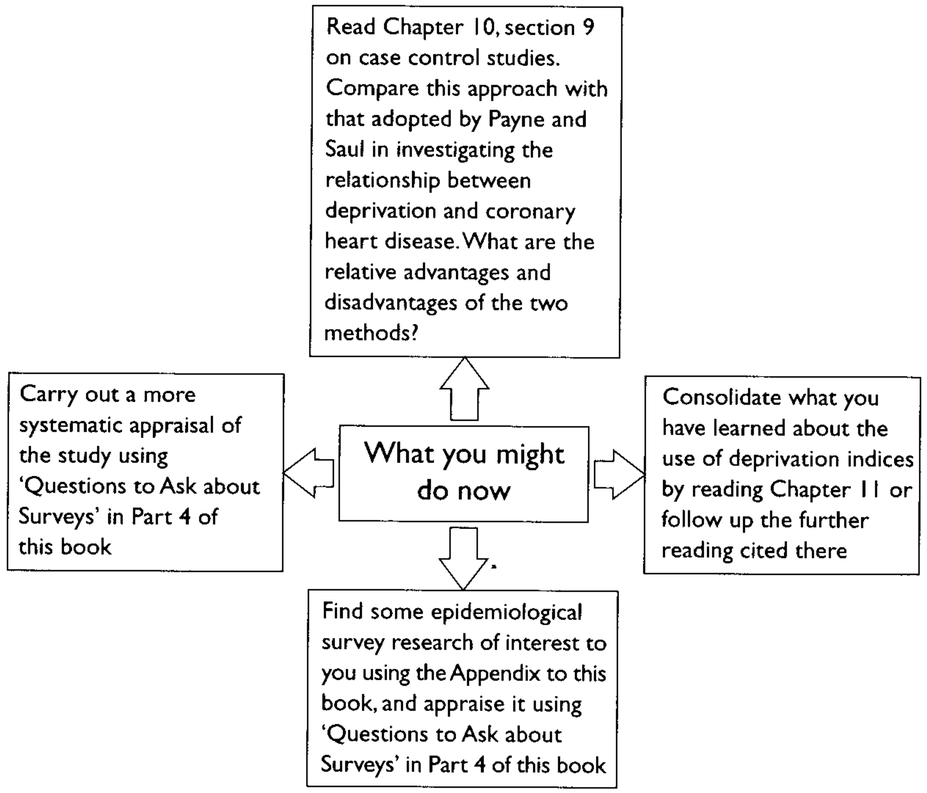
26 Thomas M., Goddard E., Hickman M., Hunter P. *General Household Survey 1992*. London: HMSO, 1994. (OPCS Series GHS No 23.)
27 Roberts H., Dengler R., Zamorski A. *Trent Health Lifestyle Survey Report to Sheffield Health Authority 1993/94*. Nottingham: Department of Public Health Medicine and Epidemiology, 1994.
28 Elder A.T., Shaw T.R.D., Turnbull C.M., Starkey I.R. Elderly and younger patients selected to undergo coronary angiography. *BMJ* 1991; 303: 950–3.

What you might do now



RESOURCE CHAPTERS

CHAPTER 10

SURVEYS AND CASE CONTROL STUDIES

Introduction

This volume contains two exemplar studies with surveys as important components. Chapter 8 offers an exercise by Geoff Cohen and his colleagues in validating the sampling procedures and the instruments (questionnaires) used in Scottish surveys of consumer satisfaction with the NHS. General remarks about the validation of instruments are made in Chapter 6. Chapter 9 presents an exemplar of service evaluation by Nick Payne and Carol Saul using, among other research techniques, a survey to chart the social distribution of angina symptoms. This chapter now provides some general comments about survey technique and how to read the results of surveys.

The purpose of a survey is to chart frequency distributions in a population. These might be the percentages of people in the population of the UK of different ages with limiting and long-standing disabilities or perhaps the numbers of people of different types who are satisfied with the primary care they receive (Chapter 8). Such data may be put to use in:

- Planning services, for example, how many people of what different types would benefit from and appreciate this kind of care?
- Evaluating policies and interventions, for example, how many people of what types engaged in health-damaging behaviours before the health promotion campaign and how many afterwards

(Tudor Smith et al., 1998)? Are the people receiving a service also those with the greatest need for the service (see Chapter 9)?
- Measuring time trends, for example repeating the same survey after a period of time to see whether people are more or less satisfied with the NHS (see Chapter 8). A *repeated survey*, such as the NHS Users' survey, usually repeats the same survey with different people at different time periods.
- Investigating causality, for example, is there something about being black in Britain which undermines mental health (Nazroo, 1997)?

It is perhaps worth noting that in medical and nursing research surveys are sometimes described as *observational studies*. This is part of a classification of research methods which distinguishes experimental research on the one hand from survey and qualitative research on the other, both of the latter being called 'observational'.

1 Sampling and representativeness

Most surveys are *sample surveys*. The major exceptions are the national ten-year censuses which attempt to collect data about everyone, and in-house evaluation exercises where an agency will attempt to poll all its clients. For a sample survey, the sample needs to be selected so that it represents some wider population. If this is accomplished successfully the survey researcher can claim that what was found in the sample will be true also of the population from which the sample was drawn; that is, the results will be *generalisable* to that population, at least at some point in time close to when the survey was carried out. Box 10.1 gives a synopsis of various techniques used to select representative samples. The same techniques might be used to select samples for experiments, although experiments often use convenience samples (see Chapter 5, section 12).

Representativeness is not an all or nothing matter. A relatively small sample can be representative of, for example, the sex ratio in a wider population. Here a random sample of 384 would be sufficient to find the sex ratio of a population of one million, to an accuracy of plus or minus 5 per cent with a 95 percent chance of being right (see Table 10.4 below).

The size a sample needs to be in order to achieve representativeness depends on the level of diversity which is of interest to the researcher. The minimum sample size is thus given by the number of possible unique combinations of responses which the survey will generate and in which the researcher is interested. Thus, for instance, in a three-question questionnaire, each with two possible responses (say, Yes and No), there are eight possible permutations of answers: YYY, YYN, YNY, YNN, NYY, NYN, NNY, NNN. But if the researcher is interested

Box 10.1 Representative samples for surveys

There are two basic ways in which representative samples for surveys are collected and variations on these.

1. Probability samples (random and systematic sampling) (Bowling, 1997: 163–6; Alston and Bowles, 1998: 83–9). Each person in the population of interest has an equal chance of being chosen – or as near equal as possible. This presupposes some listing (or *sampling frame*) from which people can be chosen which lists everyone in the population. Since complete sampling frames are rare, some kinds of people get excluded at the outset, reducing the representativeness of the sample. **Random selection** is usually done with a table of random numbers or a computer program that generates random numbers. Sometimes **systematic sampling** is used: for example, every seventeenth name on a list (Layte and Jenkinson, 1997: 48–51). In this example, 17 would be the **sampling interval**. So long as there is no feature of the list which makes, say, every seventeenth name more or less likely to be a particular kind of person, systematic samples are as good as random samples, and they have an added advantage of spreading the sampling evenly within the population (Arber, 1993: 79–80).

Since not all people chosen will be contactable and not all will cooperate, a sample which starts out as representative may become unrepresentative through **non-response**. Deviation from representativeness is often checkable by comparing the sample of respondents with the population from which it was drawn in terms of the known demographic characteristics of the population, such as its age profile or gender ratio.

Very large **unmodified** (or **simple**) **probability samples** are needed to represent diversity within subgroups or to recruit adequate numbers of people to represent groups which are in the minority in the population. To ensure adequate representation of minority groups **random stratified sampling** (or **non-proportional random sampling**) is sometimes used: for example, collecting samples of the same size from each ethnic group in the area, even though some ethnic groups only make up a small percentage of the population (Layte and Jenkinson, 1997: 48–51). The same principle may be used to ensure that the sample adequately represents people across a geographical area, or represents people from a wide range of agencies.

Two advantages of probability sampling over quota sampling (below) are that the probability sample is more likely to be representative with regard to previously unknown characteristics and that results can be subjected to statistical analysis in ways that those from quota samples cannot.

2 Quota Sampling (Bowling, 1997: 166–7; Alston and Bowles, 1998: 91–2). Researchers need to have a good knowledge of the structure of the population in advance of doing the research. Quotas are lists specifying the respondents who need to be recruited in order to build a sample that is a small-scale model of the population. Thus, if the population has 5 per cent black males between the ages of 15 and 25

then the instruction will be to find the number of such people needed to make up 5 per cent of the sample. Filling the quotas is on a first-found, first-in basis, so there is no non-response. Despite this, however, there are problems of deciding how far those who fit the criteria for a quota and are included are representative of all those who would fit the criteria of a quota. For example, those who become respondents may be unrepresentative in being more accessible or more cooperative than others, and in other ways associated with their accessibility or cooperativeness. The problem of representing minorities adequately in a quota sample can be solved by setting quota percentages disproportionate to percentages in the population, creating the same effect as random stratified sampling (see above).

The advantage of quota sampling over random sampling is that it requires smaller samples and is thus cheaper. Hence it is the method used by most public opinion polling companies. The major disadvantage is that only a limited amount of statistical analysis is possible with samples that have not been drawn using probability principles (Pett, 1997: 13–17).

See also **cluster sampling** (section 4), **staged sampling** and **phased sampling** (section 5).

in the distribution of these permutations between people of different genders, there are 8×2 possible unique permutations, and if interested in the pattern of answers by gender, age (five age groups), social class (three classes) and by ethnicity (five ethnic groups), that is $8 \times 2 \times 5 \times 3 \times 5 = 1{,}200$. If each of these permutations cropped up with equal frequency in the population, then a sample size of around 60,000 would be needed to represent the frequencies with which these patterns were to be found in the wider population. But if some permutations were rare, and they were none the less of interest to the researcher, then an even larger sample would be needed to give rare permutations a chance of being captured by simple random sampling. But a sample of 60,000 is already much larger than most survey researchers can afford, and few restrict themselves to asking just three questions, or allowing only two answers each. There are two main approaches to this problem.

1 A survey may be analysed as if it were a series of surveys all conducted at the same time with the same sample: one survey studying the relationship between responses and gender, one between responses and age, one between responses and ethnicity: or one survey looking at age, gender and ethnic differences in responses to question 1, one looking at these factors in relation to question 2, and so on; the analysis being unable to say what relationships exist between giving a particular combination of answers on questions 1, 2 and 3, and being, say, male, African Caribbean and between 25 and 45 years old.

2 Survey researchers may (and may have to) concentrate on broad patterns and ignore fine detail; for example, by looking only at common answers and consigning the remainder to the category 'other answers', and/or looking only at larger groups and consigning the remainder to the category 'other groups', as in White/African-Caribbean/South Asian/Other.

