be used in this case cannot be made on statistical grounds. If the overall population parameters are the main interest then proportionate sampling will give the best estimates of these. But is this likely to be the case? A breakdown by sex has been requested and, assuming that the school leavers are about half female and half male, the oldest leavers will be represented by about 44 boys and 44 girls. Considering that the researcher will break these down into those who are employed, unemployed, or in further education, then group size will be getting very small. And what about other groups of potential interest, such as race, or new immigrant categories: is it likely that the client will want information on sex differences within these as well?

When making decisions on sampling, the researcher will have to balance these various potential needs, and there will have to be compromise. In general, if there is a risk that subgroups of interest will be insufficiently represented in some strata of the sample, then sample size will need to be increased, and the cost of this will be less if the increase is made only in the strata where it is thought to be needed. This could then compromise some other aspect of the findings, by reducing sample size in other strata if the same total sample size needs to be maintained, and mostly there will be no completely satisfactory answer to problems of this kind in practice.

For the present example the preferred method would be simple random sampling, with a sample size big enough to provide adequate numbers in the smallest groups of interest. As an adequate size for the smallest subgroup could mean 40 or more children, this would be unrealistic. Consequently, disproportionate random sampling would be used. This would cost more per element sampled because of the need to include age strata as a sampling criterion, but overall it would be cheaper because fewer elements would be required. In the full sample, however, there would be over-representation of minority groups and the sample would not accurately reflect population characteristics, although this could be taken into account in calculating and interpreting statistics.

Cluster Sampling

Cluster sampling improves on stratified random sampling by further reducing costs, but with a risk of increasing sampling error. A *cluster sample* is a probability sample in which the elements are all the members of randomly selected sampling units, each of which is a collection or cluster of elements from the population sampled. Cluster sampling is advantageous when a sampling frame listing population elements is not available, or is not easily obtained, or is likely to be very inaccurate, and also when the cost of conducting the survey will be unduly increased by the distance separating the elements. This distance is usually in the geographical sense, but it could also be in time; for example, when supermarket customers, or people brought into a police station or casualty clinic, are sampled at designated time periods during a week. For a genuine probability sample, both the time periods, or any other form of cluster, and the individuals surveyed should be chosen at random. Simply accepting all the individuals who turn up or pass by at some specified time or times until the required number has been obtained would not constitute cluster sampling, which is a probability method. Cluster sampling can be proportionate or disproportionate, as described above for stratified random sampling. Further, in many contexts there will be another level of selection within clusters, either by random selection or by additional clustering.

In a study of teacher/parent relationships within one large local education authority, interest centred on the final two years of infant education and the first two years of junior school. This is an age range when there is an emphasis on learning to read, and on improving reading skill and vocabulary. Some infant and junior classes were within one school, under one headteacher, but in other cases infant and junior schools were separately housed and administered, although intake at the junior school was always from its related infant school. Further, the schools were of different sizes, ranging from one to three parallel classes.

For sampling purposes, children attending school in the authority in the four relevant school years were the population, and clustering was in non-overlapping blocks of all classes within the two infant and two junior years at the same school or schools related by intake. The survey was, in fact, to be conducted four times, at yearly intervals, so that children in the first infant year at the beginning of the study could be followed in subsequent years and compared with parallel previous and future year groups, both within and between schools. The study thus matched a design described in Chapter 1, but it is important to understand that the sampling plan for a repeated (longitudinal) survey, or for an intervention project, could be basically the same as for a once-only survey.

For the purpose of this example, assume that six of these clusters were chosen at random for the survey. This would give clusters containing disproportionate numbers of classes, i.e. ranging from a total of four to 12 depending on school size. A random sample of children could have been chosen from each of these clusters or, if intact classes were required, one class could have been chosen by lottery from each school year within each cluster.

As mentioned above, cluster sampling is cheaper than other methods because the cost of data collection can be greatly reduced. In the example, a sample size of between 600 and 700 children would be expected in the first year. Clearly, interviewing or testing would have been more time-consuming and costly if these had been randomly selected from all the authority's schools, instead of being concentrated in just a few. The task of following and individually testing these children as they progressed across the four school years would also have been considerable.

In this hypothetical research (modified from Tizard et al., 1982), there would be the risk that the clusters had unfortunately fallen by chance on a small group of schools that were unrepresentative, either totally or to some significant extent, of the other schools under the authority's care. There is no way that this risk could be ruled out without taking into account information from outside the survey findings, since these would only include material on the selected schools. Clearly, any such information could be taken into account in the first place and used to formulate selection criteria. These criteria could then be used to select clusters larger than needed and from which random selection could take place. Sometimes cluster samples are non-probability samples at every level, except that if there are treatment and control groups then these are allocated to those conditions randomly. It is essential that this one element of randomization is preserved, and even then the sampling is not strictly cluster sampling but is of the non-probabilistic, opportunity variety.

Sometimes cluster sampling is the only option realistically available – for example, for surveying the unemployed and homeless, when compilation of a sampling frame would in practice be impossible or in cases where lists of the elements in a population may exist, but are unavailable to the survey researcher or are known to be unredeemably defective.

Much policy-related survey research is undertaken in developing countries and can influence the funding policy of aid agencies. Sample findings which did not generalize to the population in need of aid could have disastrous consequences, yet random or stratified sampling from population lists is not likely to be possible. Cluster sampling will usually include at least some element of randomization which, in contrast with a totally haphazard method, will permit qualified estimates of the extent of sampling error. Common sense and professional judgement are likely to be needed even more than usual when evaluating research results in such circumstances.

Quota Sampling

In the most widely used method of non-probability sampling, the population is split up into non-overlapping subgroups, as for stratified sampling. Quotas of the desired number of sample cases are then calculated proportionally to the number of elements in these subgroups. These quotas are then divided up among the interviewers, who simply set out to find individuals who fit the required quota criteria. They continue doing this until they have filled their quota. A given interviewer, for example, might have to find five middle-class and five working-class women over the age of 30, with no other control over who these people are or how they are located, as long as they fill the criteria. This method is called *quota sampling.*

One reason for using quota samples is, again, to reduce costs, but another is that this method seems to have intuitive appeal to some survey practitioners. For example, quota sampling is widely used in market research. Thus, if a population is known to have 60 per cent females and 40 per cent males, it might require a comparatively large sample to reflect this proportion exactly. It might, however, seem important to the researcher that it should be so reflected, whatever the sample size. This can be achieved, for the present example even in a sample of 10, by selecting quotas of six females and four males.

