



The Magenta Book

Guidance Notes for Policy Evaluation and Analysis

Chapter 5: What is sampling?

Published: March 2004

Government Chief Social Researcher's Office
Prime Minister's Strategy Unit
Cabinet Office
Admiralty Arch
The Mall
London SW1A 2WH

Contact: phil.davies@cabinet-office.x.gsi.gov.uk
Telephone: 020 7276 1862



Chapter 5

What is sampling?

5. Introduction

In a social researcher's ideal world all data would be collected by census. Then, for example, if it proved important to know what percentage of the population had experienced a crime in the last year, or what percentage of the population smoked, all members of the population would be 'approached' (either by an interviewer, by post, or by telephone) and asked to provide the information required.

Good as this may sound, it does not happen in practice (the single exception being in the decennial population census). The reasons for not taking such a bold approach include cost (the UK census is an extremely expensive exercise), practical problems, including the fact that the general population would get very tired of answering all the questions put to them, and, perhaps most importantly, because by using sampling methods it is usually possible to get close to the estimates we need by approaching only a very small percentage of the population. This means that we collect the data we need, but the burden of providing that information is spread thinly across the population.

This chapter describes how sampling works. We start with the basics, but in later sections move on to some of the complications that arise in real-world situations. By and large we concentrate on surveys of the general population, but some reference is made to surveys of other populations such as businesses.

5.1 How and why sampling works

General population survey sampling relies on the notion that it is not necessary to collect data from *all* people in order to generate statistics about that population. Intuitively this seems correct. For instance, if we wish to know the average systolic blood pressure of adults in GB, then it seems reasonable that if we measure the blood pressure of a random 1 in 1000 adults

then the average we get will be pretty close, if not exactly the same, as the average we would get if we measured all adults. Or, putting the problem another way, it seems reasonable to assume that if we measure the blood pressure of, say, 10,000 adults selected at random and then take the average, then repeating the exercise with 10,000 more adults and adding the two samples together, is unlikely to generate a very different estimate. After a certain sample size, there seems to be diminishing returns in collecting more data.

In order to understand why, and when, this intuitive understanding of sampling works it is helpful to turn the problem on its head, and ask when, and how, might sampling fail? In other words, what are the circumstances under which the estimate we get from our sample will be a poor approximation to the true 'all population' value?

One of the most damaging sampling failures occurs when the sample we select is systematically different to the population we are trying to represent. So, if our sample for measuring blood pressure included a higher percentage of young people than the population as a whole, then we can be reasonably sure our sample will give a blood pressure average that is too low (since blood pressure is known to increase with age). So, as a primary rule, the sample must be a fair representation of the population we are interested in.

The second potential failure is if we take too small a sample. To get an estimate of average blood pressure that is close to the population average, it seems clear that we can be more confident in our estimate if we take a sample of 10,000 than if we take a sample of 100. One of the major contributions of sampling theory is that, not only can we demonstrate that our intuition here is correct, we can also quantify how much more confident we can be with the larger sample size. Hence the ubiquitous 'confidence interval'. 'Confidence' in this sense, is written in terms of statements about how far away from the true estimate we think our sample estimate might possibly be: even with larger sample sizes we can never be absolutely sure we will match the population value, but there is a smaller probability of being a long way off this figure than there is with a smaller sample size.

The third factor that can make an impact is variability in the population. To take an example, imagine we are trying to estimate average height and average weight among adult women using a sample survey. Then, it is known that the vast majority of the female adult population is of a height within the range 145 to 175 cm, the standard deviation being about 6cm and the

mean 161cm. Whereas weight is about twice as variable, the range being about 45kg to 100kg, with a standard deviation of about 14kg (mean=63kg). Now, if you imagine the scenario where a sample is selected but we are very unlucky in our selection and just happen to over-sample taller women. Then it is clear that even under the very worst case scenario, where the sample *only* includes women who are at the top of the height range (no pun intended) the largest mean our sample can give us is about 175cm. In which case the difference between the true mean and the sample mean is 14cm. In other words, the largest possible margin between the sample mean and the population mean is 14cm. For weight however, if we were unlucky enough to select a sample with only the very heaviest women, then we would have a sample mean of about 100kg. So, for weight, the largest possible margin between the sample mean and the population mean is 37kg. This is a margin more than twice as large as the worst error margin for height. It follows from this type of argument, that if we want our sample to give an estimate that is close to the population value, we need to take into account how much variability there is in the variable we are trying to measure. All else being equal, the greater the variability the larger the sample size we need¹.

