

# 8

## Sampling

It is often decided to study only a part, or sample, of the study population (the 'sampled' or 'parent' population). Samples may be chosen, for example, of residents of a neighbourhood, of people with (or without) a given disease, or of people exposed (or not exposed) to a suspected causal factor. The decision to sample may be forced on the investigator by a lack of resources. The procedure may make for better use of available resources – because of the restricted number of individuals to be studied, it is possible to investigate each of them more fully than might otherwise have been possible, and to make greater efforts to ensure that information is in fact obtained from each individual. Frequently, it is decided to have the best of both worlds, by obtaining easily acquired types of information about the total study population, but limiting certain parts of the study, which require more intensive investigations, to one or more samples.

Provided that certain conditions are met, there is no difficulty in applying the results yielded by a sample to the parent population from which it has been selected, with a degree of precision that meets the investigator's requirements. Statistical techniques are available that make it possible to state with what precision and confidence such inferences may be made. The conditions to be met are:

1. The sample must be *well chosen*, so as to be representative of the parent population.
2. The sample must be *sufficiently large*. If a number of representative samples drawn from the same parent population are investigated, it can be expected that, by chance, there will be differences between the findings in each sample; this problem of *sampling variation* is reduced if the sample is large.
3. There must be *adequate coverage* of the sample. Unless information is in fact obtained about all or almost all members of the sample, the individuals studied may not be representative of the study population.

Mere size is not enough. A sample that is badly chosen or inadequately covered remains a biased one, however big it may be. This was strikingly shown by the notorious poll conducted by the Literary Digest in 1936, which, although based on 2,000,000 ballots, dismally failed to predict Roosevelt's landslide victory in the presidential election. These ballots constituted 20% of the 10,000,000 that had been sent out to an unrepresentative sample comprising Literary Digest and telephone subscribers.

How can a survey finding that 22% of sword swallowers have a history of perforation be relied upon, if responses were received from less than half of the sample?<sup>1</sup>

This chapter deals with sampling methods and sample size, followed by remarks on substitutions and random numbers.

## Sampling Methods

A sample chosen in a haphazard fashion, or because it is 'handy', is unlikely to be a representative one. Such samples have been termed 'chunks' or 'accidental' or 'incidental' samples, or *samples of convenience*. Their use has no place in community medicine research, except possibly in exploratory and other surveys where the investigator is doing no more than obtaining a 'feel' of the situation, and in some qualitative studies (see pp. 147–149 and 318).

The recommended method is *probability sampling*, the distinctive feature of which is that each individual unit in the total population (each *sampling unit*) has a known probability of being selected. Generalizations can then be made to the 'parent' population with a measurable precision and confidence (see p. 260).

First, however, a word on the nonprobability sampling methods that are sometimes used: quota, purposive and snowball sampling. In *quota sampling*, the general composition of the sample, e.g. in terms of age, sex and social class, is decided in advance; quotas, or required numbers, are determined for, say, men and women of different ages and social classes, and the only requirement is that the right number of people be somehow found to fill these quotas. The disadvantage of this method is that the persons chosen may not be representative of the total population in each category, and generalizations made from the findings may be incorrect. *Purposive samples* are those selected because the investigator presumes that they are typical of the study population. In a study of general practices, for example, what is believed to be a representative cross-section of practices may be selected; subsequent generalizations from the findings may or may not be valid. In some qualitative studies (see p. 147) subjects are purposively selected not in order to represent the study population, but in such a way that they will express a wide range of the beliefs, practices or experiences under study. In *snowball sampling*<sup>2</sup> (*chain referral sampling*), people who meet the criteria for inclusion in the study are asked to name others who meet these criteria. This may be a useful way of identifying hard-to-find individuals, e.g. those with deviant or illegal behaviour, or homeless people. But the sample, say of drug abusers, will not necessarily be representative of all drug abusers.

We will discuss four types of *probability sampling* (random, systematic, cluster and stratified) – with special mention of random digit dialling – and two-stage and multi-stage sampling.

It is important to set up the sampling rules in advance and to avoid any possibility that selection may be influenced by whim or convenience. The interviewers in a household survey, for example, should be told in advance which homes to visit. If inclusion in the sample depends on information that is collected during the visit, then the interviewer should be given precise instructions for making the choice.<sup>3</sup>

### *Random sampling*

Random sampling (or *simple random sampling*) is a technique whereby each sampling unit has the same probability of being selected: the laws of chance alone decide which of the individual units in the parent (or 'target') population will be selected. To avoid confusion with the colloquial meaning of the word 'random', i.e. 'haphazard' or 'without a conscious bias', the term *strict random sampling* is sometimes preferred.