The major problem with quota sampling is that attempts to deal with one known source of bias may well make matters worse for others not known, or at least not

Table 2.2 *Interlocking quota sampling design for a survey project*

	Odd student number		Even student number	
Sex	18–34	35+	18–34	35+
Male	Middle	Working	Working	Middle
Female	Working	Middle	Middle	Working

Social classes ABCI on the Social Grading Schema are shown as 'middle', and social classes C2DE as 'working'.

known until after the data have been collected and it is too late. Further, as there is no element of randomization, the extent of sampling error cannot be estimated. For example, in an Open University research methods course, students undertook a research project on class attitudes in the UK. A questionnaire designed and piloted by the course team was administered by each student to a small group of respondents. The student then collated the data for her or his small sub-sample, and sent them off to the university to be merged with the data from all other students. For 1991, this produced a national sample of well over 900 respondents, as did the preceding year. Questions were included on class consciousness, class awareness, and aspects of class structure. There were questions intended to identify class stereotypes, and questions seeking biographical information on such matters as sex, age, educational attainment, housing and previous voting behaviour. The intended population was all adults in the UK.

Activity 2.3 (5 minutes)

Write a very brief note saying which sampling method would, in an ideal world, have been best for this study, and what the main advantage of this would be.

In fact, the method chosen, for practical reasons, was quota sampling. Each student was asked to collect data from four respondents in an interlocking quota design, which took into account three criteria: social class, sex and age. This design is shown in Table 2.2. Thus, a student with an even OU student number (right-hand side of the table) found one respondent for each of the following categories:

- male/working class/18–34 years
- male/middle class/35+ years
- female/middle class/18–34 years
- female/working class/35+ years

As you can see from Table 2.2, a student with an odd student number also chose four respondents, but with the position of the social class categories reversed.

When the course was being developed, a pilot study was undertaken using this quota design. The results revealed what appeared to be a strong bias in the sample.

For example, there was an unexpectedly high number of respondents in social class A or B who appeared to have 'left-wing' attitudes. Further, the pilot study did not find some of the differences between the social classes expected from other research. The pilot interviewers had correctly followed the quota procedure but had selected middle-class individuals who were not representative of the national population. The same may well have applied to the selection of working-class respondents. It is easy to think of many possible sources of bias when respondents are selected by interviewers solely to fill sampling quotas.

Although the bias in the pilot study was noted when the results were analyzed, it was decided that in a large sample collected by OU students across all parts of the UK there would be no such problem. In the event, as the years went by, successive waves of students collected and analyzed the pooled data, both regionally and nationally. Invariably it was found that, although students had selected individuals to fill their quotas according to the set criteria, the working-class individuals sampled by OU students differed in material ways from what was expected for that social class in the population at large, and the same was true for the middle-class component of the sample. Random sampling, if it had been possible, would have avoided this persistent and pervasive selection bias, whereas increasing sample size did not. In general, selection bias will never be overcome by increasing sample size, which will merely inflate costs to no avail.

Other non-probability methods – such as 'opportunity' sampling, the simple expedient of including as subjects whoever happens to be available from the population of interest – have already been mentioned. These methods, or lack of methods, are sometimes referred to as 'haphazard' sampling, but the term 'opportunity' is preferred because it implies usually what is the case; that is, the necessity of accepting whatever is available, with no realistic alternative. Thus, in a study of new admissions of the elderly to institutional care, the sample might be all admissions until adequate sample size has been achieved at the only institution, or institutions, available to the researcher.

Alternatively, data collection could be continued until certain predetermined quotas were achieved. For example, a quota could be set for the minimum number of persons diagnosed as suffering from senile dementia. Note that this differs from regular quota sampling, where there is an element of choice in that the fieldworker selects individuals to fill the quotas. It also is not *sequential sampling,* which is a method assuming both independence and random selection, but in which sample size is not predetermined. Instead, random sampling continues sequentially until a pre-established criterion has been met; for example, that the sample includes 30 individuals diagnosed as having senile dementia. The purpose of sequential sampling is to find out what sample size will be needed to reach the set criterion.

Estimation of Population Parameters

To follow this section you need to understand in principle the main measure of central tendency – the mean – and measures of dispersion such as the variance and standard deviation. You will also need to understand how probability can be defined as relative frequency of occurrence, and that it can be represented by the area under a curve – by a frequency distribution.

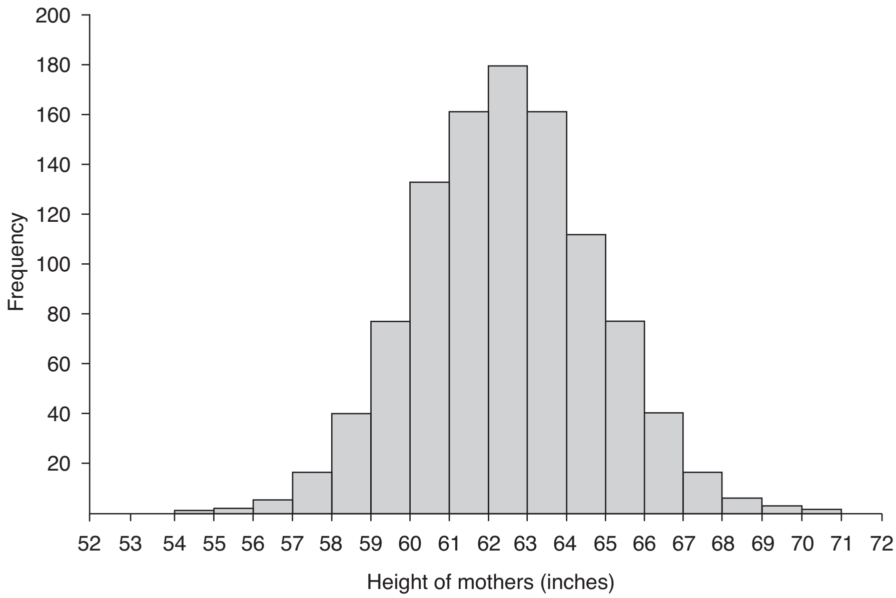

Figure 2.1 *Distribution of heights in a sample of 1,052 mothers (Pearson and Lee, 1902)*

Means, Variance and Standard Deviations

Figure 2.1 is a histogram showing the heights of a sample of 1,052 mothers. The central column of this histogram tells us that about 180 mothers in the sample were between 62 and 63 inches high. From the column furthest to the right we can see that almost none was over 70 inches high. Histograms are simple graphical devices for showing frequency of occurrence.

Figure 2.2 shows another way of representing this same information – this time by a continuous curve – but the histogram has been left in so that you can see that the curve does fit, and that either method provides a graphical representation of the same information. Mothers' heights are a continuous measure, and in a way the curve seems more appropriate than the histogram. But the histogram makes it clear that what is represented by the area within the figure, or under the curve, is frequency of occurrence. Thus the greatest frequency – the most common height for the mothers – is the 62–63 inch interval. Reading from the curve, we can more accurately place this at 62.5 inches. Check this for yourself.

