2

SAMPLING

OBJECTIVES

After reading this chapter, you should be able to:

1. Select a random, simple, systematic, stratified, cluster, multistage sample, or targeted (risk based) sample—given the necessary elements.

2. Recognise the advantages and disadvantages of each sampling method.

3. Select the appropriate sampling strategy for a particular situation, taking into account the requirements, advantages, and disadvantages of each method.

4. List the elements that determine the sample size required to achieve a particular objective, and explain the effect of each upon the sample-size determination.

5. Compute required sample sizes for common analytic objectives.

6. Understand the implications of complex sampling plans on analytic procedures.

7. Select a sample appropriately to detect or rule out the presence of disease in a group of individuals.

2.1 INTRODUCTION

2.1.1 Census vs sample

For the purposes of this chapter, we will assume that data are required for all individuals, or a subset thereof, in a population. The process of obtaining the data will be referred to as measurement.

In a census, every individual in the population is evaluated. In a sample, data are collected from only a subset of the population. Taking measurements or collecting data on a sample of the population is more convenient than collecting data on the entire population. In a census, the only source of error is the measurement itself. However, even a census can be viewed as a sample, because it represents the state of the population at one point in time, and hence is a sample of possible states of the population over time. With a sample, you have both measurement and sampling error to contend with. A well-planned sample, however, can provide virtually the same information as a census, at a fraction of the cost.

Note The outcome in any study (*eg* disease status) is often determined by the use of diagnostic tests (see Chapter 5). For the sake of simplicity, in this chapter we will assume that the outcome is measured without error.

2.1.2 Descriptive versus analytic studies

Samples are drawn to support both descriptive studies (often called surveys) and analytic studies (often called observational studies).

A **descriptive study** (survey) aims to describe population attributes (frequency of disease, prevalence of an exposure). Surveys answer questions such as, 'What proportion of people in a defined population had diarrhea over a 1-month period?' or 'What is the average body mass index (BMI) of students in Grade 12?'

An **analytic study** is done to estimate the magnitude of an association between outcomes and exposure factors in the population. Analytic studies contrast groups and seek explanations for the differences between them. An analytic study might ask a question such as, 'Is water source associated with the incidence of diarrhea?' or, 'How does time spent playing video games affect the BMI of Grade 12 students?' Establishing an association is the first step to inferring causation, as was discussed in Chapter 1.

The distinction between descriptive and analytic studies is discussed further in Chapter 7.

2.1.3 Hierarchy of populations

There is considerable variation in the terminology used to describe various populations in a study. In this text, we will adopt terminology consistent with that used in the text *Modern Epidemiology* (Rothman *et al*, 2008) with 3 populations of interest: the target population, the source population, and the study sample or group. These will be discussed with reference to the dataset used for examples in this chapter—a study designed to evaluate the impact of water cisterns on the incidence of diarrhea in the semi-arid region of Pernambuco State in Brazil.

The **target population** is the population to which it might be possible to extrapolate results from a study. It is often not clearly defined and might vary depending on the perspective of the individual interpreting the results of the study. For example, the investigators conducting the study referred to above might be interested in the impact of rainwater cisterns in Pernambuco State, while someone from outside the state might want to extrapolate the findings to all semi-arid regions of Brazil.

The **source population** is the population from which the study subjects are drawn. Conceptually, all units in the source population should be 'listable' and have a non-zero probability of being included in the study. For example, in the Brazil diarrhea study, families included in the study were drawn from the list of households participating in the One Million Cisterns Project (OMCP). An alternative strategy would be to randomly select communities within the study region, and then sample households within selected communities (assuming a list of households exists).

The **study sample** (or group) consists of the individuals (or groups of individuals) that end up in the study. Usually this group is some form of sample from the source population. Prior to conducting the study, the researchers would determine the necessary sample size (perhaps planning to sample only some of the households and some of the individuals within each household). Data would be collected from eligible study subjects and the study sample would consist of the households and individuals who agreed to participate (and whose data were adequate for inclusion in the study). These individuals are referred to as a sample or group of individuals rather than a population because they do not constitute an easily defined population.

The concept of validity is discussed at length in Chapters 12 and 13, but validity relates to the populations in the following ways. The **internal validity** of a study relates to whether or not the study results (obtained from the study sample) are valid for members of the source population. Essentially, this indicates whether or not the study has obtained the 'correct' answer for the source population. Much of this book is dedicated to methods used to ensure the correct answer is obtained.

The **external validity** relates to how well those results can be generalised to the target population. Evaluation of external validity involves a subjective assessment of whether or not the source population is broadly representative of the target population. Given that the target population may be defined differently by different readers, assessment of external validity is much more difficult. However, it is easier to generalise the results from an analytical study (one which evaluates associations) than results from a descriptive study (which describes the level of a disease or other characteristics in a population). For example, the monthly incidence of diarrhea (a descriptive result) may be different in Pernambuco State than in other states in Brazil. However, an observed association between water source and the risk of diarrhea (an analytic result) is more likely to be generalisable.

2.1.4 Sampling frame

The **sampling frame** is defined as the list of all the **sampling units** in the source population. Sampling units are the basic elements of the population that is sampled (*eg* households, individuals). A complete list of all sampling units is required in order to draw a simple random sample, but it might not be necessary for some other sampling strategies. The sampling frame is the information about the source population that enables you to draw a sample. In our example, a suitable sampling frame would be the list of all households eligible for participation in the

OMCP. Once households were selected, we would devise a strategy for selecting individuals within those households (if we decided to sub-sample within households).