Table 10.1 gives a synopsis of the more common kinds of problems which may arise from the unrepresentativeness of samples. In each

Table 10.1 Problems of representativeness in surveys

Shortcomings in design	Problems in interpreting the results
Poor selection of clusters in cluster sampling – often adopted as a way of avoiding the expense of interviewing large numbers of people scattered widely across the country (see section 4)	There is usually a problem in deciding how far the clusters are representative of the wider population of areas, agencies, institutions etc., which they are supposed to represent, and hence a problem of deciding how far individuals selected from within clusters make up a representative sample of the wider populations of individuals
An incomplete sampling frame. Those omitted might be different in significant and relevant aspects from those included (see section 2) A sample drawn in a way other than those which ensure representativeness (see Box 10.1): for example, a convenience sample of clients known to services presented as standing for all people with a particular problem, or a poll of members of a service user pressure group presented as standing for all service users. Case control studies show these problems too (see section 9) A high non-response rate. The non-responders may be different in significant and relevant respects from the responders (see section 8) Non-reponse to some questions Attempts to generalise from the sample to a population that is not the population from which the sample was drawn (see also **cluster sampling** above)	There will be problems about generalising the results of the survey to the population from which it was drawn. That is, what was true for the sample *will not be true* for the population at which the generalisation is directed. The extent of the problem will depend on the size and composition of the group who should have been included, but were excluded, and/or the extent of the difference between the population and the sample. Problems arising from non-response to some questions apply to those questions only
A sample too small adequately to represent the population with regard to relevant characteristics (see sections 6 and 7) Asking questions allowing for more responses than sample size will cater for	The necessary size of a sample depends on the degree of diversity made relevant by the analysis attempted (see this section and section 7). The more sub-categories the sample is to be divided into, and/or the more options are available for answers, and/or the smaller the differences of interest between categories, the bigger the sample needs to be

case the problem is one of generalisability because only insofar as a sample is representative of the population from which it is drawn can the results for the sample be regarded as true for the population (within the confidence limits cited). On a smaller scale only if a sample is representative for its sub-groups will what is true for the sub-groups in the sample be true for the same sub-groups in the population.

2 Sampling frames

A *sampling frame* is a listing from which a sample can be chosen. The comprehensiveness and accuracy of the list influences who might be included in the sample. For example, in the Lothian survey reviewed in Chapter 8, the sampling frame was a listing of all adults registered with GPs. That would exclude the 4 or 5 per cent of people not registered, as well as include some people who were dead or had moved away where this had not been corrected on the register. Since these would be uncontactable, they would become part of the non-response of the survey (see section 8).

Two of the most common kinds of sampling frame used in health and social care research are various service registers (including GP practice lists and medical registers) and the Post Code Address File (PAF) maintained by the post office (Wilson and Elliot, 1987). Researchers tend to use the sub-set of the PAF called the 'Small User File', which identifies addresses receiving mail, but receiving less than 25 items per day, thus excluding most business addresses. This was used as the sampling frame for the NHS Users' Survey reviewed in Chapter 8. The sampling unit for the PAF is an address, hence a further sampling decision has to be made as to which person living at the address becomes the respondent. A 'Kish grid' is the tool usually used to make a random selection of household members (Kish, 1965). Post code addresses can be identified with the territorial units from which census (and other) data are collected. It is census data from which deprivation indices are constructed (see Chapter 11, section 4). Thus, using the PAF as a sampling frame makes it convenient to stratify a sample in order to include a range of addresses representing the spectrum from very affluent to very poor areas (section 4). This was the way in which the NHS Users' survey (Chapter 8) stratified respondents by social class.

3 Stratified (or non-proportional) samples

Probability sampling (Box 10.1) is more widely used in health and social care research than quota sampling, because of the amenability

of the results to statistical analysis. The principle of probability sampling is that each member of the relevant population has an equal chance of being chosen. However, this means that minorities in a population have a lesser chance of being included in the sample. In a population of 50,000 where there are only 500 Chinese people, a sample of 1,000 only gives 10 chances for a Chinese person to be included in the sample. It is impossible for ten Chinese people to be representative of all Chinese people in the population in terms of age, gender, opinions and so on. This problem is often handled by *stratifying* the sample, that is by taking non-proportional random samples. Thus the problem of getting an adequately sized sub-sample of Chinese people could be solved by stratifying by ethnicity.

The large scale Office of Population, Censuses and Surveys (OPCS) survey of psychiatric morbidity (Meltzer et al., 1995) did not stratify by ethnicity. The overall sample size was over 10,000, but this still only included about 460 from all ethnic minorities of colour. This accurately represents the number of such adults in the population of Great Britain (about 4.6 per cent), but it is much too small a sample to give meaningful results when this category is subdivided by ethnic group, gender and mental health status. By contrast, Nazroo's survey (1997), which was designed to emulate features of the OPCS survey, did stratify by ethnicity, selecting a random sample of 5,106 members of ethnic minorities of colour, and another of 2,867 white people. Even so, the sample was too small to provide an adequate representation of people of Chinese origin, or of white people with family origins outside Britain. The NHS consumer surveys reviewed by Cohen and his colleagues (Chapter 8) also used stratified approaches to recruiting samples. For example, the sample for the Scottish Users' survey was stratified by health board areas. This was to ensure a sample giving adequate coverage to all the health boards equally in the face of the fact that some have bigger populations than others. It was also stratified by social class, to ensure an equal representation of people from different social classes when different social classes make up different percentages of the population.

When sampling involves stratification, it is important to check whether the results are quoted for the sample as recruited or after re-weighting them back to proportionality. For example, the Lothian health survey reviewed in Chapter 8 recruited equal-sized samples from a range of age groups, despite the fact that these age groups made up different percentages of the Lothian population. In order to compare the results of this survey with that of the all-Scotland NHS Users' survey, Cohen and his colleagues had to *(re-)weight* the results of the Lothian survey so that each age group only contributed to the overall results in proportion to the percentage of the population each constituted.

Stratification in sampling is sometimes referred to as 'weighting'. However, as Box 10.2 shows, this term is used in a variety of ways in survey research.

Box 10.2 'Weighting' in survey research

The term 'weighting, as in 'age-weighting', is used in at least four different ways in survey research:

- Weighting a sample to ensure that important categories of respondent get included in sufficient numbers – this is better termed *stratification* or *disproportionate sampling* (see above).
- (Re-)weighting the results of a survey to compensate for higher levels of non-response by some categories of respondents (see section 8).
- Weighting the results of a survey in order to apply the results to another area with different demographic characteristics. Where this involves a reference population this is better referred to as *standardisation* (see Chapter 11).
- Weighting the results of a survey in order to control for the effects of some variable; better referred to as *statistical control* (see section 11 of this chapter and Chapter 11).

A case study of stratified sampling is given in Section 7.

4 Cluster sampling

Imagine a researcher wanting to select a national sample of 2,000 people living in nursing homes. Using an unmodified probability sample (Box 10.1) would produce a list of people to be interviewed scattered across the UK. The scatter might be no problem if postal questionnaires, or telephone interviewing were to be used, but if the research required face-to-face interviewing the scatter of respondents would make the research very costly. Similarly, the time and effort required to get research approved by research ethics committees is a considerable cost against research budgets. A simple probability sample of 2,000 might require gaining permissions from several hundred ethics committees. Even without ethics committees there may be a need to negotiate access to respondents via service managers and/or clinicians. To reduce such problems *cluster samples* are sometimes used. For the nursing home example, perhaps, a selection would be made first of nursing homes and then a selection would be made of residents only within those nursing homes selected. All individuals for interview would then be in one of a few clusters selected. The selection of nursing homes might be made on a quota basis (see Box 10.1), with quotas defined by criteria such as size, statutory, voluntary or private

Box 10.3 The trade-off between cost and precision with cluster sampling

Aim: To select a nationally representative sample of 2,000 hospital nurses
First stage: Select hospitals
Second stage: Select nurses within sample hospitals
Options available:

Number of hospitals selected	Number of nurses selected within each	
5	400	lower cost, lower precision
10	200	
20	100	
40	50	
50	40	
80	25	
100	20	
200	10	
400	5	higer cost, higher precision

Source: After Arber, 1993: 89

status, speciality or generality, urban or rural catchment areas, or on a random basis – simple or stratified. However, with cluster sampling it is very difficult to ensure a sample of respondents who are representative of all such people in the population, when the opportunities for selecting individuals have already been severely limited by the selection of clusters. Box 10.3 shows the trade-off between the convenience of clustering on the one hand and the loss of 'precision' on the other. Here precision means the extent to which results from the survey can be taken as accurate for the population as a whole.

While stratification (section 3) improves the extent to which a sample can be representative of a population, clustering makes this less likely to be achieved. In terms of the table in Box 10.3, a sample of 2,000 nurses drawn from five hospitals is only doubtfully representative of all hospital nurses because five hospitals are unlikely to be representative of all hospitals, in terms of their recruitment patterns, the experiences they provide for nurses, and so on. The table in Box 10.3 might be extended upwards to suggest a sample of 2,000 nurses from one hospital. Putting aside the fact that this would have to be a very large hospital, and unrepresentative in this regard, this illustrates a different kind of trade-off. Such a study would be a *case study.* Because of the clustering of nurses in one hospital, it would be possible to use their responses to build up a detailed picture of the experience of nursing in that hospital. But it would remain a puzzle as

to how far the nursing experience in that hospital was representative of other hospitals elsewhere. Alternatively, a sample of 2,000 nurses drawn from 400 hospitals would have a better claim to represent the nursing experience nation-wide, but since the hospitals selected would be very diverse, it would be difficult to read the results as representative of the nursing experience in any one of them in particular.

In some texts a study involving interviews done in, say, five agencies might be described as a 'multi-site case study' (Yin, 1994), and in others the same design might be described as a 'survey with a two stage sampling design' – the first stage establishing clusters in the shape of five agencies, and the second consisting of interviews with people selected from within each agency.

5 Staged sampling and phased sampling

Sampling is sometimes said to be *staged*, as in 'two-stage', 'three-stage' or 'multi-stage' sampling. Box 10.3 gives an example of two-stage sampling, where the first stage is the selection of a cluster sample and the second, perhaps, a simple random sample of nurses within each hospital. Or, with a small number of hospitals, the second stage might be a random stratified sample of nurses designed to represent both genders and all grades. A staged sampling design might involve any combination of cluster sampling, stratified sampling, simple probability sampling and/or quota sampling (see Box 10.1).

Sometimes the term 'stage' is used as a synonym for 'phase'. More narrowly defined, however, a *phased* sampling design is one in which the first phase includes a search for respondents of particular kinds, and the second phase is the collection of data about them, rather than about other kinds of respondents. This strategy is often used where the respondents of interest are rare and there is no convenient way of identifying them apart from using survey technique. For example, the OPCS national survey of psychiatric morbidity (Meltzer et al., 1995) was designed to estimate the prevalence of mental illness irrespective of whether the cases were known to services or not. That objective in itself made it important to use a sample drawn from the general population, rather than to rely on health service case records. However, the survey was also interested in details about people who had symptoms of severe mental illness. Thus the first phase of the survey identified people as having mental health problems or not, and the associations between this and age, gender, socio-economic status and so on. The first phase also served to *screen* respondents in terms of their mental health and hence allowed for the selection of a sub-sample of people with more severe problems for more detailed investigation. In this survey the sub-sample were invited for a second interview, which included a clinical assessment.