Now look at Figure 2.3. Here only the curve is shown, together with the scale on the axis, plus some additional scales and markings. Take your time, and study these carefully. Note first of all that the total area under the curve from its left extreme to its right extreme represents the variation in height of all 1,052 mothers. It is important to understand the idea of the area representing, in the sense of being proportional to, the variation in the mothers' heights. Some other matters need revision before the additional scales shown on Figure 2.3 are explained, and we will also, for the moment, defer explanation of the areas on the figure marked off as 68 per cent and 95 per cent of the total area under the curve.

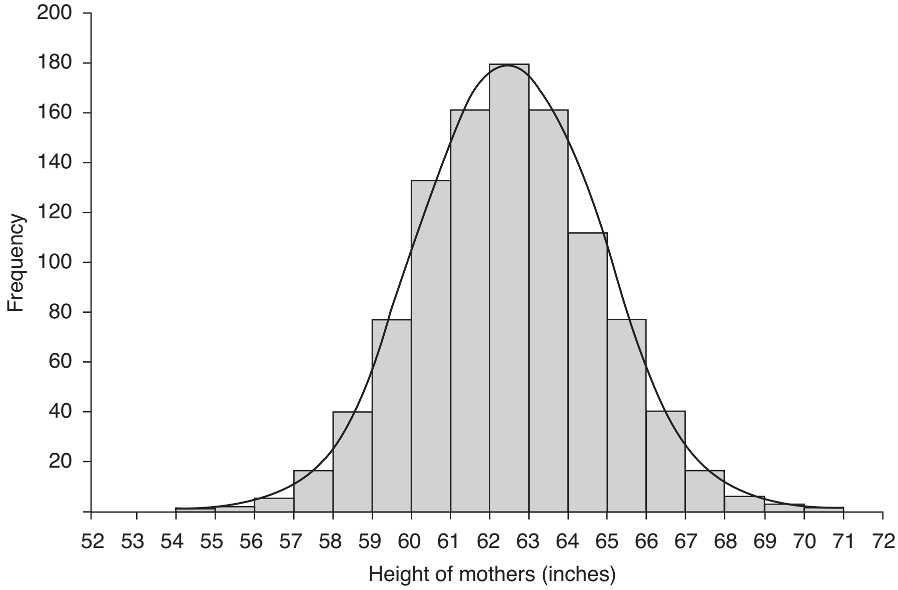

Figure 2.2 *Mothers' heights: histogram and superimposed normal curve*

The *mean* is simply the total height of all mothers, added up and divided by the number of mothers, i.e. the arithmetical average. The mean is at the centre of the distribution shown in Figure 2.3. Half of the mothers' heights (50 per cent or 0.5 as a proportion) are above the mean and half are below it. The *standard deviation* is also an average, but not an average representing where the centre of the distribution is but how it is spread out, i.e. an average of the dispersion. The standard deviation of a population is found by calculating the mean; finding the difference between each value and the mean; squaring each of the differences (deviations) so obtained; adding them all up; dividing by the number of values averaged; and finding the square root of the final answer to get back to the original scale of measurement. If you read through the preceding sentence quickly it might seem complicated, and standard deviations may remain a mystery. If you are still unsure on this idea of averaging the dispersion then re-read slowly, and perhaps use a pencil to re-express the procedures described in your own words.

Activity 2.4 (15 minutes)

Better still, using only the information given in the preceding two paragraphs, calculate the mean and the standard deviation of the following numbers:

(1, 2, 3, 4, 5)

and also for

(1, 1, 0, 1, 12)

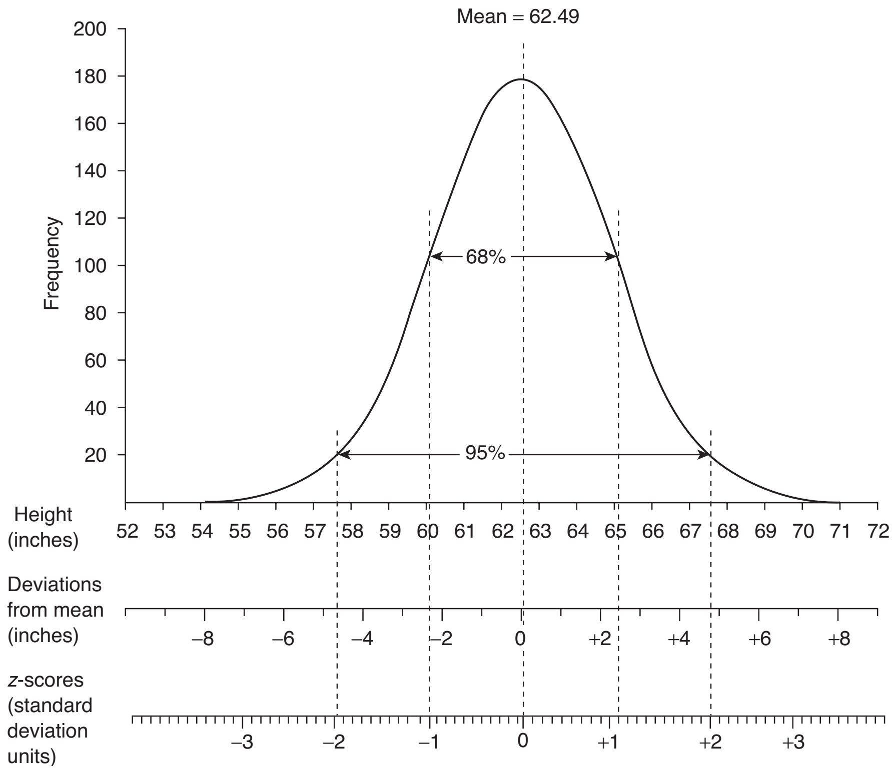

Figure 2.3 *Mothers' heights expressed in deviation and z-score units*

Compare the two means and the two standard deviations, and make a note of any way in which they are similar, or different. Check your results against those given at the end of the chapter.

In the preceding paragraphs, to provide a clear account of what the mean and the standard deviation (s.d.) represent, it has been assumed that the simple set of numbers

$$(1, 2, 3, 4, 5)$$

form a complete population. If, however, they represent a sample from a population, then an adjustment would be needed to allow for the fact that the standard deviation is a biased estimator of the population standard deviation. It does seem intuitively obvious that the dispersion in samples drawn from a population must on average be less than the dispersion in the total population. Also, that the smaller a sample is, then the greater the comparative effect of this bias will be. For that reason, when the s.d. of a sample is calculated, rather than that of a population, the divisor in forming the average of the squared deviations from the mean is not the sample size n, but n–l (This adjustment will have greater effect on small samples than on large ones).