2.1.5 Objectives of the study

The objectives of a study will influence the sampling strategy employed. Descriptive studies are usually aimed at determining the prevalence (or incidence) of disease in a population or demonstrating that a population is free of disease. Analytical studies are focused on establishing associations between factors (*eg* risk factors) and an outcome (*eg* disease). Unless otherwise specified, this chapter will focus on sampling to support prevalence estimation or analytical objectives. The issue of sampling to detect the presence of disease (or alternatively to declare a population free of disease) will be discussed in Section 2.12.

2.1.6 Types of error

In a study based on a sample of observations, the variability of the outcome being measured, measurement error, and sample-to-sample variability all affect the results we obtain. Hence, when we make inferences based on the sample data, they are subject to error. Within the context of hypothesis testing in an analytical study, there are 2 types of error:

Type I (α) error: You conclude that the outcomes in the groups being compared are different (*ie* an association exists), when in fact they are not.

Type II (β) error: You conclude that the outcomes are not different (*ie* no association between the exposure and outcome exists), when in fact they are.

A study was carried out to determine if an exposure had an effect on the probability of disease occurrence or not. Table 2.1 presents the possible decisions that can be made based on the study and their relationship to the 'truth.'

Table 2.1 Types of error

		True state of nature	
		Effect present	**Effect absent**
Conclusion of statistical analysis	Effect present (reject null hypothesis)	Correct	Type I (α) error
	No effect (accept null hypothesis)	Type II (β) error	Correct

Statistical test results reported in the medical literature are aimed at disproving the **null hypothesis** (which is that there is no difference between groups). If differences are found, they are reported with a P-value which expresses the probability that a difference as large (or larger) than the one observed could be due to chance, if the null hypothesis is true. P is the probability of making a Type I (α) error. When $P \leq 0.05$, we are 'reasonably' sure that any effect detected is not due to chance.

Power is the probability that you will find a statistically significant difference when it exists and is of a certain magnitude (*ie* power=1-β). The probability of making a Type II (β) error, or failing to detect a difference, is sometimes not stated because of the general preference for reporting positive results in the literature. So-called **negative findings** (failure to find a difference) are less likely to be reported. There are a number of reasons why a study might find

no effect of the factor being investigated.

- There truly was no effect of exposure on the outcome.
- The study design was inappropriate.
- The sample size was too small (low power).
- Bad luck.

An evaluation of the power of the study will at least determine how likely you are to commit a Type II error for a given alternative hypothesis.

2.2 NON-PROBABILITY SAMPLING

Samples that are drawn without an explicit method for determining an individual's probability of selection are known as **non-probability samples**. Whenever a sample is drawn without a formal process for random selection, it should be considered a non-probability sample, of which there are 3 types: judgement, convenience, and purposive. Non-probability samples are inappropriate for descriptive studies except in the instance of initial pilot studies (even then, use of non-probability samples might be misleading). However, non-probability sampling procedures are often used in analytical studies.

2.2.1 Judgement sample

This type of sample is chosen because, in the judgement of the investigator, it is 'representative' of the source population. This is almost impossible to justify because the criteria for inclusion and for the process of selection are largely implicit, not explicit.

2.2.2 Convenience sample

A convenience sample is chosen because it is easy to obtain. For instance, households in close proximity to a research centre might be selected for study. Convenience sampling often is used in analytical studies where the need to have a study group that is representative of the source population can be relaxed. Convenience sampling was used partially in the Brazil diarrhea study to overcome some limitations with the sampling frame. Once households with a rainwater cistern had been selected in a community, households without a cistern were selected by choosing the closest household that was eligible for the OMCP but which did not yet have a cistern installed.

2.2.3 Purposive sample

The selection of this type of sample is based on the study subjects possessing one or more attributes such as known exposure to a risk factor or a specific disease status. This approach is often used in observational analytical studies. If a random sample is drawn from all sampling units meeting the study criteria, then it becomes a probability sample from the subset of the source population.

2.3 PROBABILITY SAMPLING

A **probability sample** is one in which **every** element in the population has a known **non-zero** probability of being included in the sample. This approach implies that a formal process of

random selection has been applied to the sampling frame. The following sections will describe how to draw different types of probability sample. Procedures for analysing data derived from the samples will be discussed in Section 2.10. A much more complete description of sampling procedures can be found in general sampling texts such as Levy and Lemeshow (1999) and Heeringa *et al* (2010).

2.4 SIMPLE RANDOM SAMPLE

In a **simple random sample**, every study subject in the source population has an equal probability of being included. A complete list of the source population is required and a formal random process is used (random is **not** the same as haphazard). Random sampling can be based on drawing numbers from a hat, using computer-generated random numbers, using a random-numbers table, flipping a coin, or throwing dice.

For example, suppose you are investigating the influence of season on the length of wait times in a hospital emergency room and you estimate that you need to review 1,000 records to obtain the data you need. If all of the data already exist in hospital records (*eg* time of arrival, time of examination), then you can randomly select 1,000 records from the past year's admissions for detailed review. If there have been 13,000 admissions over the past year, you would randomly generate 1,000 numbers between 1 and 13,000. These numbers would identify the records that you would pull for review.

2.5 SYSTEMATIC RANDOM SAMPLE

In a **systematic random sample**, a complete list of the population to be sampled is not required provided an estimate of the total number of individuals is available and all of the individuals (or their records) are sequentially available (*eg* people entering a hospital emergency room). The **sampling interval** (j) is computed as the study population size divided by the required sample size. The first study subject is chosen randomly from among the first j study subjects, then every j^{th} study subject after that is included in the sample. It is a practical way to select a probability sample if the population is accessible in some order, but bias might be introduced if the factor you are studying is related to the sampling interval. Consequently, a simple random sample would be preferable, but might not be feasible if the logistics of the data collection (*eg* time required by nursing staff for additional data collection) precludes the use of a simple random sample which might generate a series of consecutive numbers.