So, in summary, samples will give estimates reasonable close to population values if the sample we select has no systematic bias, and if the sample size is large enough. Furthermore, the more variable the population is the larger the sample size that will be needed.

One way to think about sample design is the art of minimising the chance of getting a skewed, or extreme, sample, whilst minimising the survey cost. The main factor that the sample designer can adjust is the sample size. But there are many other factors that play a part: sampling frame, stratification, clustering. All of these are discussed in the chapter, but we begin with the most fundamental, namely the *sampling frame*.

5.2 Sampling frames

All samples involve, at least conceptually, the notion of a sampling frame, the frame being a list that covers (hopefully) all the ‘elements’ (that is ‘persons’ usually) in the population that we are sampling from.

¹ This discussion assumes that you wish to have an estimate of mean weight that is of similar precision to your estimate of mean height. In practice this may not be the case, and your sample size calculations would need to take this into account.

The ideal sampling frame is a straightforward list of the elements we are trying to sample. So, for a population sample, a comprehensive list of members of the population would be ideal. And for a household survey a comprehensive list of households would be ideal.

In practice the ideal sampling frame hardly ever exists. In the UK, for instance, there is no population register that can be used for sampling purposes. The closest is the electoral register, but because inclusion on the electoral register is voluntary large sections of the population are excluded. Furthermore, it is now possible to include oneself on the electoral register but to refuse for your name to be used for any other purpose. This means that the electoral register tends to give biased samples.

The default alternative to the electoral register in the UK is the small-user Postcode Address File (PAF) which is the Post Office's list of all addresses in the UK, which receive less than 25 items of post per day. The list is primarily residential addresses (about 94%), and the list covers approximately 99% of all residential addresses.

The PAF reasonably closely approximates to a sampling frame of households, which makes it ideal for surveys of households (once non-residential and unoccupied addresses have been screened out). It can also be used as a sampling frame for individuals, but to do so the sample has to be selected in two stages: a sample of households (first stage) from which a sample of individuals is then selected (second stage). Depending upon the needs of the survey, it is usual to either select all adults at a household for an interview, or to select just one adult at random. How this choice is made is described in Section 5.3.

Use of the PAF is now very well established for face-to-face interviews of the general GB population. It does have one major shortcoming in that it tends to exclude institutions (such as care homes or university halls of residence) and most GB surveys have now become surveys of the 'general household population' rather than the 'general population'. To include institutions special boost samples have to be selected.

It is worth noting that using PAF as a sampling frame for individuals is relatively unproblematic *as long as* the survey is being carried out face-to-face, because the interviewer can do the selection of an individual within the household. For postal surveys of individuals

PAF sampling is really not possible because there is no good way of controlling who completes the questionnaire. Instructions, such as asking a household to select the person who has most recently had a birthday, or to distribute a questionnaire to all household members, are not adhered to strictly enough to be robust. So for population based postal surveys the default-sampling frame is still the electoral register, even though its problems are well known.

The third 'sampling frame' of the general population worthy of mention is more virtual than actual, namely the list of all possible residential telephone numbers. Although undoubtedly BT holds such a list, only the directory list is released to researchers: ex-directory numbers are excluded. Nevertheless, sampling methods *do* exist based on the entire list. In essence this involves selecting 11 digit numbers at random, the first 7 digits being randomly selected from the published list of prefix numbers (such as 020 8693 XXXX) and the last four digits being generated entirely at random. The selected numbers are then dialled and numbers not in use, and business numbers, are screened out to leave just residential numbers. This method of sampling is known as Random Digit Dialling (RDD).

The above discussion probably suggests that, for general population samples, the choice of sampling frame is largely determined by survey mode (PAF for face-to-face interviews, ER for postal surveys, and RDD for telephone surveys). This *is* largely the case, but the issues get more complicated when the aim is to target sub-samples of the population. For instance, a face-to-face interview survey of children could use a sample based on a PAF sample with screening out of adult-only households. But this would involve interviewers approaching more households than will be included in the survey. An alternative, more cost-effective approach, *might* be to use an alternative-sampling frame such as child benefit records. Or, depending on the nature of the survey, it might be possible to sample via schools (i.e. select a sample of schools and then a sample of children within schools). A fourth option would be to sample children from another survey: for instance the Health Survey for England might be used as a sampling frame for a follow-up survey of children. In practice, when there are various sampling options some will be ruled out quite easily, but others will involve hard choices between cost-effectiveness and potential biases introduced by using imperfect sampling frames or frames where the sampling mechanism is likely to create non-response problems.