The basic procedure is:

1. Prepare a *sampling frame*. This is usually a list showing all the units from which the sample is to be selected, arranged in any order. For example, it might be a list of the registered patients in a particular practice. The preparation of a sampling frame may sometimes require considerable effort; it is seldom easy, for example, to obtain an up-to-date list of the elderly people living in a neighbourhood. If a population registry is maintained in a country or city, this may constitute the sampling frame; but such registers are often out of date, especially with regard to addresses; moreover, with the increase in concern for the individual's right to privacy, registers of this kind are becoming less accessible to investigators. Voters' lists, telephone directories, or lists of people with driving licences may be used, but these may tend to leave out some categories of people.<sup>4</sup> If the frame is an incomplete and biased representation of the study population, the sample will inevitably be biased too, however strictly the rules of random sampling are applied.
2. Decide on the size of the sample (see pp. 83–85).
3. Select the required number of units at random, by drawing lots or using random numbers. The use of random numbers is explained on pp. 86–87. When one matched control has to be chosen randomly from a small group of suitable candidates, it is often simplest to draw lots or (if there are up to six candidates) to throw a die.

The ratio 'number of units in sample/number of units in sampling frame' is referred to as the *sampling ratio* or *sampling fraction*. It is usually expressed either in the form '1 in  $n$ ' (e.g. '1 in 3', '1 in 4', etc.) or as a percentage or proportion.

Random sampling does not ensure that the characteristics of the sample and the population will coincide exactly; chance differences will exist, but by the use of appropriate statistical methods<sup>5</sup> it is possible to calculate the probability that these divergences lie within given limits (see p. 260).

Random sampling may be applied not only to the selection of subjects from a population, but to the selection of times or locations. In the latter instance, geographic areas or the coordinates of points are used as the sampling units; the sampling frame may be a map rather than a list.

### *Random digit dialling*

In a region where nearly everyone has a telephone at home, such as the United States, where under 5% of households were without a landline telephone in 2004, *random*

*digit dialling* – i.e. phoning numbers selected at random – is a convenient way of selecting a random sample, either for telephone interviews or for subsequent home interviews or other investigations. Phone numbers are kept in the sample only if they turn out to be for residential addresses. If there is no reply, the call is repeated a number of times, at different times and on different weekdays. A two-stage procedure may be used, whereby a sample of households is first selected by random digit dialling and information is then obtained about the members of the household, the subsequent selection of subjects being determined by age, sex, or other eligibility criteria, or by using a random or systematic selection rule, such as the choice of the member with the latest birthdate in the year. The detailed procedure<sup>6</sup> is designed in a way that reduces the proportion of wasted calls; unlisted numbers are not excluded. Random digit dialling is generally regarded as preferable to sampling from telephone directories, which exclude unlisted numbers.

High success rates have been reported. In some studies in the United States, information on household composition was obtained for over 90% of the residential numbers phoned, and over 80% of the eligible subjects were subsequently interviewed. Samples selected by random digit dialling have been reported to be reasonably representative of the general population; but the possible selective exclusion of underprivileged population groups and overrepresentation of households with more than one phone line may be important in some studies.

The utility of random digit dialling has, however, been impaired by new technologies,<sup>7</sup> such as cell phones, answering machines, voicemail, and caller identification, and by the public's resentment of telemarketing and opinion polls, which have led to a drop in response rates. Response rates for the University of Michigan's Survey of Consumer Attitudes, for example, declined from 72% in 1979 to 48% in 2003. A comparison of 17 North American studies using random digit dialling to find controls for childhood cancer cases revealed a decrease in the response rate from over 80% in the 1980s to 50–67% after the mid-1990s, mainly due to a drop in the percentage of households in which the phone was answered.