Although the term 'phase' may not be used, phased designs of this kind are very common in survey work, with a little information being collected from a large number of people at one phase, and, at another phase, more information being collected from a sub-sample selected on the basis of information collected in the earlier phase. Sometimes the effect of phasing is accomplished within a single interview or questionnaire by using screening or routing questions which select some respondents to answer more or different questions from others.

6 Confidence intervals and surveys

Most of what appears in Chapter 7 (sections 1 to 9) on testing experimental results for their statistical significance applies equally to surveys and will not be repeated here. If you have not read sections 4 and 5 of Chapter 7 it would be a good idea to read them before you proceed further.

As with the results of experiments, so the results of surveys should always be regarded as estimates as to the true state of affairs, estimates that are likely to be influenced by chance factors associated with sampling – *sampling bias*. The results are often cited with confidence intervals. These indicate the frequency with which various results might have been expected had the survey been repeated again and again with different probability samples of the same size. The confidence intervals provide a quick check as to whether the sample size was adequate. However, it seems to be fairly common that where sample sizes are large and more than adequate, survey researchers do not bother to provide confidence limits; they rely on their readers knowing enough about sampling to see at a glance that a sample size is big enough to produce very narrow confidence limits; that is, very precise estimates of the frequency of some phenomenon in the population. Thus, in Chapter 8 Cohen and his colleagues are working with sample sizes that are much larger than the minimum necessary and they present their results without confidence intervals. However, they do present enough data for anyone interested to calculate them for themselves (Box 10.4).

As with experiments (Chapter 7, section 8), so with surveys, sample size will determine the *statistical power* of a piece of research. With surveys, statistical power refers to the capacity of a research design to distinguish between those differences between sub-groups shown for a sample which reflect real differences for the same sub-groups in the population, and those differences between sub-groups in the sample which simply result from the chanciness of sampling. All other things being equal, a bigger sample allows for the more confident detection of smaller differences. Wide confidence intervals indicate too small a sample. Accuracy, however, comes expensive. A sample size of 384

Box 10.4 How to calculate and interpret confidence intervals in surveys

Table 10.2 below repeats part of Table 1 in the exemplar presented in Chapter 8, which is a study of NHS consumer satisfaction surveys by Cohen and colleagues.

Table 10.2 Patient dissatisfaction rates in three population surveys[a]

Statement no.	Aspect of patient care	Lothian Health Survey (1993)	NHS Users' survey (1992 and 1994 combined)
3	Sensitivity to patients' feelings[b]	6.3%	5.2%
8	Encouraged to ask questions*	23.9%	5.6%
Sample size		2,058	2,685

[a] For example, 6.3% means that 6.3% of the Lothian sample were dissatisfied about the sensitivity with which patients' feelings were treated.
[b] No statistically significant difference between the Lothian Health survey and the NHS Users' survey.
* Statistically significant difference between the Lothian Health survey and the NHS Users' survey at $p < 0.01$.

The actual result obtained is treated as an estimate. A confidence interval expresses the likely extent to which the estimate is wrong.

The formula for calculating the 95% confidence interval for percentages is:

$$1.96 \times \sqrt{\frac{(P \times Q)}{N}}$$

Where 1.96 is the 'magic number' for the 95% confidence intervals. P is the percentage you are interested in. (Don't confuse this with p meaning probability (see Chapter 7, section 2).) Q is $100 - P$ (all the percentages in which you are not interested), and N is the size of the sample. Thus for item 3 (for Lothian: $P = 6.3$, $Q = 100 - 6.3 = 93.7$ and $N = 2058$. The calculation goes as follows:

$P \times Q = 6.3 \times 93.7 = 590.31$ (1)

$(P \times Q)/N = 590.31/2058 = 0.28684$ (2)

Square root of $(P \times Q)/N = 0.53557$ (3)

$1.96 \times$ square root of $(P \times Q)/N = 1.96 \times 0.53557$
$= 1.0497258$ rounded to 1.05 (4)

The confidence interval is 1.05

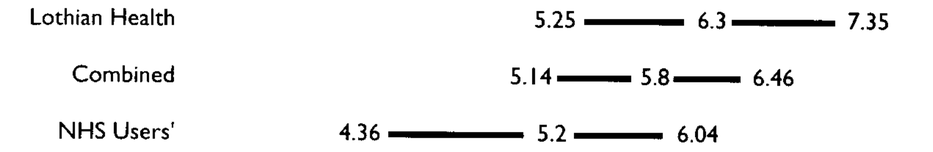
Now for the confidence limits. The actual result was 6.3% The estimate is thus 6.3% plus or minus 1.05%:

$6.3 + 1.05 = 7.35$ and
$6.3 - 1.05 = 5.25$

The confidence limits are from 5.25 to 7.35, meaning that we can be 95 per cent sure that the true percentage for the population lies somewhere between these two points.

The corresponding confidence limits for the NHS Users' survey were 5.2 plus or minus 0.84 = 4.36 to 6.04. The confidence limits for the two figures overlap (Figure 10.1).

Figure 10.1 Confidence limits for the Lothian Health survey and the NHS Users' survey separately and for combined results for Statement No. 3, illustrating a difference that is not statistically significant

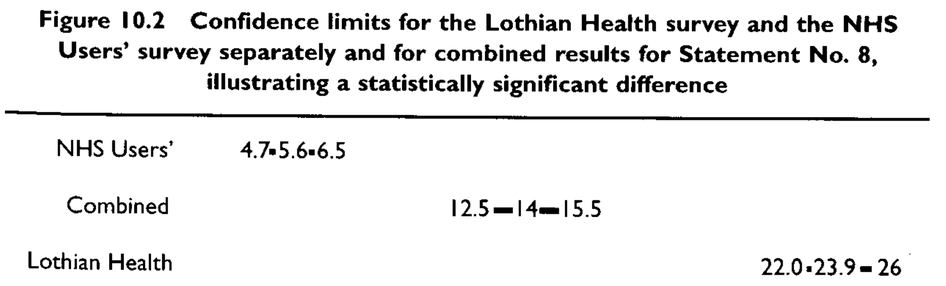


The middle line on Figure 10.1 shows what happens when the results for Lothian and the NHS Users' surveys are combined with a sample of 4,743 and a dissatisfaction rate of 5.8 per cent.

Note (b) to Table 10.2 says that there is no statistically significant difference in dissatisfaction rates for this issue between the Lothian Health and the NHS Users' survey. That conclusion came from doing a statistical test (see Chapter 7, sections 1 and 2), but the same conclusion might be drawn from looking at the way the confidence intervals overlap. The logic goes as follows. If the difference between the Lothian Health survey and the NHS Users' survey is just due to chance (not statistically significant), then the confidence intervals for both should overlap with the confidence intervals for the combined results of both surveys. In statistical texts this is often expressed in terms of the differences between the two being no greater than might be expected to occur in 95 per cent of probability samples drawn from the same population. Thus, here there might be a single population with a dissatisfaction rate of 5.8 per cent from which two probability samples were drawn, one giving a dissatisfaction rate of 5.2 per cent and one of 6.3 per cent, both of which scores are within the 95% confidence intervals for the whole population.

By contrast, the other line on Table 10.2 shows results that are highly statistically significant. Plotting the confidence limits, as in Figure 10.2, shows no overlap.

Figure 10.2 Confidence limits for the Lothian Health survey and the NHS Users' survey separately and for combined results for Statement No. 8, illustrating a statistically significant difference



In Figure 10.2 there is no overlap of the confidence limits at all. We can be 95 per cent sure that neither the NHS Users' survey result of 5.6 per cent dissatisfaction, nor the Lothian Health survey result with 23.9 per cent dissatisfaction, were random samples drawn from a population with a 14 per cent dissatisfaction rate.

would produce an estimate of the sex ratio (or any other two-value variable) in a population plus or minus 5 per cent, but it would need a sample of 1,536 to be 95 per cent sure of getting it right plus or minus 2.5 per cent (Bernard, 1994: 75–80).

Care must be taken in interpreting confidence limits when quoted for the results of a staged sampling design (section 5) using clustering (section 4). For a survey along the lines of that suggested in Box 10.3, with random samples of nurses within hospitals as a second stage, the confidence intervals would only give an estimate of how accurate the results would be *for nurses in the hospitals chosen*, and not for nurses in all hospitals. Introducing clustering divides a sample into as many samples as there are clusters and confidence interval calculations should really be done for each cluster, rather than for the grand sample. Since each cluster will be a smaller sample, the confidence intervals for each cluster will be wider, and the estimates shown will be less precise than would be the case for confidence interval calculations done at the level of the sample as a whole. This reflects the reality of the situation, but it is not uncommon to find researchers using clustering and misleadingly citing confidence intervals for the grand sample.

7 Selecting a stratified sample of adequate size: an example

The research by Payne and Saul reported in Chapter 9 used a survey to estimate the prevalence of angina symptoms in Sheffield. Their interest was in whether the need for coronary care services (indicated by angina symptoms) varied according to the affluence or poverty of the electoral ward in which people lived, and whether the availability of treatment varied likewise. Since their measurement of poverty or affluence was the poverty or affluence of wards, rather than of individuals, they needed a sample which represented the prevalence of angina accurately for each ward. There are 29 electoral wards in Sheffield. Angina is known to be both age- and gender-related. Older people and males are more likely to have symptoms. This means that for any particular ward a sample not representative of the ward for age and gender might give a misleading picture. Over-representing older males, for example, could give a higher angina figure for that ward because of the age and maleness of the sample, irrespective of whether there was a higher rate of angina in this ward compared with others for same age/same gender groups. Thus Payne and Saul needed a sample of people that was not only representative of each ward, but also representative of each age–gender group within each ward. To make matters more difficult, only a small percentage of people experience angina symptoms, so the sample had to be large enough to capture

Figure 10.3 Stratification by ward, age and gender in the Sheffield angina survey (see Chapter 9)

	Ward 1		Ward 2		Ward 29		Total (12,239)
	Male	Female	Male	Female	Male	Female	
18–34 years	Random sample	Random sample	Random sample	Random sample	Random sample	Random sample	3,738
35–54 years	Random sample	Random sample	Random sample	Random sample	Random sample	Random sample	3,837
55–94 years	Random sample	Random sample	Random sample	Random sample	Random sample	Random sample	4,664

a representative percentage of those few in each ward experiencing symptoms. Figure 10.3 shows the way their sample was stratified.