Now, to return to Figure 2.3: the point about this curve is that it follows a well-known mathematical model, the normal distribution, and is completely described by its mean and standard deviation. Once you know the mean and s.d. of any normally distributed variable, then you can say precisely what shape the distribution will be, and whereabouts within this shape any particular value will fall. This will be a statement about probability of occurrence.

Thus, if you had the height of just one mother there would be a 100 per cent (near enough) probability that it would fall somewhere under the curve shown. There would be a high probability that its actual value would be somewhere between 1 s.d. above and 1 s.d. below the mean, since 68 per cent of the area of the curve in *any* normal distribution is in this range. There would be a low probability of it being greater than 2 s.d. above the mean, because less than 2.5 per cent of the area under the curve is that far above the mean. You can see this in Figure 2.3, where the proportions of area within the ranges of 1 and 2 s.d. above and below the mean are shown.

Before moving on, let us consider one further example. Suppose the mean IQ of a sample is 100 and the s.d. is 15. Then, one individual with an IQ of 145 would be 3 s.d. above the mean. This person's IQ would be located at the extreme right-hand side of the curve. This far out the area under the curve is a very small proportion of the total area. Finding an IQ this high in a sample with mean 100 and s.d. 15 is a rare event. The probability for the occurrence of this event is very low, and can be calculated precisely.

Finally, Figure 2.3 includes two new scales drawn beneath the horizontal axis. The first simply replaces each value by its deviation from the mean. These are *deviation scores.* If you summed them, they would add to zero. They are expressed in inches, as that is the scale of the variable represented, i.e. height. Negative values give inches below the mean; positive values are inches above the mean. Do not read on until you have looked back at Figure 2.3 to check the deviation score scale.

Below the deviation scores is a further scale for the horizontal axis. This is marked out in *z-scores.* These have been obtained by dividing each deviation score on the line above by the standard deviation, i.e. the deviations from the mean are no longer expressed in inches, but in standard deviation units.

The crucial point to grasp is that, whatever units are used on the horizontal axis, the frequency distribution above the axis remains unchanged. All we have is three different ways of representing the same thing. These are mothers' height in inches, mothers' height in deviation scores, and mothers' height in *z*-scores (sometimes called standard scores). You can read from these scales that the average mother is 62.5 inches tall, or that her height as a deviation from the mean is zero, or that her height on a scale of standard deviations from the mean is zero.

These three different scales for reporting mothers' height go beyond just a matter of convenience, such as changing from inches to centimetres. To say that a mother is 62.5 inches high tells us just that. But to say that a mother's height expressed as a *z*-score is 0 tells us that in this sample the mother is precisely of average height. A mother with a height of +2.5 on this scale is a very tall woman, and we could say what the probability would be of finding someone that tall by working out what proportion of the total area under the curve is 2.5 s.d. above the mean. Similarly, a mother with a *z*-score of −2.5 would have a low probability of occurrence. In practice, we will not need to calculate these probabilities because they will be the same for any normal curve and can be found in a table in most statistics textbooks. This is an advantageous characteristic of z-scores, not present when the measurements were expressed in the original scale of inches. Another advantage of z-scores is that they provide a common scale for measurements initially made on different scales.

To conclude, consider that Figure 2.3 represents a random sample drawn from all mothers in the UK early in this century. We have calculated the sample mean, and by using the standard deviation and a mathematical model – the normal distribution – we have found a method of calculating the probability for the occurrence of any individual data item in that sample, which at the same time provides a common scale of measurement for any normally distributed variable, i.e. standard deviation units, or z-scores.

How Sample Means are Distributed

In the previous sub-section a method was developed for calculating the probability for the occurrence of any individual data item in a sample for which the mean and the s.d. were known. This involved the simple expedient of re-expressing the scale of measurement as one of z-scores. We have seen what z-scores are in s.d. units. A z-score of +1.96 is 1.96 standard deviation units above the mean. A z-score of −1.96 is 1.96 standard deviation units below the mean. For *any* normal distribution these values mark off the lower and upper 2.5 per cent of the area under the curve. Thus, a value which is outside the range of + or −1.96 s.d. from the mean has a probability of occurring less than five times in every 100 trials. This is usually written as $P < 0.05$.

The important point to keep in mind for the remainder of this section is that we will no longer be considering individual items of data, from which one mean value is to be calculated, but just *mean values* themselves. We are concerned with mean values because we want to know how confident we can be that they are accurate estimates of population parameters.

Can we use the method of the previous sub-section for finding the probability, not of encountering one individual data element of a particular value – one woman 62.5 inches high – but for finding the probability for the occurrence of a *sample mean* of that, or any other, size? Well, obviously, what would be needed to do this is not a frequency distribution of data values, but a frequency distribution of sample means. Using a statistical package called Minitab, I have generated just such a distribution (Minitab, 1985). I have used a random number generator to draw 100 samples, each with n = 1,052, from a population with the same mean and standard deviation as shown in Figure 2.1, i.e. mean = 62.49 and s.d. = 2.435. This is as if I had measured the heights of 105,200 mothers, 1,052 at a time. Just one of these samples is illustrated in Figure 2.4. It has a mean of 62.49 and s.d. is 2.45. Both are close to the population values. A histogram including this mean and means for the other 99 samples can be seen in Figure 2.5 with the horizontal scale (inches) the same as for Figure 2.4, but with the vertical scale (frequency) reduced so that the columns fit on the page.

The mean of the means given in Figure 2.5 is 62.498, much the same as for the full sample, but the s.d. is dramatically reduced to only 0.074. Clearly, a distribution of sample means is more closely grouped around the central value than is a distribution of data values. So that you can see that within this narrower range the individual means do follow a normal distribution, Figure 2.5 has been redrawn as Figure 2.6, with more bars. Note that the range of mean values in Figure 2.6 is from a minimum

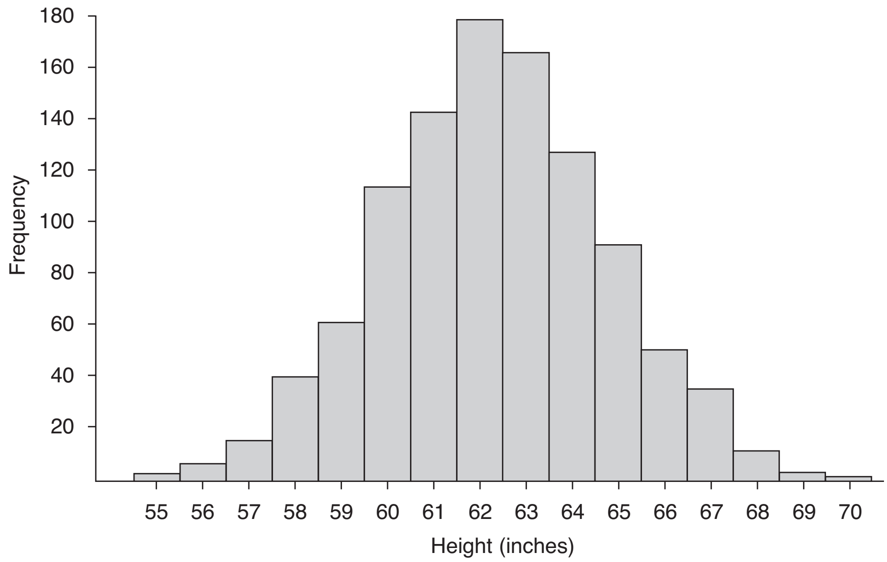

Figure 2.4 *Distribution of mothers' heights in a single simulated random sample, mean = 62.49, s.d. = 2.45*

of 62.25 to a maximum of 62.75, in contrast to the sample data in Figure 2.4, where it is from 55 to 70: a range of less than 1 inch, compared to a range of 15 inches.