Of particular importance is the exponential growth in the use of cell phones (at the end of 2005, 62% of households in the United States had more than one cell phone). In many countries, this has been accompanied by an increase in the proportion of households without landline phones. Random digit dialling surveys in the United States – where people who have only a cell phone tend to be either young and relatively well off, or poor members of minority groups – do not include cell phones in their sampling frame. The reasons include difficulties in finding sampling frames, the fact that cell phones are personal and not linked to households (making the selection of a random sample difficult), the receipt of calls while driving and in other awkward situations, and the subscriber's obligation (in many cell phone plans) to pay for incoming calls. The inclusion of cell phones in random digit dialling surveys is less problematic in countries where cell phone users do not have to pay for incoming calls, such as Brazil and Finland. Telephone survey researchers have been urged to see the cell phone problem as an opportunity rather than a roadblock, and to find solutions – basically, the use of mixed-mode and multiple-frame approaches – to maximize coverage.<sup>7</sup> In studies of

groups with a high usage of cell phones, text (SMS, short message service) messaging may be a useful way of stimulating and obtaining survey responses.<sup>7</sup>

### *Systematic sampling*

Instead of selecting randomly, a *predetermined system* may be used. The usual technique requires a list, not necessarily numbered, of all the sampling units. Having decided on the size of the required sample, the investigator divides the number in the list by this required size in order to calculate the sampling ratio, expressed as '1 in  $n$ ', rounds  $n$  off to the nearest whole number, and uses this figure  $k$  as a *sampling interval*. Every  $k$ th item in the list is then selected, starting with an item (from the first to the  $k$ th) selected at random. This technique is often easier than simple random sampling.

Such a sample can be considered as essentially equivalent to a random sample, provided that the list is not arranged according to some system or cyclical pattern. If a 1-in-30 systematic sample is selected from a list of persons arranged according to decreasing age, there may be an appreciable age difference between a sample where the first member selected was the first on the list and one where the first person selected was the 30th. If the list is one of dwelling units, listed in such a way that ground-floor and upper-floor dwellings alternate, then a 1-in-2 systematic sample (or any systematic sample using an even number as the sampling interval) will contain either ground-floor dwellings only or upper-floor dwellings only.

Other methods of systematic sampling may be used that do not require prior listing of the sampling units. For example, it may be decided to select every third patient admitted to a hospital, or every patient whose personal identity number, social security number, hospital registration number, or birthdate (day of month) ends with a predetermined and randomly selected digit or digits. These methods are usually chosen because of their convenience.

### *Cluster sampling*

In cluster sampling, a simple random sample is selected not of individual subjects, but of groups or clusters of individuals. That is, the sampling units are clusters and the sampling frame is a list of these clusters. The clusters may be villages, apartment buildings, classes of schoolchildren, schools, general practices, housing units, households or families (note that these latter terms are not synonymous),<sup>8</sup> etc.

This is often a convenient method, especially when at the outset there is no sampling frame showing all the individual subjects. It is, of course, also more convenient to investigate people living in a relatively small number of households or villages, rather than the same number of persons who have been selected randomly and whose places of residence are, therefore, more scattered.

The technique has the disadvantage, however, that if there is a degree of similarity between the people in each cluster, (*homogeneity*, *high intraclass correlation*), it

becomes difficult, without special methods of analysis,<sup>9</sup> to estimate the precision with which generalizations may be made to the parent population.

Other things being equal, a large number of small clusters is preferable to a small number of large clusters.

### *Stratified sampling*

To use this method, the population (the sampling frame) is first divided into subgroups or *strata* according to one or more characteristics, e.g. sex and age-group, and random or systematic sampling is then performed independently in each stratum (*stratified random sampling, stratified systematic sampling*).

This procedure has the advantage that there is less sampling variation than with simple random or systematic sampling. It eliminates sampling variation with respect to the properties used in stratifying; and if the strata are more uniform than the total population with respect to other attributes, it also reduces sampling variation with respect to other properties. The greater the differences between the strata and the less the differences within the strata, the greater is the gain due to stratification.

The same sampling ratio may be used in all strata. This is called *proportional allocation*, since the number of individuals chosen in each stratum is proportional to the size of the stratum. Alternatively, different sampling ratios may be used in different strata (*disproportionate stratified sampling*). This permits heavier sampling in subgroups with few members, so as to provide acceptable estimates, not only for the population as a whole, but also for each of its subgroups. In a clinical trial, for example, this can ensure that the sample will contain enough elderly subjects for separate study – the effectiveness and safety of a treatment often differs in younger and older people.

Estimates for the total population are prepared by combining the data for the various strata. If varying sampling fractions are used, then an appropriate weighting procedure is required.<sup>9</sup> If a uniform sampling fraction is used, then the sample is *self-weighting* and can, for some purposes, be treated as if it were a simple random or systematic sample. The use of varying sampling ratios greatly adds to the complexity of the analysis and should not be decided upon lightly. There is, of course, no objection to the use of different sampling ratios if the strata are to be kept separate throughout the analysis, e.g. if people of different religions are to be studied as separate groups.