Thus the stratification divided the population into 174 categories (29 wards × 2 genders × 3 age groups) and a random sample was taken from within each category, using the health authority register as the sampling frame. When the survey was conducted these categories were to be divided again into those who had, and those who had not experienced angina symptoms. Hence the survey involves 348 unique categories of respondents. A key sample size issue here is about the size of a sample that would be adequately representative for the distribution of angina symptoms *across* the age profile *within* gender groups *within each* ward. Payne and Saul selected a sample of 16,750, and had returns of 12,240 (73 per cent). For each ward–age–gender group this gives an average sample size of 70 (each cell in Figure 10.3). Is 70 large enough to provide an accurate estimate of the prevalence of angina symptoms within these groups? Does a sample of 70 provide enough statistical power to distinguish real differences in angina prevalence between wards, from chance differences that might arise in the course of selecting the samples? An answer lies in the confidence intervals. These express the extent to which an estimate is likely to be wrong. The smaller the sample, the more likely the estimate is to be wrong.

Imagine three wards: one very affluent, one middling and one very poor and the figures for males 18–34 years old (Table 10.3).

In Table 10.3 the 'actual percentage' column shows what might have been the actual percentages derived from the survey at a sample size of 70 for each of the three ward–age groups. But as with all survey data, these are only estimates. The confidence limits show how big any error might be. Since these are the 95% confidence intervals they show where we can be 95 per cent sure what the true value will be. Thus for the 'middling' ward the 'actual figure' is 1.9 per cent, but this might

Table 10.3 95% confidence intervals at sample size of 70: angina symptoms males 18–34: three wards at different levels of deprivation[a]

	Sample size	Lower confidence limit	Actual percentage from survey	Upper confidence limit
Affluent ward	70	0*	0.95%	3.2%
Ward of middling affluence	70	0*	1.9%	5.1%
Very poor ward	70	0*	3.8%	8.3%

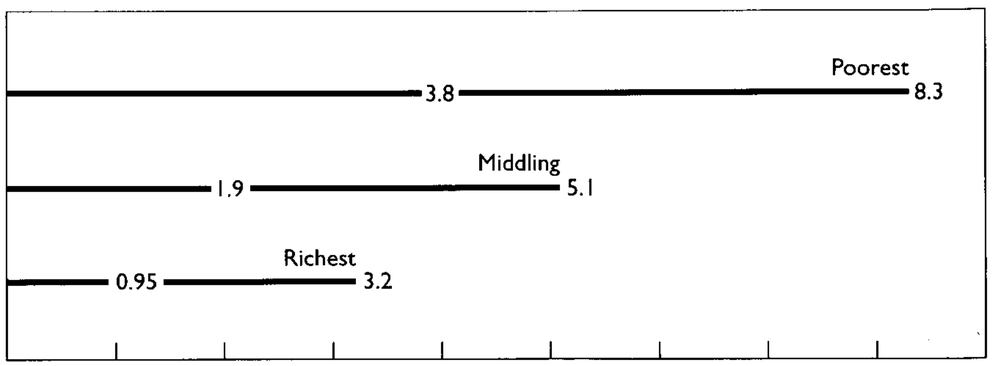
[a] Invented data based on Table 1 of Payne and Saul's study presented in Chapter 9.

* As calculated using the formula in Box 10.4, the lower confidence limits would be minus figures. But a minus percentage is impossible. In fact, zero is also impossible since in each ward *some* respondents were found who reported angina symptoms. This is an example of the way in which calculations that make good statistical sense are sometimes nonsensical in common-sense terms. It would be possible to calculate the minimum possible percentages and adjust the lower confidence intervals on that basis, but since these would be very near zero anyway there is not much point in doing so.

mean that the true value in the ward population for this age–gender group is anywhere from 0 to 5.1 per cent (though it is more likely to be closer to 1.9 per cent than to either extreme). Figure 10.4 displays the confidence intervals graphically. It shows how they overlap. So it looks possible that, despite the survey findings for the sample of different levels of angina symptoms in different wards, in fact the populations in all three wards have a similar percentage, at around 2 per cent, or any other percentage within the zone of overlap. This display illustrates a situation where we cannot see statistical significance 'at a glance' and would have to rely on statistical testing.

On this basis, had Payne and Saul only been studying three wards each with a sample of 70 for this age group of males, then their sample size was too small for comfort. But in fact they were studying 29 wards of varying affluence and deprivation and a wider age span. Hence, when all data for all poorer wards, all middling wards and all affluent wards are put together, it is highly likely that sampling errors will balance each other out. With a larger sample size thus created, the confidence intervals will be narrower and the estimates will be more precise,

Figure 10.4 Graphical display of 95% confidence intervals for data in Table 10.3



though now at the level of all males 18–94, in all poor, all affluent and all middling wards, respectively, in Sheffield. The confidence with which Payne and Saul could make claims about the prevalence of angina symptoms among males in *all poor wards together* will be much greater than the confidence with which they could make claims about the prevalence of angina symptoms among males in *any poor ward in particular*. Similarly, the confidence with which they could make claims about the prevalence of angina among all males aged 18–94 would be greater than the confidence with which they could make claims about differences in prevalence as between smaller age groups.

Payne and Saul actually cite their main results for all age groups, both genders, each ward. For this each ward has a sample size around 422. This is more than adequate for their purpose, which is to make comparisons of angina prevalence between wards. In general terms, 400 is a reasonable ball-park figure for the size of a sample needed in order to produce an accurate estimate of any dichotomous variable; that is any variable which can only take two values such as angina symptoms or no angina symptoms, the answer Yes or No, males and females. Accurate here means being 95 per cent certain that the value shown in the sample is accurate for the population with a confidence interval of 5 per cent. Thus, if the prevalence of angina symptoms found in a random sample of 400 was 4 per cent, then that would indicate a prevalence in the population of between 3.8 and 4.2 per cent (using the formula in Box 10.4) and we could be 95 per cent sure that the true value was between these limits. Note that adopting a confidence level of 95 per cent is not the same as aiming for plus or minus 5 per cent accuracy. A researcher may wish to be 95 per cent certain that an estimate is accurate plus or minus 10, 5, 3, 1 or any other per cent.

Table 10.4 shows that the ball-park figure of sample size 400 holds for samples drawn from populations of one million plus, however big they are. But with smaller populations smaller samples are possible.

Table 10.4 Size of sample required for various population sizes in order to produce an estimate of a dichotomous variable in the population, 5 per cent confidence interval

Population size	Sample size	Population size	Sample size
50	44	1,000	278
100	80	1,500	306
150	108	2,000	322
200	132	3,000	341
250	152	4,000	351
300	169	5,000	357
400	196	10,000	370
500	217	50,000	381
800	260	1,000,000+	384

Source: Krejcie and Morgan, 1970: cited in Bernard, 1994: 79

The figures given in Table 10.4 are rather 'tight' and some leeway should be allowed for non-response and unusable returns, and, if one of the values being investigated is small – as with angina symptoms – it is safer to err on the large size. One implication of Table 10.4 is that a sample size of 400 would be adequate for gaining an accurate estimate of the prevalence of angina symptoms in Sheffield as a whole with its population of 530,000 but that almost as large a sample would be needed to produce an accurate estimate for Sheffield's smallest ward, with a population of only 12,400. Moreover, while a probability sample of 400 could give a fairly accurate estimate of the prevalence of angina symptoms for Sheffield as a whole, it would not accurately show how the prevalence varied from ward to ward of the city. With a sample size this small it would not be very surprising if the people sampled from the poorer wards showed lower levels of prevalence than those sampled from the richer wards.

The important lesson from this example is that the total sample size is set by the size of the groups between which comparisons are to be made. Thus to compare the prevalence of angina symptoms between African Caribbeans in Sheffield and those of other ethnicities, a sub-sample of about 370 African Caribbeans would be needed: hence, the usual need for stratification by ethnicity when ethnic differences are of interest (see section 3).

Investigating variables which can take more than two values or attempting to achieve an accuracy greater than 5 per cent confidence intervals requires bigger samples than shown in Table 10.4. Electoral opinion polls, for example, usually use national samples of around 2,000 because there are more than two parties to vote for, and because they aim for an accuracy of plus or minus 3 per cent. The latter is because the difference in support for the leading parties is often less than 5 per cent. However, there are times when survey researchers do not need to aim for accuracy even at the plus or minus 5 per cent level. For example, in a survey designed to estimate the level of public support for a policy, the key information might be whether there was a majority for or against. In practical terms it would mean much the same if '76 per cent in favour' indicated any value between 83.6 and 68.4 per cent (plus or minus 10 per cent). Here, a smaller sample size than indicated in Table 10.4 would be adequate.

The remarks above apply to samples that are truly representative and where chance alone is likely to cause errors of estimation.

8 Under- and over-representation and their management

Having an incomplete sampling frame is one way in which samples fail to be representative (section 2). *Non-response* is the most important of the others. Among those selected to be representative of a population it will be impossible to contact some and some will refuse to co-operate. These excluded people are unlikely to be a representative sample of the sample. Subtracting an unrepresentative minority from a representative sample results in an unrepresentative sample. Cohen et al. (Chapter 8) illustrate the way in which deviations from representativeness can be checked by comparing the remaining sample with known characteristics of the population from which it is drawn. Thus, if the sample of people actually responding is on average older than the population, this is an indication that the sample of people who became respondents under-represents younger people.

The extent to which exclusion and non-response are a problem depends on a combination of four factors:

- How atypical the excludees/non-respondents are, in ways relevant to the survey.
- How great is their number.
- How large is the difference between groups in which the researcher is interested.
- How accurately the missing data can be estimated.

It is common to say that the results of surveys with non-response rates exceeding 25 per cent should be treated with suspicion. But this is slightly misleading. A larger non-response would be acceptable in a survey in which the researcher was interested in broad trends and where the non-respondents were not very atypical. By contrast, if a researcher were interested in, say, the differences between a majority population and a minority of people with disabilities, a very small non-response could invalidate the survey if disproportionate numbers of non-responders were from among the people with disabilities.

Sometimes *booster samples* are used to pre-empt or correct for initial exclusion and non-response, additional samples being taken and the results added into the main survey. The term usually implies that a different strategy for sampling is used for the booster(s) as compared with the main sample. If the strategy used for the main sample is likely to result in the under-representation of some groups then there is little point in using the same strategy in an attempt to remedy this. Adding together the results of what are, in effect, different surveys can lead to problems in statistical analysis.

Weighting is often used to manage initial exclusion and non-response. This is illustrated in Box 10.5 in terms of weighting for the under-or over-representation of age groups in a sample by comparison with the population. So long as the relevant percentages in the population are known, weighting can be done with regard to any demographic characteristic, for example, to redress the under-representation of people from a particular social class, household type and so on. For other uses of weighting, see Box 10.2.

Box 10.5 Weighting responses from a sample to manage a problem of under- or over-representation

Table 10.5 shows the percentage of males in each age group responding to a survey, compared with the percentage in the population. On each line, multiplying all the responses of the people in the sample by the 'age–sex weight' will produce a result as if there had been no under- or over-representation. These are age–*sex* weights since males make up only approximately half the totals and making these adjustments has implications for the weightings for females.