Armed with the information in Figure 2.6, we can now put a probability on different means, i.e. determine their relative frequency. We can see, for example, that a mean of 62.75 is in the top extreme of the distribution. This mean would have a z-score of +3.62, and from statistical tables I have found that the probability of obtaining a z-score with this value is only $P < 0.0001$, i.e. only 1 in 10,000. A sample from the present population with this mean would indeed be an exceptionally rare event.

This is all very well, but if a researcher has undertaken a sample survey, and has calculated a mean value for a variable of interest, what would be the good of knowing that if many more samples were randomly selected and a mean calculated for each of them to give a distribution of sample means, then probabilities could be calculated? Fortunately, these are just the kinds of problems where statisticians have come to the aid of researchers and have provided a simple but accurate method of estimating from just one sample of data – provided that it has been randomly sampled – what the standard deviation (which we will refer to as *standard error,* s.e., which is represented here by a capital S) would be for a distribution of sample means from the same population. The formula for doing this, together with an example of the calculations, is:

$$\text{s.e.} = \text{s.d.}/\sqrt{n}$$

i.e.

$$S_{\text{mean}} = s/\sqrt{n}$$

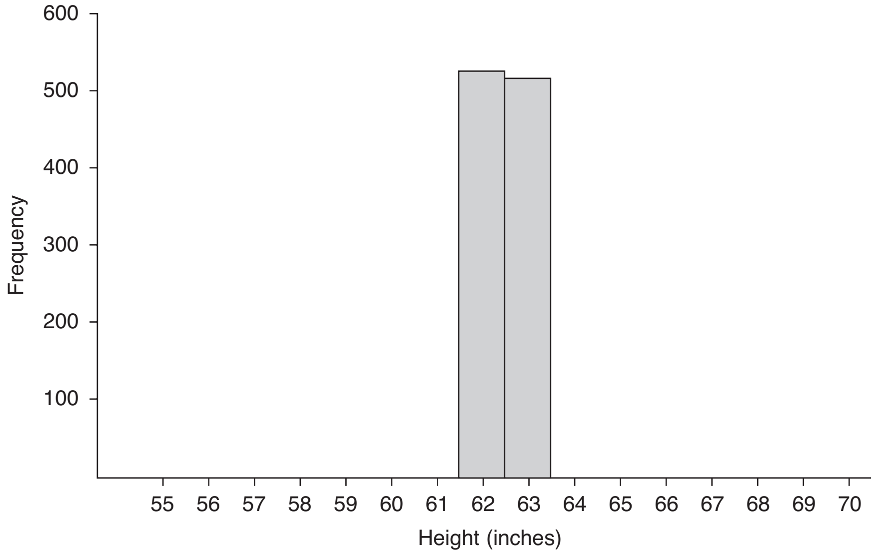

Figure 2.5 *Distribution of mean heights from a hundred simulated random samples, mean = 62.498, s.d. = 0.074*

where *s* is standard deviation and *n* is sample size. We can make these calculations for the data given in Figure 2.4 where the mean is 62.49 and the s.d. is 2.45:

$$\begin{aligned} S_{\text{mean}} &= 2.45/\sqrt{(1{,}052)} \\ &= 2.45/32.4346 \\ &= 0.076 \end{aligned}$$

I hope you will be amazed at just how simple this procedure is. The standard deviation which we have just calculated is given the special name of standard error because it is used to estimate the sampling error associated with one specific sample mean. The standard error of a mean is an estimate of the standard deviation of the distribution of sample means.

Now, since the standard error is the standard deviation of a distribution of sample means and these are normally distributed, then 95 per cent of the values in that distribution are within the range of $\pm$ 1.96 standard deviations from the mean, i.e. approximately $\pm$ 2 s.d. from the mean. In the present example that gives a range from approximately 0.15 below to 0.15 above the mean (2×0.076), i.e. from 62.34 to 62.64, from our one sample. We can be 95 per cent confident that the true population mean will be somewhere within this range. Notice how close the value we have just calculated by weighting the s.d. of one sample by the square root of the sample size, i.e. 0.076, is to the true s.d. of the distribution of sample means, which we do have in this case, i.e. 0.074.

Often, survey findings are expressed not as mean values, but as proportions or percentages. For example, a finding might be that 36 per cent of all households use brand X to wash the dishes. Assuming that this assertion is based on a sample

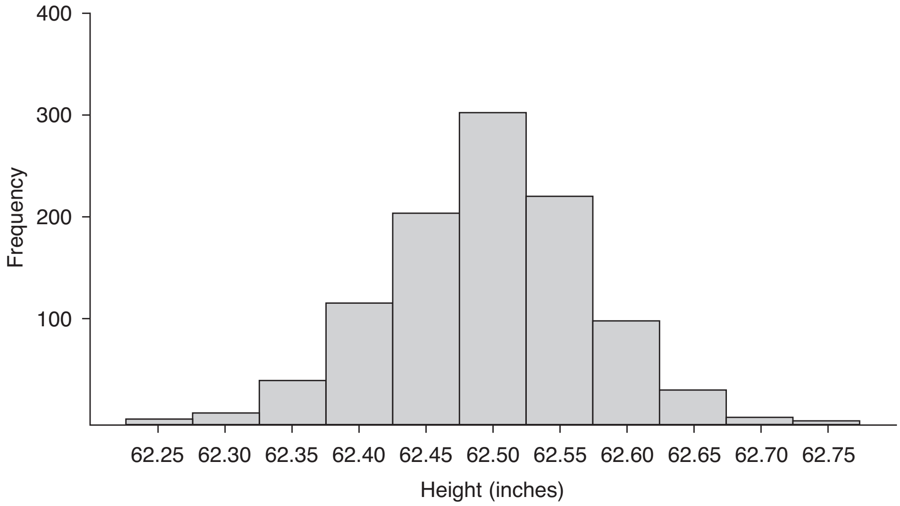

Figure 2.6 *Distribution of mean heights from a hundred simulated random samples, scaled for more detailed display*

survey for which 1,000 households were sampled, what would the likely margin of error be? Provided sampling was by a probability method – for example, simple random sampling – then an unbiased standard error of a proportion can be calculated. It will not, however, be pursued here. The theory behind calculating descriptive statistics and standard errors for proportions is harder to follow than is the case for means. This is because it involves using what is known as the *binomial distribution.* Every sample statistic – the mean, median, mode, the standard deviation itself, a total or a proportion – has a standard error which can be estimated from just one random sample when it is needed. As we have seen, knowledge of standard error enables statements about probabilities. For example, comparing how far apart two means are in terms of standard errors provides a test of whether the two means differ by more than chance.