### *Two-stage and multistage sampling*

In *two-stage sampling*, the population is divided into a set of first-stage sampling units (*primary sampling units*), and a sample of these units is selected by simple random, stratified or systematic sampling. Individuals are then chosen from each of these primary units, using any method of sampling. The sample may be biased if very few first-stage units are selected.

The first-stage units may be census tracts, villages, classes of schoolchildren, households, or other aggregations. They may be time periods, e.g. if the samples are the patients who attend a clinic on randomly chosen days. This method has the same advantages as cluster sampling: less travel by interviewers, fewer school teachers to negotiate with, no need for a sampling frame showing all individuals in the population, etc. Two-stage cluster sampling (the selection of clusters within the chosen first-stage sampling units) will be discussed in Chapter 31.

The analysis is simplified if *self-weighting* procedures are used. These ensure that each individual has an equal chance of entering the sample (the *equal probability of selection method*, or ‘*epsem*’ sampling). One method is to select primary units with a probability proportional to their size (*PPS* sampling), and then choose an equal number of individuals from each primary unit; for an example, see pp. 314–315). Another is to choose the primary units by simple random or systematic sampling and then choose samples proportional to the sizes of the primary units by applying the same sampling fraction in each primary unit.<sup>10</sup>

Multistage sampling is used in large-scale surveys. A sample of first-stage sampling units is chosen, each of the selected units is divided into second-stage units, samples of second-stage units are selected, and so on. Different methods (simple random, stratified, systematic or cluster sampling) may be used at any stage.

## Substitutions

It usually happens that, after a sample has been selected, it is found that some of the selected subjects cannot be investigated. People may have died or moved away, may refuse, or may be unavailable for a variety of other reasons. It is tempting to replace such subjects with other randomly selected subjects. This is an acceptable procedure (although an unnecessary one if expected losses were taken into account when deciding on the required sample size) provided that it is remembered that if the omissions produce a sample bias then substitutions will not remove this bias. The outcome will merely be a large biased sample instead of a small biased sample. What is important, if there are more than a few omissions, is to examine the possible bias by determining the reasons for omission and, if possible, studying the demographic and other characteristics of the subjects omitted; the relevance of this bias to the study findings can then be appraised.

## Sample Size

If numbers are too small it may be impossible to make sufficiently precise and confident generalizations about the situation in the parent population, or to obtain statistical significance (see p. 273) when associations are tested. It may thus be impossible to achieve the study’s objectives. On the other hand, it is wasteful to study more subjects than these objectives require. Moreover, if numbers are large

enough, then any difference, however small, will be statistically significant and there may, hence, be a tendency to ascribe false importance to trivial differences. ('Samples which are too small can prove nothing; samples which are too large can prove anything.')

<sup>11</sup>

'How big should my sample be?' has been likened to the question 'How much money should I take when I go on vacation?'

<sup>12</sup> (How long a vacation? Doing what? Where? With whom?) Calculations of sample size require both decisions and surmises.

For example, suppose a simple random sample is to be used to provide a confidence interval of a given width for the prevalence of a disease, i.e. to indicate the range within which it is probable (with a given degree of confidence, usually 95%) that the true prevalence lies. To calculate the size of the sample needed for this purpose, the following must be plugged into the formula or the computer program:

1. A reasonably close estimate of the actual prevalence (if in complete doubt, 50% can be used; this maximizes the sample size and, hence, errs on the safe side).
2. The maximum acceptable difference between the estimated prevalence (based on the sample) and the actual prevalence; this 'acceptable margin of error' is half the confidence interval.
3. The required confidence level (usually 95%).
4. Also, optionally, the size of the population; this is relatively unimportant – its effect on the calculated sample size (the *finite population correction*)<sup>13</sup> is small, unless the population is very small.

To calculate the sizes of the random samples required for a comparison of two groups, e.g. to test whether there is a significant difference between the rates of some outcome in two groups in a trial or analytic survey, the requirements are:

1. Whether a two-sided or one-sided testing procedure will be used. In a trial comparing two treatments, A and B, a two-sided test examines the study hypothesis, i.e. the alternative to the null hypothesis (see note 10, Chapter 27), that A and B have different effects, whereas the study hypothesis for a one-sided test is either that A is better than B, or that B is better than A.
2. A reasonably close estimate of the actual rate in one group.
3. The magnitude of the difference (or odds, rate or risk ratio) to be detected.
4. The relative size of the two samples.
5. The required significance level (e.g. 0.05).
6. The required power (e.g. 90%), of the test (its ability to detect the difference) or the required precision (the width required for the confidence interval).