Table 10.5 Age–sex weightings to (re-)weight a sample to correct for under-representation in terms of age–sex groups

	A. Proportion in population	B. Proportion in sample	Age–sex weight To correct for under-representation multiply the sample proportion by A/B
16–19	3.8	3.4	1.12
20–24	5.9	5.0	1.18
25–29	6.6	5.6	1.18
30–34	6.0	6.0	1.00
35–39	5.3	5.3	1.00
40–44	5.3	4.9	1.08
45–49	5.2	5.2	1.00
50–54	4.2	4.4	0.96
55–59	4.0	4.3	0.93
60–64	3.8	4.0	0.95
All males 16–64	50.1%	48.1%	

Source: Based on Meltzer et al., 1995: Table A3.4. Office for National Statistics © Crown copyright 1995

There are problems in the kind of weighting shown in Table 10.5. For example, what is increased by 0.4 per cent in the first row will be the contribution of the responses of those males 16–19 who did respond to the survey. Males aged 16–19 will no longer be under-represented, but the responses of the kinds of males aged 16–19 who were originally under-represented may be even more under-represented now.

Weighting in this way also assumes that researchers have an accurate knowledge of the composition of the population against which to compare the sample. For survey work in general populations, the census is usually the most accurate source of such information, but census data become progressively more inaccurate in between census dates. In the study featured in Chapter 8, Cohen and his colleagues did not have up-to-date census data on the age composition of the Lothian region, and had instead to use estimates derived from various other surveys in order both to judge the age-representativeness of the Lothian sample and to re-weight the results in order to undo the effects of stratification in the original sampling design (section 3).

9 Case control (or case comparison) studies

A survey with a sample of adequate size drawn from the general population is often the only way of producing an accurate estimate of the frequency of a condition or circumstance in a population. If the condition is rare, a very large sample is required for this. For example, in 1998 it would have required a sample of millions accurately to estimate the frequency of new strain Creutzfeldt–Jakob disease (n-sCJD) since there had only been 35 known cases in the UK in the previous 18 years (National CJD Surveillance Unit, 1999). For many purposes it is enough to know that a condition is rare, without being able to put an exact figure on the frequency. However, when it comes to investigating the causes of a rare condition a largish sample of cases is needed. One way of managing this problem would be to use a stratified sample (section 3), taking a largish sample of cases, to compare with a random sample of 'non-cases'. However, with rare conditions, or conditions that are difficult to know about such as drug abuse or child abuse, there is no adequate sampling frame (section 2) to provide the starting point for the random sampling of cases. *Case control* (or *case comparison*) studies provide an alternative approach.

Here known cases are recruited to the survey. Usually these are cases known to services, for example, patients diagnosed with sporadic CJD (which includes n-sCJD) or children on an at-risk register. Then a sample of 'controls' is recruited to match the cases according to variables such as age, gender, socio-economic status, ethnicity and so on. Ideally matching should be for all characteristics that might be relevant, except the one for which causes are being investigated. But in practice matching has to be on characteristics that are easy to know about before detailed investigation begins. Matching may be done at the individual level using a *matched pairs* design, or it may be done group on group so that, although there are no individual matches between cases and controls, the two groups have similar profiles for age, gender, socio-economic status and so on (see Box 5.1 in Chapter 5). For example, a series of case control studies associated with the Confidential Inquiry into Stillbirths and Deaths in Infancy uses all known cases of sudden infant death in three NHS regions (1993–5) – 195 deaths – and 780 matched controls (Blair et al., 1996; Fleming et al., 1996). The controls in this case were selected from health visiting records. The logic of these studies is experimental (see Chapter 5). Those variables which are found equally associated with both the sudden infant death babies and with the living controls are unlikely to be among the causes of sudden infant death, and those variables

which are associated more with the deaths than with the controls are possibly among the causative factors.

There are two problem areas for case control studies. The first concerns representativeness. Cases known to services are not necessarily representative of all cases. There is a risk that the cases recruited are a sample biased by *ascertainment bias*, meaning that those which are known about are different in significant ways from those which remain unknown. For example, offenders who have been arrested are unlikely to be a representative sample of all offenders. Similarly, problems can arise from the selection of the controls. For example, the earlier case control studies of n-sCJD used controls drawn from hospital populations. But in 1997 this was abandoned in favour of controls drawn from the general population because 'controls chosen from hospital patients may have medical histories which are not representative of the general population' (National CJD Surveillance Unit, 1999: 19). But the Unit point out that one of the costs of using controls drawn from the general population here is that fewer of them will have detailed medical histories, and more data will have to be collected directly by the researchers.

For these reasons, and simply because cases and, nearly always, controls will not have been chosen on probability principles, frequencies derived from case control studies cannot be generalised safely to wider populations.

The second problem area concerns the matching of cases and controls. The importance in experiments of creating matching comparison groups is discussed in detail in Chapter 5, section 3. Just as with experiments, faulty matching can lead to confounding and misleading results. For example, in the field of sudden infant death research, family smoking habits have been shown to be associated with sudden infant death: the more smoking, the greater the risk. However, some doubts have been expressed as to the adequacy of the matching of cases and controls for social class (Dwyer and Ponsonby, 1996). If the matching was imperfect such that the controls contained a smaller percentage of people from lower socio-economic groups than the cases, then some at least of the association found between smoking and sudden infant death may reflect the fact that people in lower socio-economic groups both smoke more, and that, for reasons not associated with smoking, their babies are more likely to experience sudden infant death.

10 Correlation, co-efficients, regression and scatterplots

In analysing the results of surveys or case control studies differences between groups are often of interest: for example, different percentages

of different kinds of people expressing dissatisfaction with the NHS (Chapter 8), different rates of angina symptoms in different wards, or between different age groups (Chapter 9), or differences expressed in terms of odds ratios (Chapter 7, section 10.4). Such differences can be tested for their statistical significance in exactly the same ways as differences between groups in experimental research (see Chapter 7, sections 1 and 2) and all the remarks about kinds of data, and appropriate statistical tests in Chapters 6 and 7 apply equally to analysing surveys.

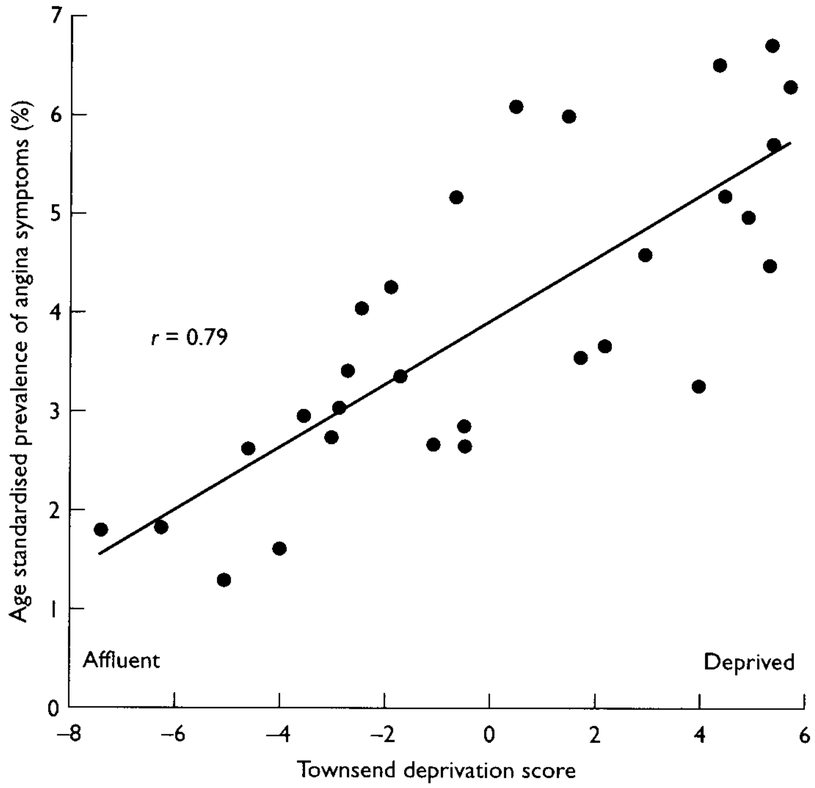
In survey work, however, correlations – measures of association – are just as important as differences. Correlation is the extent to which one thing varies with another. It is often called 'co-variance'. A good example of a correlation is given by the two ends of a see-saw. Since one end always goes up when the other end goes down there is a perfect *negative* (or *inverse*) correlation between the positions of the two ends. The movement of any two points on a roundabout show a perfect *positive* correlation. For every metre one point moves, the other will always make a corresponding movement; not necessarily a movement of a metre but always the same distance for every metre moved by the other point.

Correlations are usually expressed in terms of correlation co-efficients. Usually, but not always, these express a perfect positive correlation as 1, and a perfect negative (or inverse) correlation as -1. Zero designates no correlation at all. Seesaws and roundabouts apart, perfect correlations are rare, and 0.8 is usually a high positive correlation and -0.8 a high negative one.

Chapter 6 (sections 6 and 7) deals with the validation of research instruments which usually entails statistical correlation. Many such instruments are questionnaires and the questionnaires used in surveys are often validated in the same way for much the same criteria.

Another research tool used in surveys is the deprivation index, exemplified in this volume by the study by Payne and Saul (Chapter 9 exemplar), and described in more detail in Chapter 11, section 4. Deprivation indicators or indices are widely used in both research and social administration. They use easily obtainable population data of the kind which most people would agree indicates socio-economic deprivation; for example, percentage of people unemployed, percentage of single parent families, number of households without cars, and so on. Then from such indicators a score is produced for an area which is a good basis for predicting matters that are much more costly to find out about: for example, numbers of people suffering from depression, numbers of children who will suffer from glue-ear next year, and so on. Validation of a deprivation indicator means doing research to find out how well the unknown (and difficult to find out about) can be predicted from the known (and easy to find out about). The best deprivation indicators are the ones which show the highest

Figure 10.5 Prevalence of angina symptoms compared with Townsend index of deprivation (Payne and Saul, 1997: Fig. 1: 258)



correlations between the deprivation score, and whatever else researchers were trying to predict.

Correlations are often displayed in terms of *scatterplots* or *scattergrams*. Figure 10.5 comes from the exemplar study by Payne and Saul in Chapter 9. Along the bottom of Figure 10.5 is a scale expressing the affluence or deprivation of areas according to the Townsend deprivation index. Zero is the middle score. Positive scores are an indication that an area is more deprived than average: negative more affluent than average. On the vertical axis there is a scale for measuring the proportion of people in an area experiencing angina symptoms. Each plot on the diagram represents a ward of Sheffield located according to its deprivation index score (horizontal axis) and the percentage of people in the ward experiencing symptoms (vertical axis). These latter are expressed after *age standardisation*. That is, after accounting for the fact that different wards have different age profiles. Age standardisation is dealt with in Chapter 11.

From eyeballing the distribution it is possible to see that the more deprived the area, the greater the percentage of people experiencing angina symptoms: or, to be precise, the greater the percentage of

people in a sample of people from each area reporting that they experienced such symptoms in a survey conducted for that purpose.