Assume once again that you want a sample of 1,000 patients attending a hospital emergency room. Given that you know there will be approximately 13,000 patients over a 1-year period, you need to sample every 13^{th} patient to achieve the desired sample size. To start, randomly pick a number between 1 and 13 to choose your first patient, and then have the required data collected on every 13^{th} patient after that. Data from a systematic random sample are analysed as though they were derived from a simple random sample.

2.6 STRATIFIED RANDOM SAMPLE

In this approach, prior to sampling, the population is divided into mutually exclusive strata based on factors likely to affect the outcome. Then, within each stratum, a simple or systematic random sample is chosen. The simplest form of stratified random sampling is called

proportional (the number sampled within each stratum is proportional to the total number in the stratum). There are 3 advantages of **stratified random sampling**.

1. It ensures that all strata are represented in the sample.
2. The precision of overall estimates might be greater than those derived from a simple random sample. The gain in precision results from the fact that the between-strata variation is explicitly removed from the overall estimate of variance.
3. It produces estimates of stratum-specific outcomes, although the precision of these estimates will be lower than the precision of the overall estimate.

For example, assume you believe that wait times are different for males and females. You could stratify the existing records on the basis of sex, and then randomly sample within each sex (either equal numbers of males and females or proportional to the distribution of the sexes in the whole set of records).

2.7 CLUSTER SAMPLING

A **cluster** is a natural or convenient collection of study subjects with one or more characteristics in common. For example:

- a household is a cluster of people
- a city block is a cluster of households
- a clinic is a cluster of patients
- a day is a cluster of emergency room visits
- a classroom is a cluster of students.

In a cluster sample, the **primary sampling unit** (PSU) is larger than the unit of concern. For example, if you wanted to estimate the proportion of Grade 12 students who smoke (in a small city), you could use a cluster sample in which you randomly selected Grade 12 classes, even though the unit of concern is the student. In a cluster sample, every study subject within the cluster is included in the sample (*ie* all students in the selected classes).

Cluster sampling is done because it might be easier to get a list of clusters (Grade 12 classes) than it would be to get a list of individuals (students), and it is often less expensive to sample a smaller number of clusters than to collect information from selected individuals within many different clusters.

In this example of cluster sampling, a survey to determine the proportion of students smoking was conducted. Of 47 Grade 12 classes in the city, 10 were randomly selected from a list provided by the school board and every student in each of the 10 classes was asked to complete the questionnaire. A cluster sample is convenient because it may be impossible to get a complete list of all Grade 12 students, but it is relatively easy to get a list of the Grade 12 classes. It is also more practical to sample all students in 10 classes than it is to visit all 47 classes and sample a few individuals in each class. Of course, students within a given class are probably more alike than students from different classes, so the sampling variation for a given number of individuals is greater than if they had been chosen by simple random sampling. The impact of sampling at the cluster level is discussed further in Sections 2.10.3 and 2.11.6.

When a group is not a cluster In cluster sampling, a group is a cluster of individuals. A sample is a cluster sample if the group is the sampling unit and the study subjects within the group are the unit of concern. When the group is both the sampling unit and the unit of concern, then by definition the sample is not a cluster sample. For example, the following is **not** a cluster

sample: a sample of households to determine whether or not anyone in the household smokes indoors (in this case, the household is the unit of concern, not the individual residents).

2.8 MULTISTAGE SAMPLING

A cluster might contain too many study subjects to obtain a measurement on each, or it might contain study subjects so nearly alike that measurement of only a few study subjects provides information on the entire cluster. **Multistage sampling** is similar to cluster sampling except that, after the PSUs (*eg* households) have been chosen, then a sample of **secondary sampling units** (*eg* individuals) is selected. Once again, assume that you are interested in the smoking habits of Grade 12 students, but you would like to use urine cotinine as a measure of smoking (or exposure to tobacco smoke). Given the difficulty of collecting the samples, and the cost of testing, you might only sample a small number of individuals in each class.

If you want to ensure that all individuals in the population have the same probability of being selected, 2 approaches are possible. First, the PSUs chosen might be selected with a probability proportional to their size. In other words, if the class size is known ahead of time, large classes should have a higher probability of being chosen than small ones. After the number of classes to be sampled is chosen, you select a fixed number of students in each class from which to get urine samples. If class size is not known ahead of time, take a simple random sample of the PSUs (classes) and then sample a constant proportion of the students in each class. Either approach will ensure each individual has the same probability of selection. If this is not the case, the probability of selection needs to be accounted for in the analysis (see Section 2.10.2).

How many classes and how many students to sample within each class depend upon the relative variation (in the factor(s) being measured) between classes, compared with within classes, and the relative cost of sampling classes compared with the cost of sampling individuals within classes. In other words, when the between-class variation is large relative to the within-class variation, you will have to sample many more classes to get a precise estimate. Multistage sampling is very flexible where cost of sampling is concerned. If you are like most researchers, you are working on a limited budget and, when it is expensive to visit and sample classes, you will want to sample as few as possible. On the other hand, if the cost of processing samples from an individual is high relative to the cost of visiting the class, you will want to sample fewer individuals per class. Usually researchers desire to have the most precise estimate of the outcome for the lowest possible cost. These 2 desires can be balanced by minimising the product of the variance and the cost. Regardless of the total sample size for the study (n), the variance*cost product can be minimised by selecting n_i individuals per group according to the following formula:

$$n_i = \sqrt{\frac{\sigma_i^2}{\sigma_g^2} * \frac{c_g}{c_i}} \qquad \textit{Eq 2.1}$$

where n_i is the number of individuals to be sampled per group, and σ^2_g and σ^2_i are the between- and within-group variance estimates, and c_g and c_i are the costs of sampling groups and individuals, respectively. The value for n_i needs to be rounded to an integer value and cannot be less than 1. Once the number of individuals per group has been determined, the number of households to be sampled is then $n_g = n/n_i$.