When calculating the sample sizes for a comparison of two groups, a two-sided testing procedure is usually stipulated. But the sample sizes required for a one-sided test are somewhat smaller, and, especially in a trial, smaller samples are to be preferred for both ethical and practical reasons. Therefore, it has been suggested that sample sizes should be calculated for a one-sided test whenever this is appropriate.<sup>14</sup> Specifically, this would apply to a trial comparing treatment A with treatment B (or a placebo),



where the research question is whether A is clearly better than B with respect to a defined clinical end-point and where there is no interest whatever in knowing whether B is better than A, because such a finding would have no practical implications. Although this seems a reasonable suggestion, some authors object to it, on such grounds as that a one-sided test might not detect harmful effects of treatment A.<sup>15</sup>

Computation of sample sizes can be done by using computer programs<sup>9</sup> or by manual calculation or the use of tables or nomograms.<sup>16</sup> The computed size should be increased to allow for the loss of members of the sample; a larger sample will be needed if separate analyses of subgroups are intended.

Consideration must be given to practical constraints. A large sample may be difficult or impossible to find, or there may be an insufficiency of resources or time. A balance may have to be struck between the cost and the usefulness of the sample. The larger the sample, the less the sampling variation, i.e. the less the likelihood there is that the sample will be a misleading one. As a very rough guide, the usefulness of a sample is proportional not to its absolute size but to the square root of its size. To double the usefulness of a sample, its size must be increased fourfold; above a sample size of about 200, the absolute size of the sample must be augmented considerably to make an appreciable difference to its usefulness. This means that it may be necessary to balance increased cost (largely determined by the size of the sample) against increased usefulness (largely determined by the square root of its size). A *sensitivity analysis* may be helpful, i.e. a series of calculations of sample size based on different assumptions and requirements.

Samples that are to be compared with one another, e.g. in case-control studies and clinical trials, are usually kept approximately equal in size, since (for a given total sample size) this provides the most precise results (i.e. a measure of association that has a narrow confidence interval). But equal groups are by no means essential, and there may be good reasons for having unequal ones.<sup>17</sup> The relative size of the groups must be taken into account when calculating sample size.

In some therapeutic and prophylactic trials in which the subjects enter the investigation serially, as they become available, no initial decision is made about the sample size. Instead, rules are set up in advance whereby at any stage it can be decided, on the basis of the findings to date, whether enough subjects have been studied to give a sufficiently definite answer so that the trial can be stopped. This procedure is termed *sequential analysis*.<sup>18</sup>

A basic difficulty in calculations of sample size, whether they are done in advance or by the sequential method, is that the result depends on the attribute that is to be measured or compared. Samples of very different sizes are needed to study differences between two groups in their blood lipid levels, in their incidence of coronary heart disease, or in their mortality rates. It is seldom that a study is conducted to investigate only a single characteristic, and the real question often becomes not 'How many subjects do I need?' but 'With such-and-such a sample size (determined by practical considerations), about what variables and about what associations can I expect to get useful findings? – and in these circumstances, is the study worth doing?'

*Cluster samples* present a special case.<sup>19</sup> The required sample size is generally larger than for a simple random sample in the same study population, for the same maximum acceptable difference and confidence level.

The *power* of a test (i.e. its ability to demonstrate a difference if it exists) for given sample sizes can be appraised by the same basic formulae as for calculating sample size, but used in reverse, i.e. sample sizes are entered instead of power, and power is calculated instead of sample sizes. This can be very useful information when a study is being planned. But calculating power *after* the study has been done has been called an abuse of power – it is generally inappropriate, and may be misleading.<sup>20</sup>

Random Numbers

Tables of random numbers (digits arranged in a random order) are to be found in most statistics textbooks, or can be provided by a computer.<sup>21</sup> A short specimen (provided as an illustration, and not for use) is shown here (Table 8.1), and a table for actual use is provided in Appendix B.

Readers who intend to use computer programs to select samples can safely skip this section.

To use a table of random numbers for selecting a sample, a number must first be allocated to each sampling unit (e.g. from one to the total number of sampling units). Successive random numbers are then read from the table, and the sampling units whose numbers coincide with these random numbers are chosen. This is continued until enough units have been selected. Numbers not appearing in the list of sampling units are ignored, and numbers that reappear after they have already been selected are generally also ignored. The starting point is chosen at random, e.g. by shutting one’s eyes and using a pin.