The diagonal line drawn on the diagram is the *regression line*. If there were a perfect positive correlation between ward deprivation and reported angina symptoms all the plots on the diagram would lie on this line. It would be very surprising if they did, and the regression line represents the best estimate that can be made with these data about the relationship between angina and deprivation. One way of thinking about this is to think of how far this display improves our ability to predict the prevalence of angina symptoms from knowing the deprivation score (or vice versa if you prefer). If we didn't know the deprivation index then *for any area* the best bet is that it will have the same prevalence of angina as for *all areas*. This is about 4 per cent (with 95 per cent confidence limits of 3.7 to 4.4 per cent). This will under-estimate the rate for a highly deprived area, and over-estimate it for an affluent one. The regression line improves our ability to predict. Thus, with a Townsend score of 6, the regression line predicts an angina prevalence of about 5.5 per cent. Reading from Figure 10.5, the actual scores for wards with this, or a close degree of deprivation, are (approximately) 5.6, 6.1 and 6.5. The prediction is not perfect but it is nearer the mark than a prediction that these areas will have the average score for angina symptoms (4 per cent) and it improves the prediction in the right direction.

On scatterplots like this, plots clustering around a diagonal line sloping in the opposite direction would show negative (or inverse) correlation (see Figures 3 and 4 in Payne and Saul's study in Chapter 9), and plots showing no pattern at all would be showing no, or only a very weak correlation (see Figure 11.3 in Chapter 11).

Figure 10.5 also bears the legend '$r = 0.79$'. The r in this case is Pearson's product moment correlation co-efficient. As with other correlation co-efficients it expresses, in effect, the extent to which the observations cluster round the regression line. If they were all exactly on the regression line r would equal 1 (or minus 1). If there were no pattern at all r would equal zero. Since correlations can occur by chance, correlation co-efficients are also tested for their significance, although a score of 0.79 can usually be regarded as a 'high' correlation. In this case p was less than 0.001: there was less than one chance in 1,000 that this pattern of association was the result of chance factors (see Table 7.2 in Chapter 7 for the meaning of different values of p).

11 Correlations, causes and statistical control

Surveys – including case control studies (section 9) – can demonstrate correlations: what is associated with what. (For the expression of

correlations, see section 10.) Causal links are often inferred from correlations. For example, being poor and dying earlier are correlated. One reasonable interpretation of this is that there is something about being poor which *causes* premature death. However, there is usually more than one interpretation of a correlation. For example, it is possible that the kinds of people who die earlier are also the kinds of people who are in poor health, and that it is being in bad health which both *causes* them to be poor and causes them to die earlier. In this case it is likely that both interpretations are valid; that there are two *directions of effect*. One may be more important than the other but a survey alone may not be able to establish the predominant direction of cause.

Payne and Saul in the Chapter 9 exemplar found that those with angina in the poorest wards of Sheffield were only about one-third as likely as those in the most affluent wards to receive angiography: there is a correlation between higher deprivation and lower rates of angiography. But what underlies this correlation? Is some aspect of poverty causing poorer people to receive angiography less often, or some aspect of affluence causing better off people to receive angiography more often? And, if either, or both, what exactly is, or are, the causal mechanism(s), and do they operate at the level of individual characteristics, or at the level of difference between services serving different parts of the city. These are not questions for which Payne and Saul's research design is capable of giving a final answer. This is partly because they did not collect the data to do so, but more importantly because surveys are not well designed to demonstrate causality. As noted in Chapter 5, only an experimental approach can do this, through setting up artificial situations in which variables are controlled, excluding all influences on what happens, except that which is of particular interest.

Survey researchers can, however, exert *statistical control* over variables. Age-standardising data provide a simple example. Two wards might show different levels of angina either because one ward has an older population than the other, or because it is more deprived than the other. To exclude the effect of age on differences in angina rates between wards all that is necessary is to compare the rates of angina between the wards, age group by age group, and, to exclude the effect of gender, to express the results for males separately from females. Age and gender will then have been *controlled* and any remaining differences in rates of angina between the two wards are available to be attributed to differences in deprivation. More elaborate techniques of age-standardisation are discussed in Chapter 11. This example shows how variables can be controlled statistically in analysing survey results, but it also shows a major problem with this. While the manoeuvre above will have eliminated the effects of age and gender on differences in angina rates between wards, any differences left will not

solely be due to deprivation, but to deprivation plus other uncontrolled factors, or worse, due to other uncontrolled factors rather than to deprivation and affluence.

In discussing the reasons why people in more deprived wards are less likely to be treated with angiography Payne and Saul consider an alternative to a straightforward 'poverty' explanation: poorer people are more likely to smoke, and there is anecdotal evidence to suggest that doctors are less likely to allocate expensive coronary care procedures to smokers.

Payne and Saul exert statistical control over their data in order to see how well this explanation stands up.

> The angiography rate in those with angina identified through the survey was found to be 11% (13/116) in the 10 most affluent wards and 4% (9/216) in the 10 most deprived wards. National and local data suggest that about 83% of the affluent populations are likely to be non-smokers, but only 65% of deprived populations. Even if the smokers had been excluded from treatment (that is from the numerator) and if the denominator was adjusted to reflect the likely number of non-smokers, the angiography rate in affluent wards would still be twice that in deprived wards – that is 13% (13/(116 × 0.83)) *v* 6% (9/(216 × 0.65)), respectively. (Payne and Saul, 1997: 261; see Chapter 9)

As Payne and Saul note, it would have been better if their survey had collected data about whether people smoked or not, rather than relying on estimates. In the absence of evidence from their own sample they have to assume that *all* those receiving angiography were non-smokers. But that has the advantage of modelling what the situation would be if discrimination against smokers had the strongest possible effect. They also assume that the rate of smoking among people with angina symptoms would be the same as for those in populations of smokers and non-smokers of similar socio-economic status. This is unlikely, since coronary heart disease is itself correlated with smoking. But this will not alter the comparison being made so long as the correlation between smoking and coronary heart disease is of much the same strength irrespective of social class. If it is, any over- or under-estimate has the same effect on both groups being compared. Despite its speculative nature the analysis quoted above does illustrate the way in which a 'discrimination-against-smokers' explanation can be tested by controlling for smoking, that is, by comparing angiography rates just between *non-smokers* with angina symptoms from affluent areas and *non-smokers* with angina symptoms from deprived areas.

Notice that statistical control presents a version of a problem which may arise wherever comparisons are made in the search for important differences between groups. In experiments (Chapter 5) it is important to create comparison groups as similar to each other as possible

prior to the intervention – otherwise any outcome differences may reflect prior differences between groups, rather than the effects of the different ways in which the groups were treated. In case control studies (section 9), imperfect matching of cases to controls may result in mismatches between the groups being mistaken for 'risk factors'. Similarly, sorting the data to exert statistical control may create groups which do not match for variables that should be controlled. For example, we might try to control for age by citing results in terms of age bands. But this still leaves it possible that in one 18–25 age group most of the subjects are between 18 and 20, and in another most are between 20 and 25. A difference shown between the groups may be due to this imperfect control of age.

Sample size limits the possibilities for exerting statistical control over survey results. This is because statistical control requires the respondents to be divided into sub-groups which will then be compared with each other. Payne and Saul started with a large sample, but for the analysis quoted above, they have deleted the results from nine wards of middling affluence, and from all the people without angina symptoms, divided the remainder into the richest 10 and the poorest 10 wards, then again into those receiving and those not receiving angiography, and then again into smokers and non-smokers. The result is that the differences between sub-groups they are comparing are very small (13/116 and 9/216). It would be impossible for them to take a further step and, for example, investigate gender differences in angiography rates between non-smokers according to ward affluence or deprivation. By this time the sample has run out of statistical power. Many surveys have smaller samples than this, and therefore much less capacity for testing the possible reasons for correlations.

Correlations from surveys can be useful even if the causal sequences which they reflect are unclear. For example, knowing that there is a correlation between socio-economic status and coronary disease is useful for planning services even if it is unclear as to the mechanism which links higher rates of coronary disease to higher rates of poverty. In many areas of practice the 'risk factors' identified result from correlations derived from surveys.

12 Contemporaneous and longitudinal surveys

Most surveys, including most case control studies, are *contemporaneous*: snap-shots of a state of affairs at a particular point in time, and therefore gather only *retrospective* data about events which happened in the past. Data that rely on people's memories must be regarded with less confidence than data about current matters. *Retrospective*, or *recall bias*, is not only a question of forgetting; it is a characteristic of

human memory that people constantly revise their memories in the light of what has happened subsequently and in terms of the context in which they are asked the questions. For example, in the case control studies of sudden infant death (section 9) it is possible that since some people questioned have already heard that there are associations between sleeping posture and sudden infant death that they mis-remember what happened in a way that shows that they did what was right (Dwyer and Ponsonby, 1996). Again, mothers whose children have become delinquent may remember events in a child's past differently from those mothers whose children have not become delinquent, since they will have reconstructed their memories in a search for an explanation (West, 1969). This is to say nothing about the deliberate falsification of answers. Because data about what happened before and what happened afterwards all have to be collected at a single point in time, contemporaneous surveys are particularly weak designs for investigating causality and issues of effectiveness. Accurate service records can, of course, compensate for recall bias. Payne and Saul (Chapter 9), for example, check their respondents' answers about treatments received against medical records.

Longitudinal surveys (or prospective surveys) are able to produce more convincing evidence about what causes what by collecting data from people *before* the events of interest happen, and then following them up with one or more further studies later. Surveying at different points in time ameliorates *direction of effect* problems. A longitudinal survey, for example, can establish whether those who died prematurely were poor before they became ill, or got poor because of their illness (Goldblatt, 1990). Similarly, an evaluation study may take a longitudinal survey design, with a survey before the implementation of a policy, and a survey some time afterwards (Tudor Smith et al., 1998). Most longitudinal surveys are in fact a time series of snapshot surveys, perhaps just two, or perhaps more where the same *panel* of respondents is questioned on numerous occasions.

The terms *longitudinal surveys, prospective surveys* and *cohort studies* are often used as synonyms for each other. Usually all these terms imply that the same subjects are surveyed at different points in time, with the term *repeat survey* being used to refer to the same survey being conducted at different points in time with different subjects, as with the NHS Users' survey in Chapter 8. However, usage is imprecise in this field. 'Cohort study' is sometimes used more narrowly to refer to medical research where only a small range of factors are of interest and where a convenience sample rather than a probability sample is the starting point. A classic example is that by Doll and Hill begun in 1951 to investigate the link between smoking, lung cancer and coronary heart disease. This is illustrated in Box 10.6, which also explains the notions of *relative risk* and *attributable risk*.

Box 10.6 An example of a medical cohort study: relative risk and attributable risk

In 1951 Doll and Hill (1964) sent a questionnaire to all 59,600 doctors on the UK Medical Register asking about their smoking habits. Sixty-eight per cent returned questionnaires. From 1951 to 1961 the deaths of doctors and the causes of their death were monitored, mainly through the death registration process. There were 4,963 deaths. Some of the results are shown in Table 10.6.