Formerly a lot of time could be spent learning formulae and calculating confidence limits and significance levels using standard errors. However, in research, calculations are now done by computer. Thus the details of statistical formulae are not important to survey researchers. What is important is that the method being used and its assumptions are fully understood. In this case you need to understand thoroughly what is meant by a *distribution of sample means,* and how the standard deviation of this distribution can be estimated from one sample. Further, you need to understand the use to which this s.d. is put in its role as standard error of just one sample mean.

A final important point should be mentioned, although space does not permit it to be developed in any way. Very often in social and behavioural science, data distributions do not look at all normal, and it might seem that procedures which assume a normal distribution cannot be used. However, first, various things can be done about this, including transformation of the scale to one which is normal for the measure concerned. Secondly, the relevant mathematical models require that the *underlying* variable which is being measured is normally distributed, and it is to be expected that individual samples (especially small samples) will look rather different. Thirdly, and this is the most fortunate point of all, even if the population distribution is far from normal, as sample size increases the distribution of sample means from that population will move closer and closer to the desired normal form, thus permitting valid statistical inferences to be made about those means. This statement rests on a fundamental theorem in statistics (the central limits theorem). This theorem justifies much of the data analysis undertaken when quantifying and estimating the reliability of survey research findings (Stuart and Ord, 1987).

Error, Sample Size and Non-response

As was mentioned in the discussion of Activity 2.1, there are two categories of error in survey research: sampling error and non-sampling error. Conceptually, the terms sampling error and non-sampling error refer to different entities, and it is theoretically important to consider them as such, but in practice we can never have a true measure of sampling error, but only an estimate of it, and the influence of non-sampling error is hopelessly confounded within that estimate. Both researcher and research evaluator have to ensure that non-sampling error is avoided as far as possible, or is evenly balanced (non-systematic) and thus cancels out in the calculation of population estimates, or else is brought under statistical control. As has been shown, the difference between sampling error and non-sampling error is that the extent of the former can be estimated from the sample variation, whereas the latter cannot. Further, we have seen that sampling error can only be reliably estimated if the selection of respondents has been random. At best, random sampling will allow unbiased estimates of sampling error; at worst, quota and opportunity sampling will provide little or none.

In practice, researchers often overlook or are unaware of these difficulties and quote standard errors, i.e. estimates of sampling error, even for samples where their use is not justified in theory. But at least sampling error *can* be calculated, whether appropriately or not, and if sufficient information is provided then judgements can be made about just how much reliance can be placed on it. The various sources of error grouped together as non-sampling errors are another matter – not because they will be necessarily greater in extent, although this could well be the case, but because they are difficult to control, or even detect. The great virtue of randomization is that it takes care of potential sources of bias, both known and unknown. If it can be assumed that error, whatever its source, will be randomly spread across a sample, and will cancel when statistics are computed, then one does not even need to know what it is that is cancelled. The problem is systematic, non-random error, which will not cancel.

Non-sampling error is often overlooked when survey findings are evaluated, and if an estimate of sampling error is given, then it is often wrongly assumed that this shows the likelihood of total error. For example, in the 1992 General Election in the UK, one survey predicted a Labour Party vote of 42 per cent ± 3 per cent. Presumably the figure of 3 per cent represents approximately twice the standard error, and thus the 95 per cent confidence range for this result would be from 39 per cent to 45 per cent. This says that, if the same pollsters drew sample elements in exactly the same way, and questioned and recorded in exactly the same way, from the same population, a sample of the same size, then they could expect to obtain a value in that range 95 times for every 100 samples so drawn and tested.

However, this statement tells us nothing whatsoever about whether the sampling frame truly represented the voters of the UK overall, let alone the more restricted set of those who actually did vote. It tells us nothing about interviewer bias, or procedural reactivity, or untruthfulness on the part of respondents. If one took a guess and allowed another 3 per cent for all of these, then the predicted range would increase to 36–48 per cent, which would greatly decrease the usefulness of the survey finding, since in a moderately close election it would be very unlikely to predict successfully which way the outcome would go, because the estimates for the two major parties would always overlap. An advantage the pollsters do have, however, is replication. Many polls are taken, and by different organizations. Taking all into account might give some possibility of balancing some sources of non-sampling error – but not all. It could, of course, be the case that all the polls suffered from similar defects, in which case pooling would not cancel the bias and prediction would be highly unreliable.

Major sources of non-sampling error related to the sampling process itself include: sampling-frame defects, non-response, inaccurate or incomplete response, defective measuring instruments (e.g. questionnaires or interview schedules), and defective data collection or management. Some of these are the subject of other chapters in this book, but their relevance here should also be kept in mind. Many of these effects are, or can be, controlled by proper randomization in sampling. For example, in a large survey the error related to small differences in technique on the part of interviewers (perhaps consequent upon personality differences) will be randomly spread across the data, and will cancel out. Any residual effect should be small and would be lost in the estimates of standard errors, possibly here balancing with other small residual effects.

Sample Size

Often, selecting an appropriate sample size has been a hit and miss business of choosing a size which can be managed within the resources available, or a size similar to that used in earlier published work. There is a misconception that sample size should be related to the size of the population under study. As has been shown above, the precision of sample estimates depends very much on sample size (the sample s.d. is divided by the square root of the sample *n)* and no reference is made to the size of the population sampled.

Assuming that for a sample survey the 95 per cent level of confidence is required *(P* < 0.05), and the maximum error is set to 5 units on the scale of measurement, then the following formula will provide the estimated sample size:

$$\text{Sample size} = 2 \times 1.96(\text{s.d.})^2/5^2$$

If the estimated s.d. is 10, then the required sample size would be approximately 16. If the limit for the difference of interest was reduced from 5 to 2 points, then estimated sample size would increase to close to 100, assuming that the s.d. remains unchanged. Note that the researcher had to provide an estimate of the s.d., although the actual value will not be known until the research is concluded.