Keep in mind that cluster and multistage sampling almost always require more subjects for the same precision than simple random sampling. Example 2.1 describes a stratified, multistage sampling approach. Multistage sampling, as the name suggests, can be extended to more than the 2 levels discussed above.

2.9 TARGETED (RISK-BASED) SAMPLING

Although disease surveillance programs based on targeted sampling plans are more common in veterinary medicine (and much of the literature on target sampling relates to veterinary disease surveillance), they have applications in human medicine as well (Robinson *et al*, 2006). It involves the stratification of the source population into strata based on one, or more, characteristic(s) which are thought to be associated with the probability of disease occurrence. However, unlike stratified sampling, targeted sampling may involve sampling only from strata in which the probability of finding cases of disease is highest (Salman, 2003; Stärk *et al*, 2006), or at least weighting the sample heavily in favour of high risk strata. Consequently, some individuals may have a zero probability of being included in the sample. Methods for targeted sampling have recently been developed and are an active area of research.

In targeted sampling, individuals are assigned point values based on the probability of them having the disease of interest, and sampling is proportional to that estimate of risk (Thurmond, 2003). Sampling proceeds until individuals with the predetermined number of points have been sampled. Population inference from a targeted sample requires 2 key epidemiological parameters: an estimate of how the characteristic used to create the strata relates to the probability of disease (*ie* an estimate of the risk ratio (see Chapter 4) for the characteristic), and

Example 2.1 Multistage sampling
data = brazil_smpl

A study of the impact of rainwater cisterns on the incidence of diarrhea was conducted in a semi-arid region of the Agreste Central Region of Pernambuco State in Brazil. These data are a subset of the full dataset which is described in Chapter 31. The study had the following characteristics:

- The target population was all households in the region.
- The **source population** was households participating in (or eligible for participation in) the One Million Cisterns Project.
- The **sampling frame** consisted of a list of all households within each community with an installed cistern. (Non-cistern households were selected by field investigators identifying non-cistern households that met the eligibility criteria and which were in close proximity to cistern households.)
- Sampling was **stratified** by municipality, with cistern households within a municipality being identified first. (In fact, stratification often took place at the community level although it was not always possible to obtain the required sample within a community, so the data will be analysed with stratification at the municipality level.)
- In general, sampling was carried out as **cluster sampling** with the households being selected first and then all residents of the household being surveyed. In a small subset of households, not all individuals were surveyed and this was taken into consideration in the computation of sampling weights.
- The **study sample** consisted of the individuals selected for participation in the study for which data were recorded.

These data will be used in Examples 2.2 through 2.4.

an estimate of the distribution (frequency) of the characteristic in the source population (Williams *et al*, 2009a). The advantage of targeted sampling is that it will require a much smaller sample size than other forms of sampling if the outcome of interest (disease) is rare and characteristics that strongly influence the probability of an individual having the outcome can be identified. A disadvantage is that key epidemiological parameters might not be known. Specifically, the effect of the characteristic of interest (*ie* the risk ratio) is often not known for the population being studied and must be derived from evidence in other populations. In addition, the proportion of individuals with the characteristic of interest also might not be known. Uncertainty in these 2 estimates should be taken into account when planning a targeted sampling program (Williams *et al*, 2009a). Poisson sampling is an unequal probability sampling strategy that can be used for targeted sampling programs (Williams *et al*, 2009b).

In veterinary medicine, targeted sampling has been used extensively in bovine spongiform encephalopathy (BSE) surveillance programs (Prattley *et al*, 2007a; 2007b). In this instance, sampling is focused on the following strata (also called 'streams'): cattle with clinical signs compatible with BSE, dead stock (cattle that die on the farm), and casualty slaughter (unhealthy cattle slaughtered at the slaughter house). A simulation study used to evaluate the performance of targeted sampling for disease prevalence estimation concluded that targeted sampling is appropriate provided justifiable estimates of the key epidemiological parameters are available (Wells *et al*, 2009).

2.10 ANALYSIS OF SURVEY DATA

The sampling plan needs to be taken into account when analysing data from any research project involving a complex sampling plan. (**Note** Although referred to as 'survey' data, the concepts discussed in this chapter apply equally to the analysis of data from analytic studies based on complex sampling plans.) There are 3 important concepts that have been raised in the above discussion of various sampling plans: stratification, sampling weights, and clustering. In addition to these, the possibility of adjusting estimates derived from finite populations must be considered.

2.10.1 Stratification

If the population sampled is divided into strata prior to sampling, then this needs to be accounted for in the analysis. For example, in the Brazil diarrhea study, the population was initially stratified into municipalities with approximately equal numbers of cistern and non-cistern households being selected in each municipality. The advantage of such stratification is that it provides separate stratum-specific estimates of the outcome of interest. If the factor upon which the population is stratified is related to the outcome (*eg* incidence of diarrhea in the strata), then the standard error (SE) of the overall prevalence estimate might also be lower than if a non-stratified sample was taken. Correctly accounting for the stratified nature of the sample requires that the total population size in each stratum be known in order to compute the appropriate sampling weights (Section 2.10.2).