Table 10.6 Deaths of doctors by smoking behaviour 1951–1961: some results from a cohort study

	Deaths per 1,000 persons (doctors) per year			
Cause of death	All doctors in survey	Non-smoking doctors	All cigarette smoking doctors	Doctors smoking more than 25 cigarettes a day
All causes	14.05	12.06	16.32	19.67
Lung cancer	0.65	0.07	1.20	2.23
Coronary heart disease	3.99	3.31	4.57	4.97

Source: Based on Unwin et al., 1997: 38

Two ways in which data like this are often expressed are in terms of *relative risk* and *attributable risk*.

Relative risk

$$\frac{\text{Rate in the group with the attribute or exposure}}{\text{Rate in the group without the attribute or exposure}}$$

For example: 1.20/1,000 doctors who smoked died per year of lung cancer in the 10-year period, and only 0.07/1,000 non-smoking doctors died of lung cancer

$$\frac{1.20}{0.07} = 17.1$$

meaning that those smoking were 17 times more likely to die of lung cancer than non-smokers.

Attributable risk There are two versions of this: attributable risk (exposed) and attributable risk (population).

Attributable risk (exposed) is the rate among those with the attribute or exposure *minus* the rate among those without the attribute or not exposed. For example: 1.20/1,000 doctors who smoked died per year of lung cancer in the 10 year period, and only 0.07/1,000 non-smoking doctors died of lung cancer.

$$1.20 - 0.07 = 1.13 \text{ per 1,000 persons per year}$$

meaning that out of the 1.20 deaths from lung cancer per 1,000, 1.13 can be attributed to smoking. This can be expressed as a proportion:

$$\frac{1.20 - 0.07}{1.20} \times 100 = 94 \text{ per cent}$$

94 per cent of deaths from lung cancer were due to smoking.

Attributable risk (population) is the rate in the population *minus* the rate in the group with the attribute or exposure. For example, deaths per 1,000 from lung cancer in the population were 0.65/1,000 and for smokers were 1.20/1,000. In this case, 'the population' is the sample of doctors responding to the questionnaire.

$$0.65 - 0.07 = 0.58 \text{ per 1,000 per year}$$

meaning that 0.58 deaths per year from lung cancer in this population were due to smoking. Expressed as a proportion:

$$\frac{0.58}{0.65} \times 100 = 89 \text{ per cent}$$

89 per cent of deaths from lung cancer in this population were due to smoking.

The same results might also be expressed in terms of other expressions of effect size (see chapter 7, sections 10.1–10.4).

Definitions and calculations based on Unwin et al., 1997: 38–40

As the material in Box 10.6 indicates, longitudinal studies sometimes start with convenience samples, particularly those called cohort studies. In Doll and Hill's study this was all UK doctors, and in the so-called Whitehall study of mortality and morbidity (Marmot, 1995) the sample was taken from civil service employees. For longitudinal work there are some merits in starting with a sample of people who are likely to be easy to stay in touch with, rather than using probability sampling. This has the same effect as 'clustering' respondents (section 4). As with clustering *per se*, the cost of this convenience lies in the problems of generalising from a sample to a wider population. To generalise from the Doll and Hill study (Box 10.6) we have to make 68 per cent of all doctors stand as representative of all adults with regard to the influence of smoking on death rates. It is reasonable to assume that the research demonstrates an elevation of the risk of death for smokers everywhere – and other research has shown this. But it is not reasonable to assume that the extent to which the risk of death is elevated among doctors is the same as for other occupational groups. Thus it would not be safe to generalise the relative risk and attributable risk figures calculated in Box 10.6 from the sample of doctors

to the population of all adults in Great Britain. The results of the Whitehall study (Marmot, 1995) show a continuous gradation of ill-health and premature death from the highest to the lower grades, but although the civil service covers a wide range of pay rates and work circumstances, it cannot be entirely representative of the social class spectrum of Great Britain.

Perhaps the best-known longitudinal surveys are the national birth cohort studies, which are so important in the knowledge base for practitioners in child health and child social work and education (Wadsworth, 1991,1996). These are studies through life of all or a large sample of the children born in a particular week in 1946 (5,362 children), 1958 (17,000 children) or 1970 (16,000 children). For the 1946 cohort the sample was fixed, but for the other two samples immigrant children with the same birth date have been identified and added in – an example of using booster samples (see section 8) However, in all the cohorts children from ethnic minority backgrounds remain under-represented in terms of today's population. The 1946 birth cohort study has been extended to cover the children of the original children (Wadsworth, 1996).

The sampling strategy for these studies might be regarded as a kind of systematic sampling (see Box 10.1), though with a single sampling interval, or as a form of clustering, by time rather than by place (see section 4). Systematic samples are as good as random samples so long as there is an arbitrary relationship between the sampling interval and the data collected. But there are important differences between children with summer birthdays and others, and hence no single week of birth will produce a sample of children representative of all children born that year for characteristics which depend on season of birth. This is a limitation of these studies, though for many purposes an unimportant one.

The longer a study goes on, the greater the opportunities to lose contact with respondents. However, in longitudinal surveys the problem of non-response is reduced somewhat by the fact that much data will have been collected about people before they disappear so that similarities and differences between those who get lost and those who don't can be specified, and the results interpreted accordingly, unlike the situation in a snapshot survey where it may be difficult to know about the characteristics of those people sampled, but not entering the survey. In fact these birth cohort studies have been very successful in retaining subjects. For example, in 1985 the birth cohort of 1958 was still scoring a 76 per cent response rate, and the loss includes people who died as well as those who lost contact for other reasons (Shepherd, 1985).

Longitudinal surveys conducted over a long time period also exaggerate another problem of generalisation; that of *generalising through time*. In March 2000 the children who entered the national birth cohort study in 1946 were 54 years old. The study does and will continue to provide an enormous amount of information on how circumstances in infancy and childhood in the 1940s relate to someone's life in their 50s at the turn of a new century. But how far will the findings also be true for someone born in the 1960s, or 1970s, or 1980s and in their 50s in 2010, 2020 or 2030 ? The circumstances of childhood in the 1940s are long gone. There can, of course, be time generalisation problems with any kinds of research as the research ages.

13 The ecological fallacy in interpreting survey results

To commit an ecological fallacy is to make unjustifiable assumptions about the behaviour or conditions of individuals on the basis of the characteristics of the areas they come from. Robinson (1950), who invented the term, gave as an example the possibly fallacious argument that unemployment causes high rates of crime, based on a survey finding that areas with high crime rates also have high rates of unemployment. Unless a survey has been specially designed to establish that it is unemployed people who commit the crimes, this would not be a safe interpretation of the survey results. For most of their study of the relationship between the need for coronary care services and the receipt of coronary care services, Payne and Saul (Chapter 9) take the electoral ward as their unit of analysis. They find both that people in poorer wards report more angina symptoms and that people in poorer wards are less likely to receive more advanced forms of coronary care. But since they did not collect information about the economic conditions of individuals their data *does not* actually show that it is poorer people in Sheffield who are most at risk of coronary heart disease and least likely to be treated for it. With their data it would remain possible – though very unlikely – that it is the richest people in the poorest wards who are at greatest risk, and the poorest people in the richest wards who receive most care. While their assumptions seem reasonable enough, there is a gap between ward-level data which they have collected and individual socio-economic conditions, about which they have no data and the gap has to be bridged by making assumptions. Since Payne and Saul were particularly interested in providing intelligence for planning services, and services have to be provided on a territorial basis, analysis at the level of the ward, rather than at the level of the individual seems justified. They address the problem of ecological validity in their article.

In terms of applying research results to practice, the ecological fallacy can sometimes lead to poor policy making. For example, in many rural areas poor people are in the minority and their deprivation gets lost in expressions of the socio-economic status of areas

(Abbott et al., 1992). Or again, the educational priority area programme of the 1960s was designed to improve educational achievement. Additional resources were targeted to schools and educational welfare services in areas with high levels of socio-economic disadvantage, since these were the *areas* with the lowest educational results. But the majority of under-achieving children, and the majority of poor children, actually did not, and do not, live in areas of high socio-economic disadvantage, but spread across all areas of the country. Nor is the category 'under-achieving children' quite identical with the category 'deprived children' (Bernstein, 1970).

14 Questionnaires, reliability and meaningfulness

The main instrument used in surveys is the questionnaire. Usually the questions are posed to be answered by the people who are selected for the survey. Sometimes, however, the questions are actually answered by a practitioner or practitioner-researcher, who carries out a clinical examination or assessment on the people selected for the survey. This can lead to problems if different practitioners make judgements in different ways (see Chapter 6, section 5).

However questions are posed, the way they are posed will shape the answers given. In Chapter 8 Cohen and his colleagues illustrate this with regard to differences in response to what look like similar questions.

Most of the issues raised about research instruments in general in Chapter 6, apply equally to questionnaires, but in Part 4 of this book there is a set of critical appraisal questions to ask about surveys which includes questions to ask about questionnaires.

The results of a survey are a composite of comparisons and contrasts. For this reason it is important that data about the same matters are collected in the same way from each respondent so that like can be added to like and contrasted with unlike. Put another way, survey researchers usually place a great emphasis on *reliability*. This follows from a long history of survey work which has developed a substantial research base showing that unless safeguards are used there is a strong danger that responses will reflect more about interviewers than about interviewees, or more about questionnaires than about those who fill them in (for example, Bradburn, 1983). *Social desirability bias* is a particularly important source of unreliability (Fielding, 1993: 147–50). While most respondents will probably want to show themselves in a good light, what responses they think will do this will depend on what cues they can pick up from the wording of the questions or the demeanour, persona or social status of the interviewer, or from the location of the interviews (Davies, 1997: Chapter 4). Interviews and interviewers may provide different cues as

to social desirability to different types of respondents, and different respondents may read the same cues differently. By contrast with questionnaire studies the possibility for social desirability bias is much greater in loosely structured ('in-depth') interviews, where interviewers disclose far more of themselves and hence give more cues as to what would be a socially desirable performance by the interviewee, and where interviewees have much more opportunity to probe the views of their interviewer (see Chapter 16).

Unreliability undermines the credibility of comparisons made in analysing survey results. For example, different levels of dissatisfaction shown by the clientele of different agencies might turn out to be merely the result of their being questioned in different ways (see Chapter 8, for example). Apparent changes over time might be the result of a second survey asking questions in a different way from the first. Table 10.7 gives a synopsis of the main problems arising from unreliability of method in survey research and the main safeguards used to avoid them.

Table 10.7 also indicates one of the major trade-offs in research; here between reliability and meaningfulness. Following the policy suggested in the table to maximise reliability may mean, for example:

- asking people questions that may not be important or meaningful to them;
- asking some people questions that are important and meaningful to them, when the same questions will be less meaningful or important to others;
- not asking people, or some people, questions relevant to the survey topic which may be much more meaningful and important to them than the questions actually asked;
- forcing people to opt among responses which are pre-decided, when they might prefer to give a response which has not been provided for;
- offering some people the opportunity to give exactly the response they would like to give, but ruling this out for others;
- giving a performance as an interviewer which is congenial to some kinds of people, but not to others.