5.3 Sample size

The main decision needed in deciding on a sample design is sample size. To decide on this a number of questions need to be decided on:

- (i) What are the key estimates for the study?
- (ii) How precise do those estimates need to be? (i.e. what size of standard error or confidence interval can be tolerated?)
- (iii) Are there key sub-groups for which separate estimates will be needed?
- (iv) Does the survey need to be large enough to detect change over time between surveys, or differences between key sub-groups?².

The basic formula that survey statisticians use to determine sample size is the ‘standard error for a mean from a simple random sample survey’:

$$sderr(\bar{x}) = \sqrt{\frac{S^2}{n}}$$

Where S^2 is the population variance and n is the sample size.

The 95% confidence interval for the mean is then calculated as

$$\bar{x} \pm 1.96sderr .$$

Note that, as per the discussion above, the standard error increases as the population variance increases, and decreases as the sample size increases.

In surveys being carried out for the first time S^2 has to be (gu)estimated. Life is simpler if we are interested in percentage rather than means because then S^2 becomes $p(100 - p)$ where p is the percentages. It is usually easier to estimate a value of p in advance than it is to estimate S^2 for a mean.

² Note, this last requirement needs a power calculation rather than a standard sample size calculation. Power calculations are discussed in Chapter 3

In practice simple random samples are relatively rare in practice and survey statisticians use an amended version of the basic formula:

$$sderr(\bar{x}) = deft \sqrt{\frac{S^2}{n}}$$

The multiplier ‘deft’ in the above equation is the ‘design factor’. The deft is essentially a factor that adjusts the standard error because of design features. These features include:

- (i) Stratification of the sample either to guarantee that sub-groups appear in the correct proportions (proportionate stratification) or to over-sample sub-groups (disproportionate stratification).
- (ii) Weighting of the sample to adjust for non-response
- (iii) Clustering of the sample.

Each of these is discussed below, beginning with the last on the list above: clustering.

5.4 Clustering

A ‘clustered’ sample is defined as a sample that is selected in two or more hierarchical stages, different ‘units’ being selected at each stage, and with multiple sub-units being selected within higher order units. A few examples will help to clarify this:

- A sample of children is selected by (a) sampling schools and then (b) selecting children within schools. This is a two-stage clustered sample, the clustering being of children within schools.
- On a general population survey a PAF sample is used to generate a sample of households. Within each household up to two adults are selected at random. This is a two-stage clustered sample, the clustering being of adults within households. Note that, had the instruction been to select just one adult per household, this would *not* be described as a clustered sample, because there would no clustering of the adult sample within a smaller number of households.

- The most common design for PAF samples is, at the first stage, to select a random sample of postcode sectors³. Then, at the second stage, households are selected within these postcode sectors. And then, at a third stage, individuals might be selected within households. Under this design adults are clustered within households (assuming more than one adult is selected per household) and households are clustered within postcode sectors.

Clustering, or multi-stage sampling, is adopted on surveys for a number of reasons. The two main reasons are:

- because the sampling frame units cover two or more survey units and so clustering is the only practical way of selecting a sample of the units required
- to divide the sample into manageable workloads for interviewers.

5.5 Clustering because of the sampling frame

If the sampling frame to be used for a survey consists of units larger than the survey units, then it is very common to use a clustered sample design. Generally speaking the more survey units each sampling frame unit covers the more clear-cut the case for clustering will be.

For instance, if a survey of employees is to be carried out using a sampling frame of business establishments, then the most cost-efficient solution is almost bound to be to select a sample of establishments and then to select a sample of employees per establishment. The 1998 Workplace Employee Relations Survey adopted this design, with a sample of about 2000 establishments being selected at the first stage, and then 25 employees being selected per establishment at a second stage, giving a total sample size of about 50,000. Although it would have been possible to take an unclustered sample by selecting 50,000 establishments and then one employee per establishment, the fact that this would involve negotiating access via such a large sample of establishments renders this approach completely impractical.