As an example, if five units are to be chosen out of nine, numbered from 1 to 9, one could start at (say) the ‘8’ in row 4 of Table 8.1 and read off numbers 8, 9, 3, 1 and 6 (moving horizontally). Or one could move vertically and select the units numbered 8, 6, 7, 9 and 5; the two zeros would be ignored, as there are no subjects numbered ‘0’. To choose a sample from 86 units, we would use pairs of digits. Moving horizontally from the same starting-point, we would select the units numbered 89 (ignored), 31, 62, etc. To choose a sample from between 100 and 999 sampling units, we would use

Table 8.1 Random numbers

Row	Columns		
	1–4	5–8	9–12
1	96 22	74 70	80 46
2	82 14	73 36	41 54
3	21 47	59 93	48 40
4	89 31	62 79	45 73
5	63 29	90 61	86 39
6	71 68	93 94	08 72
7	05 06	96 63	58 24
8	06 32	57 11	81 59
9	91 15	38 54	73 30
10	54 60	28 35	32 94

sets of three digits (893, 162, 794, 573, 632, and so on). With between 1000 and 9999 sampling units, we would use sets of four digits (8931, 6279, etc.).

Sometimes, many numbers have to be discarded and the process may become very tedious. For example, with 195 units to choose from, if we started from the same '8' in row 4 and moved horizontally we would find only two helpful numbers among the first 16 we looked at: 162 in row 4 and 050 (or 50) in row 7. In such instances, short-cut methods may be used.<sup>22</sup>

## Notes and References

1. A postal survey of 110 sword swallowers (members or contacts of the Sword Swallowers' Association International) revealed a history of definite or probable perforation in 22% of the respondents (10 of 46) (Witcombe B, Meyer D. Sword swallowing and its side effects. *British Medical Journal* 2006; 333: 1285). The percentage with perforations in the total sample might conceivably be as high as 67% (74 of 110) if all the nonrespondents were perforated, or as low as 9% (10 of 110) if all the nonrespondents were unperforated.

Note that the sample could not include mortally wounded sword swallowers – an example of *prevalence-incidence bias* (see p. 92). Moreover, it is unlikely that all sword swallowers are members or contacts of the Sword Swallowers' Association International (which can be joined by sending a photograph of oneself swallowing a sword and completing a membership application at the association's website). Furthermore, we do not know how valid the reports of perforation and nonperforation were (see *sensitivity* and *specificity*, p. 167). Add to this is the uncertainty related to the small size (46) of the effective sample – the 95% confidence intervals (see p. 260) of the above extreme estimates are 58–76% and 4–16% respectively – and we can conclude with certainty that we can only guess exactly how dangerous this pastime is. But some sensitive souls may feel that a risk even as low as 4% is enough to put them off trying.

2. Faugier J, Sargeant M. Sampling hard to reach populations. *Journal of Advanced Nursing* 1997; 26: 790.

It is claimed that there are statistical manoeuvres that can derive valid population estimates from snowball samples of hidden populations, e.g. of injection drug users or (surprise!) jazz musicians (Heckathorn DD. Respondent-driven sampling II. Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems* 2002; 49: 11. Salganick MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology* 2004; 34: 193).

3. Specimen instructions for interviewers: 'Ask if any children aged under 15 years live in the home. If 'yes', carry on with the interview if there is an "A" in the sealed envelope'; this requires a prior allocation of the required proportion of As, in accordance with the sampling fraction; the envelopes should be well shuffled.
4. See Smith W, Mitchell P, Attebo K, Leeder S (Selection bias from sampling frames: telephone directory and electoral roll compared with door-to-door population census: results from the Blue Mountains Eye Study. *Australian and New Zealand Journal of Public Health* 1997; 21: 127).

A New York study that used driver's licence files as a sampling frame for the selection of controls found differences (e.g. in age, income and alcohol consumption) between cases of breast cancer with and without licences (Bowlin SJ, Leske MC, Varma A, Nasca P, Wienstein JA, Caplan L. Breast cancer risk and alcohol consumption: results from a large case-control study. *International Journal of Epidemiology* 1997; 26: 915).

5. For a detailed exposition of the statistical aspects of sampling and the handling of sample data, see Cochran WG (*Sampling techniques*, 3rd edn. New York, NY: Wiley; 1977).

6. *Random digit dialling* is usually done by the procedure described by Waksberg J (Sampling methods for random digit dialling. *Journal of the American Statistical Association* 1978; 73: 40).