This is a very simple account of what might appear to be a simple subject, but which in fact is a complex one. Just how big a sample should be is a matter of balancing cost against the level of precision required. True, as sample size increases, the size of the standard error of any estimate of a statistic does decrease. But this needs to be qualified by knowledge that large samples may introduce more non-sampling error (as mentioned in the answer to Activity 2.3) than smaller ones, where measurements and management problems may be smaller. Also, the power of the statistical test to be used must be taken into account and tables have been published for doing this (Cohen, 1969; Lipsey, 1990). Many computer packages now also include routines for dealing with this.

Non-response

Estimating the required sample size needed for a stated level of precision has been discussed. There is, however, little point in reporting that sample size was formally determined to achieve maximum precision, if a sizeable proportion of the sample was subsequently lost through non-response, or because items of data were missing. This is a major source of error in many surveys.

Procedures for dealing with non-response and missing data have to be established when the research is being planned, and not left to desperate *post hoc* remedy. In establishing such procedures, total non-response should be distinguished from failure to respond to individual items in a questionnaire, and both should be distinguished from data which are simply missing (i.e. lost or inadequately recorded). Preliminary data analysis will also lead to further data loss, usually due to the dropping of elements (individuals or cases) found to have been included in the sampling frame by mistake, but which do not belong to the population studied, or because inconsistent or highly improbable values have been found on crucial variables.

Final reports should contain information on the extent of sample loss and missing data, which amounts, at least in part, and sometimes completely, to the same thing. Non-response rates as high as 50 per cent or more have frequently been reported. Some elements of the sample simply will not be found, others will refuse to participate, either completely or in part. In addition, data, and sometimes whole subjects, will be lost due to clerical inaccuracy. The extent of data lost for this reason alone is seldom reported, but is usually surprisingly high, perhaps as much as 8 per cent (Schofield et al., 1992). Response rate is influenced by such design matters as the appearance of a questionnaire, its layout, length and readability. These topics are dealt with in more detail elsewhere in this book. Information on such matters will be sought in pilot studies, in which different versions of a survey instrument can be tested. Sample loss for these reasons is likely to introduce bias because it might increase the proportion of more persistent or better educated respondents.

If the survey involves home interviews, non-response might be related to time of day at which the interview was sought. From Table 2.3 it can be seen that a higher proportion of persons over 14 years of age are at home in the early hours of the evening on Sunday, Monday and Tuesday than at any other time. This, however, is also evening meal time, and a busy time for families with young children. Again, sample loss could be systematic, and it could introduce bias.

If, when a survey is being planned, it seems likely that response rate will be low due to the nature of the information sought, or the accuracy of the sampling frame,

Table 2.3 *Proportion of households with at least one person over 14 years of age at home, by day and time of day*

	Proportion by day of week						
Time of day	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.
8.00–8.59 am	(B)	(B)	(B)	(B)	(B)	(B)	(B)
9.00–9.59 am	(B)	(B)	(B)	0.55	0.28	0.45	(B)
10.00–10.59 am	(B)	0.47	0.42	0.38	0.45	0.40	0.55
11.00–11.59 am	0.35	0.41	0.49	0.46	0.43	0.50	0.62
12.00–12.59 pm	0.42	0.53	0.49	0.56	0.45	0.55	0.60
1.00–1.59 pm	0.49	0.44	0.50	0.48	0.43	0.51	0.63
2.00–2.59 pm	0.49	0.50	0.52	0.47	0.45	0.45	0.59
3.00–3.59 pm	0.54	0.47	0.49	0.54	0.50	0.50	0.65
4.00–4.59 pm	0.52	0.58	0.55	0.57	0.57	0.56	0.53
5.00–5.59 pm	0.61	0.67	0.65	0.67	0.59	0.57	0.56
6.00–6.59 pm	0.75	0.73	0.72	0.68	0.65	0.64	0.59
7.00–7.59 pm	0.73	0.74	0.75	0.64	0.61	0.57	0.66
8.00–8.59 pm	(B)	0.51	0.51	0.59	0.74	0.52	(B)
9.00–9.59 pm	(B)	(B)	(B)	0.64	(B)	(B)	(B)

(B) = base less than 20.

Source: Weeks et al. (1980).

or the method used to contact respondents, then sample size could be increased. This might seem to be an obvious and easy solution, but it will be costly in terms of management and material and, in any case, will be unlikely to solve the problem.

Activity 2.5 (10 minutes)

In the planning of a sample survey by questionnaire sent by post it has been calculated that a sample of $n = 200$ will give the required level of precision for estimating population means at the 95 per cent confidence level or better for most items of interest. But only about 40–50 per cent of those sent questionnaires are expected to return them. The researchers propose simple random sampling with sample size increased to $n = 400$. Comment briefly on this proposal. If you were an adviser, what advice would you give?

Increasing sample size to cover expected non-response would, in fact, be more likely to increase than to decrease bias. More money would be spent to no avail. Studies have shown that non-responders tend to be: the elderly; those who are withdrawn; urban rather than suburban, or rural, dwellers; individuals who fear that they will not give the information adequately in comparison to others, or who fear that they might expose themselves, and be judged in some way by the responses they make. To lose such individuals selectively would very likely reduce the representativeness of a survey sample. To increase sample size while continuing to lose such individuals would in no way help, and could lead to comparatively stronger influence from, perhaps, initially small biasing groups.

Whether information is collected by questionnaire or by interview, positive effort should be made to follow up non-responders. Even when second copies of a questionnaire are sent out, or repeat interviews arranged, response rates above about 80 per cent are seldom achieved. The task for the researcher, who wants sample results which truly represent the population studied, plus information which will help evaluate how far this objective has been achieved, is to get as much information as possible on those individuals who are still missing when all possible action has been taken to maximize response rate.

For this reason, the records of individuals who have made only partial, or even nil, response should never be dropped from a data set. Usually, information will be available on some variables; for example, geographical region, home address, perhaps age or sex. Analyses can be made to see if the missing individuals are at least randomly distributed throughout the sample in terms of these measures, or grouped in some way which might help identify the possible direction and extent of bias on other measures for which there are no data.

Even better would be a small follow-up survey of a random sample of non-responders, possibly involving home visits and/or the offer of incentives, so that reliable predictions can be made about the likely characteristics of all non-responders. In some circumstances this could be counter-productive, in that interviewer/respondent reactivity might be increased. One way or another, however, the problem of non-response has to be tackled. Vagueness or, worse, total lack of information on this topic, is no longer permissible.

Conclusion

This chapter has dealt with methods and problems of designing sample surveys, and has related these to the wider research context, where ultimately the validity of findings will rest on how well the sample represents the population being researched. We have seen that the quality of the inferences being made from a sample will be related to both sample size and sampling method. We have seen that, provided a probabilistic method has been used, then a reliable estimate can be made of the extent to which the sample results will differ from the true population values, and that error of this type is known as sampling error. The methods discussed included both simple and stratified random sampling systematic sampling, and cluster sampling, and also non-probabilistic methods such as quota sampling. Selecting the best method for any particular research will usually involve compromise, and will be a matter of balancing the level of precision required, in terms of the width of the error estimates, against feasibility and cost.