An instance where the merits of clustering are less clear-cut is the selection of individuals within households. On some general population samples several adults are selected per household (to give a clustered sample) whereas in other surveys just one adult is selected per household (giving an unclustered sample).

³ As an example, the postcode EC1V 0AX is part of the postcode sector EC1V 0.

5.6 Clustering to give manageable interviewer workloads

Almost all face-to-face interview surveys of the GB population use geographical clustering at the first stage of sample selection. The usual procedure is, as was noted above, to select a sample of postcode sectors and then to select PAF addresses within these postcode sectors. For instance, the main Health Survey for England sample is selected by firstly selecting 720 postcode sectors and then 20 addresses per sector.

The rationale behind this approach is that the sample of addresses in each postcode sector becomes the workload for one interviewer. So an interviewer is given 20 addresses to interview at, and these are located in a relatively small geographical area. The interviewer does not then have large distances to travel between sampled addresses, and the costs of the survey are reduced.

The observant reader may notice that, on the face of it, this sampling procedure (whereby equal sample sizes of addresses are selected per sector) means that people who live in large postcode sectors (i.e. postcode sectors with a large number of addresses) will be under-represented in the sample. This is avoided by over-sampling large sectors at the first stage of selection. More precisely, sectors are selected with probability proportional to their address count (probability proportional to size sampling). Taking an equal sample per sector at the second stage then counterbalances this over-sampling, to give an equal probability of selection per address overall.

5.7 The impact of clustering on standard errors

The main objection to clustered samples is that they tend to give estimates with larger standard errors than unclustered samples. That is they give a deft greater than one. The reasoning here is that the more the sample is clustered the greater the chance we have of drawing a sample that is extreme (see Section 2 above). For instance, imagine a scenario where we are selecting a sample of 1000 people. Then if we choose to select the 1000 people by selecting just 10 postcode sectors and 100 people per sector, then if we are unlucky enough to select one or two sectors that are outliers in some sense then we will get a sample mean that is quite different to the population mean. If, instead, we choose to select 100 postcode sectors and 10 people per sector, then the impact on sample estimates of, by chance, selecting one or two outlier postcode sectors will be much smaller. Under this less clustered

design we can be sure the sample will give estimates that will be reasonably close to the population mean. However, we could be even more confident if we unclustered the sample even further, by taking, say, 500 postcode sectors and just two people per sector.

From this example it is hopefully clear that the more the total sample is spread across clusters the lower the chance of taking an extreme sample and the lower the standard error. This translates into: for a fixed sample size, the smaller the sample size per cluster the smaller the standard error. Note however, that in the example above clustering will only be a problem if there is a risk of selecting a non-representative sample of postcode sectors (described above, as over-representing outliers). This can only happen if postcode sectors differ from each other, that is, if there is between-sector variance. If all sectors are the same then no matter what sectors are selected the survey estimates will be the same. So, standard errors associated with clustering increase *if* there is between-cluster variance, and as the sample size per cluster increases.

On the other hand, as was noted earlier, clustering of a general population sample within postcode sectors tends to reduce interviewing costs because it reduces travel costs for interviewers. So, decisions on the extent of clustering involves judgement about how standard errors can be minimised for a fixed survey budget – a clustered sample may give larger standard errors than a simple random sample, but a larger sample size will be affordable with a clustered sample, so the impact of clustering can usually be more than offset. For most face-to-face interview surveys a clustered sample will be the most cost-effective.

To model the effect of clustering survey statisticians make use of a crude estimate of the design factor

$$deft = \sqrt{(1 + (m - 1)roh)}$$

Where m is the average sample size per cluster and roh is a measure of the relative between-cluster variance⁴. roh values differ from estimate to estimate, but tend to be highest for variables that are very geographically clustered (such as tenure, and to a lesser extent deprivation). roh is very small for variables that are roughly constant across clusters (e.g. sex or age). One of the difficulties of survey design is that it is necessary to estimate roh in

⁴ roh varies from 0 to 1, being zero when there is no between-cluster variance, and 1 when all the variance between population members is between clusters, and there is no variance within clusters.

advance of doing a survey – unless the survey is a repeat of an earlier survey this is more of an art than a science.