In Example 2.2, the Brazil diarrhea data have been analysed ignoring the stratification by municipality, and then by taking it into account.

Example 2.2 Analysis of stratified survey data
data = brazil_smpl

Data on diarrhea were obtained from 3,399 individuals in 21 municipalities. A simple estimate (treating the sample as a simple random sample) of the overall monthly incidence risk was 0.1462 (14.62%) and the SE of that estimate was 0.0061 (0.61%).

If the data are stratified by municipality, some of the incidence estimates are as follows.

		Monthly incidence risk	
Municipality	**Number of samples**	**Incidence**	**SE (incidence)**
1	137	0.1314	0.0290
2	145	0.2276	0.0349
3	231	0.2987	0.0302
... some estimates omitted			
19	160	0.2375	0.0337
20	188	0.1277	0.0244
21	19	0.5263	0.1177
Overall	3399	0.1462	0.0059

There are considerable differences across the municipalities in terms of the incidence of diarrhea (the range of estimates was from 0.0152 to 0.5263). The SE of the overall estimate from the stratified sample is slightly smaller than when the data were treated as a simple random sample, but the difference is minimal. Stratification alone does not change the overall point estimate of the prevalence.
Note This analysis is provided for pedagogical purposes only. It would not be correct to assume equal sampling weights (Section 2.10.2) given that non-proportional sampling was carried out across strata.

2.10.2 Sampling weights

Although probability sampling requires a formal random process to be used to select the sample, it does not imply that all units sampled have the same probability of selection. If a sample of households is selected from a source population, and a sample of individuals is selected within each of those households, then the probability of selection for any given individual can be computed as:

$$\text{p(selection)} = \frac{n}{N} * \frac{m}{M} \qquad \textit{Eq 2.2}$$

where n is the number of households in the sample, N is the number of households in the source population, m is the number of individuals that were selected from the sampled household, and M is the number of people in that household. For example, assume that 10 households are selected out of 300 in a municipality, and that in each household 2 individuals are sampled. If household A is a 5-person household, the probability that a person in that household will ultimately end up in the sample is:

$$10/300 * 2/5 = 0.013 (1.3\%)$$

Similarly, if household B is an 8-person household, the probability that an individual in that household will be in the sample is:

$$10/300 * 2/8 = 0.0083 (0.83\%)$$

These different probabilities of selection need to be taken into account in order to obtain the correct point estimate of the parameter of interest.

The most common way of forming sampling weights is to make them equal to the inverse of the probability of being sampled. This value reflects the number of individuals that each of the sampled individuals represents. For example, a person in household A would actually represent 1/0.013=77 people in total. A person in household B would have a sampling weight of 1/0.0083=120 because s/he had a smaller probability of selection.

In Example 2.3, the overall incidence risk of diarrhea has been computed taking sampling weights into consideration.

Example 2.3 Analysis of stratified and weighted survey data
data = brazil_smpl

Individuals within the study population had different probabilities of being selected for the sample. Two factors influenced this:

1) the probability that the entire household would be selected.

2) the probability that an individual within the household would be selected. (Although, in general, although the probability was 1) (*ie* all individuals sampled), this was not always the case.)

Household selection probability Within each municipality, the selection probability of a cistern household being selected was computed as the number of cistern households sampled divided by the number installed according to the OMCP. Similarly, the selection probability of non-cistern households was computed. For example, in municipality 1, there were 73 households sampled out of a total of 242 that had cisterns installed. Consequently, the probability of this household being selected was 73/242=0.3017 (30.17%).

Individual selection probability Within each household, the probability of a person being included in the analysis was the total number of people sampled within the household divided by the total number of residents of the household. For example, in household (-family-) 195 (municipality 1, community 1), there were 6 residents of the household but the investigators were only able to obtain data for 3 of them. Consequently, the probability of selection for these individuals was 0.5 (50%).

Overall selection probability The overall selection probability for an individual in household 195 in community 1 was the product of the above 2 probabilities: 0.3017*0.5=0.1508 (15.08%).

Sampling weights The sampling weight applied to the individual above was the inverse of the overall selection probability: 1/0.1508=6.63. Effectively, the results from this individual were considered to represent almost 7 people in the source population.

Taking the sampling weights into consideration, the overall estimate of the incidence of diarrhea was 0.1751 (17.51%), with an SE of .0091 (0.91%). Incorporating weights into the analysis changed the point estimate of the prevalence and also increased the SE.

2.10.3 Clustering

Cluster sampling and multistage sampling involve the sampling of individuals within groups. Individuals within groups are usually more alike (with regard to the outcome being measured) than individuals chosen randomly from the population. From a statistical perspective, this means that these observations are no longer independent and this lack of independence must be taken into account in the analysis. Failure to do so will almost always result in estimated SEs that are smaller than they should be.

Clustering may occur at multiple levels. For example, people may be clustered within a household which in turn may be clustered within a neighbourhood. In Chapters 20–22, we discuss techniques for evaluating the degree of clustering at each of the possible levels. However, when analysing survey data, one often wants to simply deal with the clustering as a nuisance factor in order to obtain correct estimates of the SEs of the parameters being estimated. The simplest and most common approach is to identify the PSU (*eg* household) and use this to adjust the estimates for all clustering effects at levels at, or below, this level (*eg* clustering of individuals within households).