In a study in Washington requiring blood tests, potential subjects were chosen by random digit dialling. The response rate was 83% in this phase, 81% in the next phase (a telephone interview) and 67% in the third phase (blood-taking). The overall rate was thus  $83\% \times 81\% \times 67\%$ , or only 45%, illustrating the effect of offering repeated opportunities for nonresponse (Brown IM, Tollerud DJ, Pottern LM, Clark JW, Kase R, Blattner WA, Hoover RN. *Biochemical epidemiology in community-based studies: practical lessons from a study of T-cell subsets. Journal of Clinical Epidemiology* 1989; 42: 561).

7. Kempf AM, Remington PL. New challenges for telephone survey research in the twenty-first century. *Annual Review of Public Health* 2007; 28: 113. Nathan G. Telesurvey methodologies for household surveys – a review and some thoughts for the future. *Survey Methodology* 2001; 27: 31. Bunin GR, Spector LG, Olshan AF, Robison LL, Roesler M, Grufferman S, Shu X, Ross JA. Secular trends in response rates for controls selected by random digit dialing in childhood cancer studies: a report from the Children's Oncology Group. *American Journal of Epidemiology* 2007; 166: 109.

A national study in the USA found that, in comparison with adults in households with landline telephones, adults who only had cell phones were more likely to smoke and have drinking binges, and less likely to receive influenza vaccine and to have medical insurance (Blumberg SJ, Luke JV, Cynamon MC. Telephone coverage and health survey estimates: evaluating the need for concern about wireless substitution. *American Journal of Public Health* 2006; 96: 926).

Daily text messaging was found to be a useful procedure in a survey that found that the frequency of mild hypoglycaemia in young diabetics was three times higher than previously recognized (Tasker APB, Gibson L, Franklin V, Gregor P, Greene S. *Pediatric Diabetes* 2007; 8: 15).

8. One research institute used the following operational definitions:

'A *household unit* is a room or group of rooms occupied or vacant and intended for occupancy as separate living quarters. In practice, living quarters are considered separate and therefore a housing unit when the occupants live and eat apart from any other group in the building, and there is either direct access from the outside or through a common hall, or complete kitchen facilities for the exclusive use of the occupants, regardless of whether or not they are used'. (The definition then goes on to explain what is meant by 'living apart', 'eating apart', 'direct access', etc.) A household is everyone who resides in a housing unit at the time the interviewer speaks to a household member and learns who lives there, including those who have places of residence both there and elsewhere. The household also includes people absent at the time of contact, if a place of residence is held for them in the housing unit and 'no place of residence is held for them elsewhere'.

'A *family unit* consists of household members who are related to each other by blood, marriage, or adoption. A person unrelated to other occupants in the housing unit – or living alone – constitutes a family unit with only one member'. If there is more than one family unit in the household, the 'primary family unit' is the one that owns or rents the home. 'If families share ownership or rent equally, the one whose head is closest to age 45 is usually considered to be the primary family'. (Survey Research Center, Institute for Social Research. Interviewer's manual. Ann Arbor, MI: University of Michigan; 1976. pp. 39, 91, 94).

9. For software, see Appendix C.
10. If there are many primary units (e.g. households) it is easier to divide them into strata according to their size, and use a different sampling fraction for each stratum, the sampling fractions being proportional to the size of the units. If a single member of each selected household is required, use may be made of a simple method described by Cochran WG (1977, pp. 364–365; see note 5).

11. Sackett DL. Bias in analytic research. *Journal of Chronic Diseases* 1979; 32: 51.
12. Moses LE. Statistical concepts fundamental to investigations. *New England Journal of Medicine* 1985; 14: 890.
13. The *finite population correction*, the factor introduced into the calculation to allow for the effect of the size of the parent population, is one minus the sampling fraction. If the sampling fraction is low then this factor is close to unity, and the correction has a negligible influence and may be omitted (Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*, 4th edn. Blackwell Science; 2002. pp. 95–96).
14. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. *Journal of Clinical Epidemiology* 2001; 54: 109. Knottnerus JA, Bouter LM. The ethics of sample size: the whole picture should be considered. *Journal of Clinical Epidemiology* 2000; 56: 207.
15. Moyé LA, Tita TN. Hypothesis testing complexity in the name of ethics: response to commentary. *Journal of Clinical Epidemiology* 2002; 55: 209. Schouten HJA. The ethics of sample size: reaction to commentary. *Journal of Clinical Epidemiology* 2003; 56: 206.
16. *Calculations of sample size*. To estimate a proportion from a simple random sample, the required sample size is

$$z^2 p (1 - p) / d^2$$

where  $z = 1.96$  for 95% confidence, 1.645 for 90% confidence,  $p$  is the estimated proportion in the study population, and  $d$  is the acceptable margin of error.