It is worth noting that the value of roh will tend to vary depending upon the geographical definition of the clusters. Generally, the smaller the cluster, in geographical terms, the larger roh will be. This is the reason survey organisations tend to use postcode sectors, which cover areas of about 2300 households, rather than smaller geographical areas such as enumeration districts.

5.8 Clustering within households

Using PAF as a sampling frame for general population surveys means that a decision is always needed on whether to select one adult per household or more than one adult per household. For instance, the Health Survey for England selects all adults per household, and the annual British Social Attitudes survey selects just one.

The arguments for and against are:

- (i) Selecting more than one adult per household introduces a second tier of clustering into the sample. This will tend to increase standard errors.
- (ii) But selecting just one adult per household means that adults from larger households are under-represented in the final sample. This has to be adjusted for by weighting the final sample (see Section 8 below). This will also increase standard errors.

So both options increase standard errors. But

- (iii) Selecting more than one adult per household means that fewer households need to be included in the final sample. This reduces survey costs.
- (iv) However selecting more than one adult per household puts a lot of burden on the household and this can increase non-response.

In practice it is usual to select more than one adult per household *if*

- (a) this is not expected to excessively burden the household; and
- (b) household members are not expected to be too homogeneous in terms of the things the survey is trying to measure (which would give large roh values); and/or
- (c) most of the analysis will be done within male/female sub-groups (which has the effect of giving relatively unclustered subgroups).

5.9 Stratification

Alongside decisions on how to cluster a sample, decisions also need to be taken on stratification.

Stratification essentially means dividing the sampling frame into groups (strata) before sampling. A simple example would be to take a sampling frame of, say, business establishments and then to sort them into size strata before sampling. The sample would then be described as a sample stratified by size. If a list of the general population was available that had age and sex recorded, then it would be possible to divide the list into age and sex strata before sampling to give a sample stratified by age and sex.

There are two methods of stratified sampling: proportionate and disproportionate. In a proportionate stratified sample the sampling frame is divided into strata but the same sampling fraction is applied per stratum. This means that each stratum is sampled from in its correct proportion. In a disproportionate stratified sample the sampling fraction differs between strata. This means that individuals from the strata with the highest sampling fractions will be over-represented in the sample. Disproportionate sampling is generally used when there is a need to boost the sample size within a particular stratum or strata.

Proportionate stratified sampling is a more powerful tool than it may appear to be at first glance. The main advantage it has over simple unstratified random sampling is that it *guarantees* that the sample drawn matches the sampling frame in terms of the strata. So, for instance, if the sampling frame is stratified by age and sex and a proportionate sample selected, then the sample will match the sampling frame in terms of the age-sex distribution. In other words the age-sex distribution is controlled. If the survey data happens to be correlated with age and/or sex, then it follows that by stratifying the sample by age and sex there is less risk of drawing an extreme sample which gives survey estimates far removed from the population values. In other words, stratification reduces standard errors.

The degree to which standard errors are reduced depends on how closely associated the strata variables are to the survey estimates. For example, in a health survey measuring physical health, stratification by age would be very powerful because physical health is so closely related to age. Stratification by sex would be useful, but less so, because there is a weaker

relationship between sex and physical health than there is between age and physical health. In contrast, for a survey of mental health, sex would probably be the better stratifier, because there is a stronger relationship between sex and mental health than there is between age and mental health.

Even though stratification is very useful in minimising standard errors, its use is often restricted by the fact that it is only possible to use variables as stratifiers if they appear on the sampling frame. So, in practice, age and sex stratification for general population samples is very rare because none of the usual sampling frames include age and sex. For PAF based samples the possible stratifiers are all geographical indicators, such as location, and characteristics of postcode sectors as derived from the most recent census (such as percentage of households with an unemployed head of household). There are no good stratifiers at the level of individual addresses.

Typically, what happens on PAF samples is that PAF is divided into regional strata (typically Government Office Region) and then, within regions, postcode sectors are divided into strata using one or two variables thought to be reasonably closely related to the survey subject matter. So a survey of income might use an area-level deprivation index as a stratifier for postcode sectors, and a travel survey might use an urban/rural stratifier.