Computation of the appropriate variance estimates in the presence of clustering and other elements of the survey design is not straightforward and requires specialised software. While the details of the procedure are beyond this text, the most common technique is **variance linearisation** (Dargartz and Hill, 1996; Kreuter and Vallian, 2007). It has the advantage that analytical solutions for SEs for most statistics computed from survey data (*eg* proportions, means) are available. However, the procedure requires a large number of PSUs to be reliable. Variance linearisation is the approach used in Example 2.4, in which the overall incidence of diarrhea has been estimated taking the within-household clustering into account (households were the PSUs and individuals were sampled within households). **Note** Survey design can be incorporated not only into the estimation of descriptive characteristics (*eg* incidence in Example 2.4), but also into many of the regression models described in later chapters of the book. Example 20.2 gives an example of the use of these procedures to account for clustering in a regression analysis.

2.10.4 Design effect

The overall effect of the sampling plan on the precision of the estimates obtained can be expressed as the **design effect** (referred to as **deff**). The deff is the ratio of variance obtained from taking the sampling plan (*eg* stratification and clustering) into account to the variance that would have been obtained if a comparable-sized, simple random sample had been drawn from the population. A deff >1 reflects the fact that the sampling plan is producing less precise (larger variance) estimates than a simple random sample would have. (Of course, a simple random sample is often impossible to obtain.) The deff of the sampling plan computed in Brazil diarrhea study is also presented in Example 2.4. If an independent estimate of the deff is available, it can be incorporated into methods to account for clustering in the analysis of survey data (see Section 20.5.5).

2.10.5 Finite population correction

In most surveys, sampling is carried out **without replacement**. That is, once a study subject has been sampled, it is not put back into the population and potentially sampled again. If the

Example 2.4 Analysis of clustered survey data
data = brazil_smpl

The Brazil diarrhea data were derived from a **cluster** sample with households being the **primary sampling unit**. If the clustered nature of the sample was taken into account (in addition to the stratification and sampling weights), the overall incidence estimate remains at 0.1751 (17.51%) but the SE increases to 0.0128 (1.28%). (Clustering was accounted for using a variance linearisation approach to computing the SE.)

A summary of the estimates of the overall incidence taking various features of the sampling plan into account is shown below.

	Incidence	
Type of analysis	**Estimate**	**SE**
Assuming it was a simple random sample	0.1462	0.0061
Taking stratification into account	0.1462	0.0059
Taking stratification and sampling weights into account	0.1751	0.0091
Taking clustering into account	0.1462	0.0088
Taking stratification, sampling weights and clustering into account	0.1751	0.0128

The last row contains the most appropriate estimates for the incidence (and SE) of diarrhea. The design effect from this analysis was $16.265*10^{-5}/3.674*10^{-5}=4.43$ which indicates that taking the sampling plan into consideration produces an estimate of the variance of the incidence which is 4.43 times larger than the estimate would have been if a simple random sample of the same size (n=3,399) had been drawn.

proportion of the population sampled is relatively high (*eg* >10%), then this could substantially increase the precision of the estimate over what would be expected from an 'infinite-sized' population. Consequently, the estimated variance of the parameter being estimated can be adjusted downward by a **finite population correction** (*FPC*) factor of:

$$FPC = \frac{N-n}{N-1} \qquad \textit{Eq 2.3}$$

where N is the size of the population and n is the size of the sample. (**Note** An *FPC* should not be applied in cases where multistage sampling is carried out, even if the number of PSUs sampled is >10% of the population.) A finite population correction can also be used when estimating a sample size (see Section 2.11.5).

2.11 SAMPLE-SIZE DETERMINATION

The choice of sample size involves both statistical and non-statistical considerations. Non-statistical considerations include the availability of resources such as time, money, sampling frames, and some consideration of the objectives of the study. Interestingly, cost can be factored into sample-size calculations, and the greater the cost per sampled study subject, the smaller the sample size when the budget is fixed.

Statistical considerations include the required precision of the estimate, the variance expected in the outcome of interest, the desired level of confidence that the estimate obtained from

sampling is close to the true population value (1-α), and in analytic studies, the power (1-β) of the study to detect real effects.

2.11.1 Precision of the estimate

Whether you want to determine the monthly incidence risk of diarrhea or to estimate the body mass index of Grade 12 students, you must determine how precise an estimate you want. The more precise you wish to be, the larger the sample size you will require. If you want to know how many people had diarrhea within ±5%, you will have to sample more people than if you were only interested in obtaining an estimate within ±10%. Likewise, if you wanted your estimate of the BMI of Grade 12 males to be within 1 unit, you would need to collect data from more people than if you only needed to be within 3 units of the true population mean.

2.11.2 Expected variation in the data

The natural variation inherent in the data must be taken into account when calculating sample size. The variance of a simple proportion is $p*q$, where p is the proportion of interest and q is (1-p). Consequently, to estimate the sample size necessary to determine a proportion, then (paradoxical as it might seem) you must have a general idea of the proportion (with the outcome of interest) that you expect to find.

The measure of variation used for the estimation of the required sample size of a continuous variable such as BMI is the population variance (σ^2). We often don't know what the standard deviation (σ) is, but we can estimate it. One way to do this is to estimate the range that would encompass 95% of the values, and then assume that range is equal to 4σ. For example, if you think that 95% of people have a BMI between 18 and 42, then a rough estimate of the σ would be (42-18)/4=6 units, and the variance would be 36 units^2. (This is based on an assumption that the data are normally distributed, which may not be true, but this approach still provides a rough estimate of σ.)

2.11.3 Level of confidence

In descriptive studies, we must decide how sure we want to be that the **confidence interval** (CI) for your estimate will include the true population value. Similarly, in analytical studies, we must decide on the certainty we want that any difference we observe between 2 sampled groups is real and not due to chance. This is referred to as **confidence** and it is most commonly set to 95% (assume a Type I (α) error rate of 5%).