Qualitative research is often offered as an antidote to the problems listed above, though solving them usually means compromising reliability, representativeness and generalisability. Thus what a survey researcher might regard as poor practice because it is unreliable, a qualitative researcher might regard as good practice because it adjusts the collection of data to the particularities of each individual respondent, and therefore produces data which are meaningful to those who provide them (Oakley, 1981). Chapter 12 in this volume is by Mildred Blaxter, one of the main researchers involved in the large-scale *Health and Lifestyle Survey*. Here she contrasts the kind of data which

Table 10.7 Problems of reliability in surveys and some safeguards against them

Shortcomings in design	Problems of interpretation	Safeguards and remedies
Different questions might be asked of different respondents	If entirely different questions are asked of each respondent this will be many surveys each with a sample of one, and no sound generalisations will be possible. More usually, the result will be to reduce a larger, possibly representative sample to a number of smaller (probably unrepresentative) samples each consisting of a cluster of respondents asked the same or similar questions about the same topic(s)	Ask each respondent the same questions in the same ways. Brief each in the same way about the purpose of the survey
The same questions might be asked but in different ways of different respondents *Or* The interviewer might make different kinds of relationships with different respondents	Unless the differences are known precisely it will be impossible to decide whether differences in responses between respondents are due to differences between them, or to the different ways in which they were asked questions and/or the different relationships they struck up with the interviewer	Prepare each interviewer to follow a standard protocol for interviewing. If several interviewers are employed, analyse the results interviewer by interviewer to detect *interviewer effects*. Even if only one interviewer is used, analyse the results of, say, early as opposed to later interviews, interviews with males and interviews with females and so on, to look for interviewer effects
Interviewers with very different characteristics are employed	There will always be some *interviewer effect*, with results differing according to different interviewers – or even perhaps between interviews done at different times by the same interviewer	
Questions are open-ended and allow respondents to determine what are appropriate answers. There may be a good case for using open-ended questions but there will be problems of interpretation none the less	There will be acute difficulties of matching the answers of one respondent with those of another. Insofar as some people will give longer, fuller answers the results will over-represent the loquacious and under-represent the reticent	Preferably use closed/forced choice questions. If using open-ended questions do not assume representativeness for the distribution of answers
In epidemiological research, different diagnosticians might be using different decision-rules for making judgements	It will be difficult to decide how far the distribution of an illness or social problem shown in the survey reflects something real and how far it reflects different ways of defining or diagnosing problems	Use standardised protocols for assessment. Establish similarity or difference of judgement by subjecting at least a sample (of the sample) of judgements to inter-rater reliability testing (see Chapter 6, section 6)
Questions asked about past events are vulnerable to *recall bias*	It will not be clear whether differences between respondents are due to differences in the way in which they reconstruct their memories	Avoid asking questions about distant events. Find means to verify factual accounts

are produced by a questionnaire survey with the kind of data which are produced by informal, unstructured interviewing, which give an insight into how the people concerned themselves make sense of the topics about which they are asked questions.

15 Questions to ask about surveys

There is a checklist of critical appraisal questions about surveys in Part 4 of this volume. It deals mainly with questions to ask about whether a survey is valid in its own terms for the population the survey claims to represent. This may well be a population different from the practice population of the practitioner reader. It may be larger, in another place, or at another time. There are also questions to ask about how to extrapolate from a survey to a practice population. Two common purposes for extrapolation are:

- *Performance bench-marking*. For example, the NHS consumer satisfaction surveys reviewed by Cohen and his colleagues in Chapter 8 might be used to set norms for local performance. NHS Trusts might be surveyed separately to judge whether they were generating more or less satisfaction than the average for Scotland as a whole. Something like this is the purpose of the series of annual consumer surveys instituted by central government in 1998 (Department of Health, 1997). Insofar as 'performance' is only a small part of the cause of satisfaction ratings, Trusts would want to know how far the populations of their catchment areas were similar to or different from those of the national population surveyed, and especially in terms of those factors likely to influence satisfaction rates. For example, it is to be expected that the younger the population of a hospital catchment area, the lower the level of satisfaction which will be recorded in a satisfaction survey.
- *Epidemiological needs assessment*. Much survey work identifies the social distribution of illness, disability and social problems. This is useful information if it can be transferred from the population surveyed to another population about which service planning decisions are to be made.

Chapter 11 gives some details about extrapolating from an epidemiological survey.

16 Further reading on surveys and case control studies

On surveys

There is no shortage of good texts on survey methodology. A good introduction is Sapsford (1999). For survey methods in health care

research in particular, see Bowling (1997, Chapters 12, 13 and 14), and in social work, see Alston and Bowles (1998, Chapters 5 and 6). For more technical information about sampling and questionnaire design, see Alreck and Settle (1995) or De Vauss (1995).

On case control studies

An excellent introduction to both case control and cohort studies is given by Mant and Jenkinson (1997).

References and further reading

Abbott, P., Bernie, J., Payne, G. and Sapsford, R. (1992) 'Health and material deprivation in Plymouth', in P. Abbott and R. Sapsford (eds), *Research into Practice: a Reader for Nurses and the Caring Professions*. Buckingham: Open University Press. pp. 129–55.

Alreck, P. and Settle, R. (1995) *The Survey Research Handbook*, 2nd edn. Burr Ridge: Irwin.

Alston, M. and Bowles, W. (1998) *Research for Social Workers: an Introduction to Methods*. London: Allen and Unwin.

Arber, S. (1993) 'Designing samples', in N. Gilbert (ed.), *Researching Social Life*. London: Sage. pp. 68–92.

Bernard, H. (1994) *Research Methods in Anthropology*, 2nd edn. London: Sage.

Bernstein, B. (1970) 'Education cannot compensate for society', *New Society*, 26th February, pp. 344–7.

Blair, P., Fleming, P., Bensley, D., Smith, I., Bacon, C., Taylor, E., Berry, J., Golding, J. and Tripp, J. (1996) Smoking and the sudden infant death syndrome: results from 1993–5 case-control study for confidential inquiry into stillbirths and deaths in infancy', *British Medical Journal*, 313: 195–8.

Bowling, A. (1997) *Research Methods in Health: Investigating Health and Health Services*. Buckingham: Open University Press.

Bradburn, N. (1983) 'Response effects', in J. Rossi, D. Wright and A. Anderson (eds), *Handbook of Survey Research*. New York: Academic Press. pp. 289–328.

Davies, J. (1997) *Drugspeak: the Analysis of Drug Discourse*. Amsterdam: Harwood Academic Publishers.

Department of Health (1997) *The New NHS: Modern, Dependable*. London: HMSO.

De Vauss, D. (1995) *Surveys in Social Research*, 4th edn. London: Allen and Unwin.

Doll, R. and Hill, A. (1964) 'Mortality in relation to smoking: ten years' observation of British doctors', *British Medical Journal*, 1: 1399–410; 1460–7.

Dwyer, T. and Ponsonby, A-L. (1996) 'Sudden infant death syndrome: after the "back to sleep" campaign', *British Medical Journal*, 313: 180–1.

Fielding, N. (1993) 'Qualitative interviewing', in N. Gilbert (ed.), *Researching Social Life*. London: Sage. pp. 135–53.

Fleming, P., Blair, P., Bacon, C., Bensley, D., Smith, I., Taylor, E., Berry, J., Golding, J. and Tripp, J. (1996) 'Environment of infants during sleep and risk of sudden infant death syndrome: results of 1993–5 case-control study for confidential inquiry into stillbirths and deaths in infancy', *British Medical Journal*, 313: 191–5.

Goldblatt, P. (ed.) (1990) *Longitudinal Study: Mortality and Social Organisation 1971–1981*. OPCS series LS no. 6. London: HMSO.

Layte, R. and Jenkinson, C. (1997) 'Social surveys', in C. Jenkinson (ed.), *Assessment and Evaluation of Health and Medical Care: a Methods Text*. Buckingham: Open University Press.

Kish, L. (1965) *Survey Sampling*. London: J. Wiley and Son.

Krejcie, R. and Morgan, D. (1970) 'Determining sample size for research activities', *Educational and Psychological Measurement*, 30: 607–10.

Mant, J. and Jenkinson, C. (1997) 'Case control and cohort studies', in C. Jenkinson (ed.), *Assessment and Evaluation of Health and Medical Care*. Buckingham: Open University Press. pp. 31–46.

Marmot, M. (1995) 'In sickness and in wealth: social causes of illness', *Medical Research Council News*, 65: 8–12.

Meltzer, H., Gill, B., Petticrew, M. and Hinds, K. (1995) *The Prevalence of Psychiatric Morbidity among Adults Living in Private Households: OPCS Surveys of Psychiatric Morbidity in Great Britain: Report 2*. London: HMSO.

National CJD Surveillance Unit (1999) *Creutzfeldt–Jakob Disease Surveillance in the UK: Seventh Annual Report*. Edinburgh: Western General Hospital.

Nazroo, J. (1997) *Ethnicity and Mental Health: Findings from a National Community Survey*. London: Policy Studies Institute.

Oakley, A. (1981) 'Interviewing women, a contradiction in terms', in H. Roberts, (ed.), *Doing Feminist Research*. London: Routledge and Kegan Paul. pp. 30–61.

Pett, M. (1997) *Nonparametric Statistics of Health Care Research: Statistics for Small Samples and Unusual Distributions*. London: Sage.

Robinson, W. (1950) 'Ecological correlations and the behaviour of individuals', *American Sociological Review*, 15: 351–7.

Sapsford, R. (1999) *Survey Research* London: Sage.

Shepherd, P. (1985) 'The National Child Development Study: an introduction to the background to the study and the methods of data collection'. London. *NCDS User Support Group Working Paper No 1*. Social Statistics Research Unit: City University.

Tudor Smith, C., Nutbeam, D., Moore, L. and Catford, J. (1998) 'Effects of Heartbeat Wales programme over five years on behavioural risks for cardio-vascular disease: quasi-experimental comparison of results from Wales and a matched reference area', *British Medical Journal*, 316: 818–22.

Unwin, N., Carr, S., Leeson, J. and Pless-Mulloli, T. (1997) *An Introductory Study Guide to Public Health and Epidemiology*. Buckingham: Open University Press.

Wadsworth, M. (1991) *The Imprint of Time: Childhood History and Adult Life*. Oxford: Oxford University Press.

Wadsworth, M. (1996) 'The survey that shocked the nation', *Medical Research Council News*, 69: 28–32.

West, D. (1969) *Present Conduct: Future Delinquency*. London: Heinemann Educational Books.

Wilson, P. and Elliot, D. (1987) 'An evaluation of the postcode address file and its use within OPCS', *Journal of the Royal Statistical Society: Series A*, 150 (3): 230–40.

Yin, R. (1994) *Case Study Research: Design and Methods*, 2nd edn. London: Sage.