We noted earlier that clustering of PAF samples within postcode sectors tends to increase standard errors, giving a deft of greater than one. The effect of selecting the sectors within strata is to reduce the impact of the clustering (in the sense that it reduces the risk of selecting a skewed sample of sectors) with the result of reducing the deft. The experience on most surveys is that, with careful selection of stratifiers the deft can be quite significantly reduced, but it tends still to be greater than one. In other words, a clustered sample with clusters selected using stratification will still give larger standard errors than an unclustered random sample of the same size. Nevertheless, stratified clustered samples have been found to be the most cost-effective sample design for face-to-face interview surveys: for the same cost, larger sample sizes can be achieved than with unclustered designs, and the increased sample size more than offsets the design factor due to clustering.

5.10 Random versus quota sampling

Most government-sponsored surveys use random sampling methods (or probability sampling as it is often called). This is the most robust form of sampling methodology.

Random sampling, in a PAF-based clustered sample, means that the postcode sectors will be selected within each stratum at random (albeit with probability proportional to size). A random sample of addresses will be selected within selected sectors, and assuming there is some selection within households, individuals will be selected from within a household using a random sampling method. The aim of random sampling is to avoid any self-selection bias in the sample, whereby areas are selected because interviewers prefer to work there, and individuals are selected who are more willing or able to take part in surveys.

In strict random samples, once the sample is selected there can be no deviation from the sample. So those who refuse to take part become 'non-respondents' and, importantly, no attempt is made to replace non-responders with others willing to respond.

The main alternative sampling method to random sampling is quota sampling. In comparison to random sampling this method is less robust, but it is used fairly extensively in market research.

Quota samplers allow substitution of non-respondents with others willing to respond, but they make considerable use of stratification principles to ensure the final sample reflects the population at least on some key variables. For instance, a quota sampler might use population estimates of the numbers within each combination of age group, sex, and social class to decide what numbers is needed within each of these combinations for a survey. Interviewers are then given quotas based on these numbers that they are asked to achieve.

The theory behind quota sampling is that, as long as the variables used to determine the quotas are selected carefully enough, then the fact that within a quota cell there is no attempt at random sampling is not important. The survey estimates will still be unbiased. The underlying assumption is that, within a quota cell, those who take part in the survey have the same characteristics, attitudes, behaviours etc. as those who do not take part.

The skill is to find variables for the quota that control for most of the survey variability between respondents and non-respondents – if this can be done then the assumption that responders are similar to non-responders becomes more credible. In reality though one can only control for a limited number of variables. There may be a number of factors that one did not take into account that produce a systematic difference between respondents and non-respondents. Furthermore, where the aim of the survey is to produce reliable estimates about the views of the population that the survey relates to it is necessary to employ random sampling. This is because everyone does not have an equal opportunity to be interviewed in a quota sample and it is not appropriate to apply standard statistical techniques which allow you to estimate the degree of confidence that you can place on the results of an individual survey.

5.11 Sampling special populations using screening

It is often the case that surveys have a particular focus on sub-groups of the general population rather than of the whole population. Some examples include:

- the 2003 Health Survey for England which had a focus on child health and included a boost sample of children
- the 1996 British Crime Survey which included a boost sample of people from minority ethnic groups
- The Low Income Diet and Nutrition Survey which specifically includes only low-income households.

The usual approach to selecting the sample in these instances is to select a large PAF based sample, to carry out a short screening survey at all households (which might be done on the doorstep if it is very short) and to carry out a full interview only in households with the relevant people.

The main issue here tends to be cost, since, in general, interviewers will have to screen out more households than they include. Boost samples of children are relatively inexpensive because a fairly large percentage of households have one or more children (about 30%). Boost samples of minority ethnic groups are far more expensive because interviewers have to screen at a large number of addresses to achieve the final sample. In fact, minority ethnic boost samples tend to include over-sampling (i.e. disproportionate stratification) of postcode

sectors where the percentage of ethnic groups is higher than average in an attempt to improve interviewer screening rates and, hence, to reduce survey costs.

The success of screening largely depends on our ability to ask screening questions quickly, and at the start of interviews. This is pretty straightforward when screening for children and minority ethnic groups – this can be done on the doorstep as long as the interviewer is careful to explain why the questions are being asked. Other screening questions are harder to ask up-front, and surveys have on occasion used proxy measures. An example of this is the Low Income Diet and Nutrition Survey where it was not considered feasible to ask detailed questions about income on the doorstep, so, instead, a proxy indicator of relative deprivation was used. This included questions on tenure, car ownership, employment, lone parent status, and benefit receipt.