We have also seen that error from many other sources – non-sampling error – will have to be taken into account when planning survey research and

(Continued)

(Continued)

when evaluating results. Major sources of non-sampling error which have been discussed in this chapter include faulty selection of the sampling frame and non-response. There are many others, including the instruments used for collecting information: schedules, questionnaires and observation techniques. The problem for researchers is that, however well they plan the technical side of sampling and calculate estimates of sampling error of known precision, non-sampling error will always be present, inflating overall error, and reducing representativeness. Estimating the extent of this is a matter not of mathematical calculation, although statistical procedures can help, but of scientific judgment, based on an awareness of what problems are likely, as well as common sense.

Key Terms

Bias aspects of measurement or sample selection which tend to increase the difference between sample statistics and the population parameters.

Census a study including (or intending to include) all elements of a population, not just a sample.

Cluster sampling sampling which selects groups of elements based on geo-graphical proximity.

Element a single case (item) in a population or a sample.

Error *see* Sampling error.

Mean the average of a distribution – calculated by adding all the values together and dividing by the number of cases.

Non-probablistic *see* Probablistic sampling.

Non-sampling error *see* Sampling error.

Population the total set of elements (cases) available for study. Your population might be people – all the people in the UK, or all the children in one school, or all the children in a specified age range in a certain district – but it could be incidents, or cars, or businesses, or whatever is being studied.

Probablistic sampling sampling in which elements have known probability of being chosen. Samples where this is not the case are known as 'non-probablistic'.

Quota sample one collected by interviewers who have been instructed to ensure that the cases they collect match a predetermined distribution on certain key variables (often the known population parameters).

Random sample one for which every element of the population is guaranteed an equal, non-zero chance of being selected.

Sample elements selected from a population, by studying which we hope to understand the nature of the population as a whole.

(Continued)

Sampling error the calculable probability of drawing a sample whose statistics differ from the population parameters. This is contrasted with 'non-sampling error', which is bias built into the design, the sampling frame or the measurement and is not a

Sampling frame a complete list of the elements in a population.

Standard deviation a measure of spread or dispersion from the mean, based on the normal distribution.

Stratified random sample one made up of separate random samples, drawn from sets which together make up the entire population.

Systematic sampling a sample that consists of every *n*th member of a sampling frame, perhaps from a random starting point.

***z*-score** the distance of an element form the mean, measured in standard deviation units.

Further Reading

Lipsey, M.W. (1990) *Design Sensitivity,* Newbury Park, CA, Sage.
A short and fairly readable text on statistical power in social science research. It includes charts for determining sample size.

Moser, C.A. and Kalton, G. (1971) *Survey Methods in Social Investigation,* London, Heinemann.
Although somewhat dated, this remains a standard text for material covered in this chapter.

Schaeffer, R.L., Mendenhall, W. and Ott, L. (1990) *Elementary Survey Sampling,* Boston, PWS-Kent.
A further elementary text, recently revised, which includes some of the mathematical derivations of sampling methods, but with many practical examples of surveys and methods. Useful later if you have the task of designing a survey.

Answers to Activities

Activity 2.1

The main advantages of a sample survey over a full census is that it will be easier and cheaper to set up, manage and analyze than a full census. Although the results based on a sample will, in theory, be less accurate than if the whole population had been included (assuming this to be possible), this might not be the case in practice. Many sources of bias – for example, management problems, faulty measurement, lost or corrupted data – will potentially be present whichever method is used, and will be easier to control in a tightly constructed and managed survey than in a full census.

Activity 2.2

Table 2.1, with the missing entries added, is shown below. The first of these, in the fourth column, is the sample size *(n)* for proportionate sampling. This was found by calculating 45.5 per cent of the total sample of 400. This gave a sample proportion of $n = 182$ for the representation of 16-year-old school leavers. Similar calculations were made to find the other sample proportions. For the disproportionate method, the total sample was divided into three equal groups, one for each school-leaving age, without taking into account the differing incidence in the population of each of these groups.

Compare your note on the sampling method you would choose with the explanation given in the three paragraphs following the activity in the text, where several non-statistical reasons are given for balancing the various alternatives.

Activity 2.3

Clearly, a probabilistic method would be preferable, since this would permit a valid estimate of the extent of sampling error. As the population of interest is all the adults in the UK, a simple random sample would be costly and difficult. Precision relative to sample size could be increased by appropriate stratification, and thus you would recommend a stratified random sample.

Activity 2.4

The mean of the first set of figures is 3, and the mean of the second is also 3. The two standard deviations are, respectively, 1.414 and 4.517. In other words, the two sets have the same mean but very different standard deviations because they differ greatly in the way the individual values are distributed about the mean. The average of this dispersion (the standard deviation, s.d.) is much greater for the second set than for the first.

Activity 2.5

You will probably want to accept the decision to use random sampling, provided that an appropriate sample frame is available, and also that there is sufficient finance to cover the cost of obtaining a sample of the size needed for the required precision. You will then point out that, with an expected response rate of 40–50 per cent, the sample is not likely to be representative of the population of interest, as the non-responders are likely to differ in important ways from those who do respond. Merely increasing sample size will be costly, and will not help. You would suggest that the additional money should be spent instead on making a second, and even a third, approach to non-responders; doing analyses on whatever limited data are available for non-responders to see how they differ from those who do respond; or setting up a small random study of the characteristics of non-responders, perhaps by visiting their homes, or offering an incentive for participation.

Research Proposal Activity 2

This chapter has introduced methods for obtaining representative samples from a population and for providing estimates of the accuracy of statistics derived from a sample. It has also examined issues and problems relating to survey sampling, sample loss and non-response. In outlining a research proposal, you should consider whether the following questions are relevant:

1 What is the population to which you want your results to apply? What are the sampling units in that population?
2 Is a sampling frame available? Is it complete, accurate and fully representative of the population you wish to describe?
3 What methods of sampling will be used (random or non-random) and why?
4 If random sampling is to be used, should stratifying factors be introduced? If so, will the sampling be proportionate or disproportionate to population frequencies?
5 Is there value in adopting a cluster-sampling approach?
6 If random sampling is not feasible, or too costly and time-consuming, is some form of non-random sampling more appropriate?
7 What sources of non-sampling error can be anticipated and how can they be counteracted at the planning stage?
8 What should the size of the sample be? What balance should be sought between cost and level of precision?
9 What steps will be taken to deal with the bias introduced by non-response or missing data?