If the finite population correction is used, the required sample size is

$$N z^2 p (1 - p) / [d^2 (N - 1) + z^2 p (1 - p)]$$

where  $N$  is the size of the study population.

For other sample size formulae, refer to a statistics text (see note 2, Chapter 26).

Formulae for use with a wide variety of statistical tests are given by Lachin JM (*Introduction to sample size determination and power analysis for clinical trials*. Controlled Clinical Trials 1981; 2: 93).

For *tables* showing sample sizes for a comparison of two proportions, see Fleiss JL, Levin B, Paik MC (*Statistical methods for rates and proportions*, 3rd edn. Wiley; 2003. Table A.4) or Schlesselman JJ (*Case-control studies: design, conduct, analysis*. New York, NY: Oxford University Press; 1982. Appendix A). For a *nomogram*, see Altman DG (*Practical statistics for medical research*. London: Chapman & Hall; 1991. p. 486).

For computer programs, see Appendix C.

*Compensating for losses*. Allowance can be made for an expected loss of  $R\%$  of the study sample (due to nonresponse, dropouts, etc.) – but of course without compensating for possible selection bias – by multiplying the computed sample size by  $(100 - R)^2 / 10,000$  (Lachin JM (1981; see above)).

17. See p. 70.
18. Armitage P. *Sequential medical trials*, 2nd edn. Oxford: Blackwell; 1975.
19. The required size of a *cluster sample* (for software, see Appendix C) depends not only on the factors influencing the required size of a simple random sample, but also on the cluster size and the evenness or unevenness of the distribution of the disease or characteristic (does it occur more in some clusters than in others?).

When a cluster sample has been used in a study, it is customary to compute and report the *design effect*, which is the ratio of the required sizes for cluster and random samples. Design effects of 2 or more are not uncommon. The simplest way to estimate sample size for a cluster-sample survey is to calculate the required size of a simple random sample and then multiply this by the design effect reported in a previous cluster-sample survey of the same disease in a similar population, using a similar cluster size, or in a previous round of the same survey. If different studies yielded different design effects, it is prudent to use the highest value.

The larger the cluster size, the larger the design effect. If a design effect  $D_1$  is based on a cluster size  $b_1$  and you wish to estimate the sample size required for a cluster size  $b_2$ , the required design effect  $D_2$  is approximately

$$D_2 = 1 + (D_1 - 1)(b_2 - 1)/(b_1 - 1)$$

Methods of estimating the design effect are described by Bennett S, Woods T, Liyanage WM, Smith DL. (A simplified general method for cluster-sample surveys of health in developing countries. *World Health Statistics Quarterly* 1991; 44: 98) and Katz J, Zeger SL (Estimation of design effects in cluster surveys. *Annals of Epidemiology* 1994; 4: 295).

For *comparisons of cluster samples*, see Kerry SM, Bland JM (Sample size in cluster randomization. *British Medical Journal* 1998; 316: 549).

For *stratified cluster randomization*, see Donner A (Sample size requirements for stratified cluster randomization designs. *Statistics in Medicine* 1992; 11: 743).

20. *Post hoc* calculations of power are often performed to explain nonsignificant test results; but this procedure is logically flawed, and reliance should rather be placed on an appraisal of confidence intervals (Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 2001; 55: 19). 'We should adopt confidence intervals for effect sizes more widely, to encourage us to think more about the range of effect sizes that are supported by the data and those that are not and think less about  $p$  values' (Colegrave N, Ruxton GD. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology* 2003; 14: 446).
21. Computer programs that generate random numbers (see Appendix C) actually produce *pseudorandom numbers*, generally using algorithms whose capacity to produce sequences of numbers that are to all intents and purposes random have been thoroughly tested.
22. *Shortcuts* can be taken when using a table of random numbers to choose a sample. For example, if there are between 101 and 200 sampling units to choose from, read the successive three-digit numbers, and subtract the largest possible multiple of 200 from every number above 200 (also, read 000 as 200). Using the example in the text (p. 86), the sampling units selected would then be 93 (893 - 800), 162, 194 (794 - 600), 173 (573 - 400), 32, 190, 18, 39, etc. If there are between 201 and 300 sampling units, subtract a multiple of 300 from numbers above 300, discarding numbers above 900. If there are between 301 and 400 sampling units, subtract 400 from numbers above 400, discarding numbers above 800. And if there are between 401 and 500, subtract 500 from numbers above 500 (take 000 as 500).