5.12 Survey Weighting

In most surveys it will be the case that some groups are over-represented in the raw data and others under-represented. This might be because of the sample design, primarily because of the use of disproportionate stratification or boost sampling, or because of sampling features that lead to unequal probabilities of selection, such as selecting one person per household on PAF samples. Alternatively some groups may be over- or under-represented because of non-response patterns.

These mis-representations are usually dealt with by weighting the data. The idea behind weighting is that members of sub-groups that are thought to be over- or under-represented in the survey data are each given a weight. Over-represented groups are given a weight of less than one; under-represented groups are given a weight of greater than one, the weight being calculated in such a way that the weighted frequency of groups matches the population. All survey estimates are calculated using these weights, so that averages become weighted averages, and percentages become weighted percentages, and so on.

The calculation of the weights for a survey is rarely a straightforward business. Weights for disproportionate sampling are relatively non-controversial, but weights to adjust for non-response biases are largely dependent upon judgement, and it is likely that no two analysts would ever calculate exactly the same set of non-response weights. Nevertheless, following

the GSS task force on weighting some standardisation is now coming into play. The main principles are:

- non-equal probabilities of selection (including disproportionate stratification) is dealt with by applying weights proportional to the inverse of the probability of selection;
- at a minimum non-response is dealt with by weighting survey data to published distributions by age, sex and region.

The actual means of calculating these non-response weights can differ from survey to survey, but the most commonly used method now is ‘calibration weighting’ (see box 5.1 for an explanation of calibration weighting and how the method has been used in the British Crime Survey, a household survey of crime victimisation).

Box 5.1: Calibration weighting and the British Crime Survey (Source: Crime in England and Wales 2002/2003, Home Office Statistical Bulletin 07/03)

The Office for National Statistics (ONS) recommended that the calibration weighting method be adopted in the British Crime Survey (BCS). The weighting is designed to make adjustment for known differentials in response rates between different age by gender subgroups and households with different age and gender composition. For example a 24 year-old male living alone may be less likely to respond to the survey than one living with a young partner and a child. The procedure therefore gives different weights to different household types based on their age/sex composition in such a way that the weighted distribution of individuals in the responding households matches the known distribution in the population as a whole (based on population estimates provided by ONS). The weights are generated using an algorithm (CALMAR) that minimises the differences between the weights implied by sampling and the final weights subject to the weighted data meeting the population controls.

The calibration weighting method is now used on the General Household Survey (ONS), the Expenditure and Food Survey (ONS and DEFRA), the Family Resources Survey (DWP), Family and Children’s Survey (DWP) the Labour Force Survey (ONS), and other surveys.

By and large weighting of survey data tends to increase the standard errors of estimates. A key issue for non-response weighting is whether the reduction in survey bias is adequate compensation for the increase in standard errors. In some surveys it will be, in others not.

Further reading

Barnett, V. (2002). *Sample Survey Principles & Methods*, Arnold (3rd Edition).

Barton, J. (1996). *Selecting Stratifiers for the Family Expenditure Survey (FES)*, in *Survey Methodology Bulletin* **39**, 21-26.

Cochran, W. G. (1977). *Sampling Techniques*, Wiley (3rd edition).

Elliot, D. (1991). *Weighting for Non-Response: A survey researcher's handbook*, OPCS.

Greenfield, T. (1996). *Research Methods: Guidance for Postgraduates*, ed. Arnold.

Kalton, G. (1983). *Introduction to Survey Sampling*, Sage.

Kalton, G. (1983). *Compensating for Missing Survey Data*, Institute for Social Research, University of Michigan.

Kish, L. (1992). *Weighting for Unequal P_i*, *Journal of Official Statistics*, **8**, 183-200.

Kish, L. (1965). *Survey Sampling*, Wiley.

Moser, C. A. and Kalton, G. (1971). *Survey Methods in Social Investigation*, Gower (2nd edition 1971).

Lynn, P. and Lievesley, D. (1991). *Drawing General Population Samples in Great Britain*, SCPR.

Lynn, P. and Taylor, B. (1995). *On the bias and variance of samples of individuals: a comparison of the electoral registers and postcode address file as sampling frames*, *The Statistician*, **44**,173-194.

Stuart, A. (1984). *The Ideas of Sampling*, Griffin.