2.11.4 Power

The **power** of a study is its ability to detect an effect (*eg* a difference between 2 groups) when a real difference of a defined magnitude exists. For example, if the real difference in BMI between people who play video games more than 10 hours a week (compared with <10 hours) is 3 units, then a study with a power of 80% would detect a difference of this magnitude (and declare it statistically significant) 80% of the time. To increase the power, it is necessary to increase the sample size. The Type II (β) error rate is 1-power.

Precision and power have been presented as 2 separate issues even though they arise from the same conceptual basis. Sample sizes can be computed using either approach, although they will produce different estimates. (See Section 2.11.8 for an expansion of this topic.)

2.11.5 Sample-size formulae

The formulae for sample size required to estimate a single parameter (proportion or mean), or to compare 2 proportions or means, are shown below the following definitions:

Z_α The value of Z_α required for confidence=1-α $Z_{0.05}$=1.96
Z_α is the (1- α/2) percentile of a standard normal distribution
Note This is the value for a 2-tailed test or 2-sided confidence interval

Z_β The value of $Z\beta$ required for power=1-β; for power 1- β, $Z_{0.2}$ =-0.84
$Z\beta$ is the (1-β) percentile of a standard normal distribution

L The precision of the estimate (also called the 'allowable error' or 'margin of error') equal to half the desired length of a confidence interval

p *a priori* estimate of the proportion
(p_1, p_2—estimates in the 2 groups in an analytic study)

q 1-p

σ^2 *a priori* estimate of the population variance

μ *a priori* estimate of the population mean
(μ_1, μ_2—if estimates are required for 2 groups)

n sample size

Estimating proportions or means (*n*=total sample size)
To estimate a sample proportion with a desired precision:

$$n = \frac{Z_\alpha^2 pq}{L^2} \qquad \textit{Eq 2.4}$$

To estimate a sample mean with a desired precision:

$$n = \frac{Z_\alpha^2 \sigma^2}{L^2} \qquad \textit{Eq 2.5}$$

Comparing proportions or means (*n*=sample size per group)
To compare 2 proportions:

$$n = \frac{\left[Z_\alpha \sqrt{(2\,pq)} - Z_\beta \sqrt{p_1 q_1 + p_2 q_2}\right]^2}{(p_1 - p_2)^2} \qquad \textit{Eq 2.6}$$

where p=(p_1+p_2)/2 and q=1-p

To compare 2 means:

$$n = 2\left[\frac{(Z_\alpha - Z_\beta)^2 \sigma^2}{(\mu_1 - \mu_2)^2}\right] \qquad \textit{Eq 2.7}$$

Note The formulae shown above are approximations and most software will compute sample sizes using more exact formulae. Particular caution should be exercised with these formulae if the resulting n is small. Example 2.5 shows the calculation of a sample size for a study comparing 2 proportions.

Sampling from a finite population
If you are sampling from a relatively small population, then the required sample size (n') can be adjusted downward using the following *FPC* formula:

$$n' = \frac{1}{1/n + 1/N} \qquad \textit{Eq 2.8}$$

where n=the original estimate of the required sample size in an infinite population and N=the size of the population.

It is useful to make this finite population adjustment when computing the sample size for a simple or stratified random sample if the sampling fraction exceeds 10%. It is only applied to descriptive studies, not to analytic studies or controlled trial sample size calculations.

2.11.6 Adjustment of sample size for clustering

In epidemiologic research, we often deal with clustered data (*eg* individuals clustered within households) with observations within the same cluster being more similar to each other with respect to the outcome than observations drawn randomly from the population. If our study is

Example 2.5 Sample size for comparing proportions
data = hypothetical

Assume that you want to determine if provision of a rainwater cistern reduces the monthly risk of diarrhea. For the rainwater cisterns to be worth the cost of installing, you would want it to reduce the risk from the current level of 15% to 10% of individuals in the population. You want to be 95% confident in your result and the study should have a power of 80% to detect the 5% reduction in risk.

$p_1 = 0.15 \qquad p_2 = 0.10 \qquad p = 0.125$

$q_1 = 0.85 \qquad q_2 = 0.90 \qquad q = 0.875$

$Z_\alpha = Z_{0.05} = 1.96 \qquad Z_\beta = Z_{0.80} = -0.84$

$$n = \frac{\left[1.96\sqrt{2 * 0.125 * 0.875} - (-0.84)\sqrt{(0.15 * 0.85) + (0.10 * 0.90)}\right]^2}{(0.15 - 0.10)^2}$$

$= 685$

Consequently, you would require 1,370 (685*2) individuals with 685 being in households with rainwater cisterns and 685 without. A sample size derived incorporating a continuity correction (see Fleiss *et al* (2003) for details) is 726 individuals per group.

taking place exclusively at the lower (individual) level, with the factor of interest distributed at the individual level independent of the household, and the outcome (*eg* diarrhea) is measured at the individual level, this clustering does not present a problem when computing the necessary sample size. Such a situation arises when conducting a controlled trial of a treatment that is randomly assigned to individuals within households (ensuring that treatment allocation is independent of household). (See Chapter 20 for a more complete discussion of this situation.)

However, if the factor of interest is something that occurs at the household level (*eg* presence/absence of rainwater cistern), then the number of households in the study becomes a more critical concern than the number of individuals (even though the outcome is measured at the person level). The total sample size will need to be increased with the magnitude of the increase depending on:

1. the degree to which observations within a household are similar (measured by a parameter called the intra-cluster (or intra-class) correlation coefficient) (Section 20.3.3) and,
2. the number of people sampled per household (having many people sampled within a household is of little value if the individuals within a household are very similar). The formula for adjusting the sample size is:

$$n' = n(1 + \rho(m-1)) \qquad \textit{Eq 2.9}$$

where n' is the new sample size, n is the original sample size estimate, ρ is the intra-cluster correlation coefficient and m is the average number of people sampled per household. See Chapter 20 for further discussion of this issue. In Example 2.6, the sample size estimate from Example 2.5 is adjusted for a group-level study. An alternative approach applicable to studies with a dichotomous outcome is to base the sample size on a beta-binomial model with the prevalence of disease within each cluster having a binomial distribution and the prevalences between clusters following a beta distribution (Fosgate, 2007).

If the factor of interest is measured at the individual level (*eg* age), but also clusters within households (*ie* some households have older residents than other households), then the required

Example 2.6 Sample size with clustering
data = hypothetical

Presence/absence of a rainwater cistern is a household-level variable. Risk of diarrhea tends to be highly clustered within households and, in the Brazil diarrhea data the intra-class correlation (ρ) for diarrhea in households is about 0.45.

Assuming that there are, on average, 6 people in each household, the revised sample size that you will need will be:

$$\begin{aligned} n' &= n(1+\rho(m-1)) \\ &= 685(1+0.45(6-1)) \\ &= 2230 \end{aligned}$$

Consequently, you will need 2,230 people per group or 2230/6=372 households within each group. This very large increase in sample size results from the fact that the intra-cluster correlation for diarrhea is quite high (ρ=0.45) and that we are using a moderate number of observations (6) in each household.

sample size can be expected to lie somewhere between the simple estimate (ignoring clustering) and the much more conservative estimate required for household-level variables. In such cases, a simulation approach (Section 2.11.8) may be the best way to estimate a required sample size or assess power.

2.11.7 Adjustment of sample size in multivariable studies

If you want to consider confounding and interaction (Chapter 13) in your study, you generally need to increase your sample size (Smith and Day, 1984). If the confounder is not a strong confounder (odds ratio (*OR*) with disease and exposure between 0.5 and 2), then about a 15% increase is needed. If it is a stronger confounder, then a greater increase in study size should be used. For continuous-scaled confounders, estimate the correlation of the confounder with the exposure variable ρ_{ce}, and then multiply the unadjusted sample size by the factor $(1-\rho^2_{ce})^{-1}$. For k covariates, the corresponding formula is,

$$n' = n\left(\frac{1+(k-1)\,\rho^2_{ce}}{1-\rho^2_{ce}}\right) \qquad \textit{Eq 2.10}$$

where ρ_{ce} is an average correlation between the confounders and the exposure variable of interest. Thus, for 5 covariates with a ρ_{ce} approximately equal to 0.3, the increase in study size is 50%.

A similar approach is to start with a simple approach to estimating sample size for the key predictor (exposure) of interest and then modify this for the multivariable situation using the **variance inflation factor** (*VIF*) (Hsieh *et al*, 1998).

$$n' = n * VIF \qquad \textit{Eq 2.11}$$

where $VIF = 1/(1-\rho^2_{1,2,3,\ldots,k})$.

Note that $\rho^2_{1,2,3,\ldots,k}$ is the squared multiple correlation coefficient (between the key predictor and the remaining k-1 variables) or the proportion of variance of the key predictor that is explained when it is regressed on the other k-1 variables. In general, as k increases, then the multiple correlation increases, as does the *VIF*. The approach to estimating the *VIF* is the same for both continuous and binary covariates.

2.11.8 General approaches to sample-size estimation

As indicated in Section 2.11.4, computing sample size for analytical studies (*eg* comparing 2 means) can be done either by specifying the desired power of the study to detect a difference of a defined magnitude, or by specifying the desired width of the CI for the difference being estimated (*ie* a precision-based approach). In simple situations, these calculations are relatively straightforward. Two approaches to generalising these calculations for more complex study designs are described below.

Precision-based sample-size computations
The general formula for the width of a confidence interval of a parameter is:

$$par \pm Z * \mathrm{SE}(par) \qquad \textit{Eq 2.12}$$

where *par* is the parameter being estimated, *Z* is the desired percentile of the normal distribution and SE*(par)* is the SE of the parameter estimate.

Note For simplicity, the standard normal distribution will be used as a large sample approximation for the *t*-distribution throughout these examples.

For linear regression models, the SE of any parameter can take the general form of:

$$\mathrm{SE}(par) = \sigma * c \qquad \textit{Eq 2.13}$$

where σ is the residual standard deviation from the model and *c* is a value which will depend on the design of the study. For example, for estimating a mean in a single sample:

$$c = \sqrt{1/n} = 1/\sqrt{n} \qquad \textit{Eq 2.14}$$

where *n* is the sample size.

For a comparison of means from 2 samples:

$$c = \sqrt{2/n}$$

where *n* is the sample size in each of the 2 groups.

The formulae for the CI can be inverted to solve for *n*. For example, to estimate the difference between 2 means with the CI of the estimate being 2*L* units long (*ie* ±*L*), then:

$$L = Z_{\alpha} * \sigma * \sqrt{2/n} \qquad \textit{Eq 2.15}$$

Based on this, the sample size required is:

$$n = \frac{2Z_{\alpha}^2 \sigma^2}{L^2} \qquad \textit{Eq 2.16}$$

Eq 2.16 is the 2-sample analogue of Eq 2.5.

Note Unlike in Eq 2.7, we have not specified a Z_{β} nor have we specified hypothesised 'true' values for the 2 means. The sample size estimated is the one required to provide a confidence interval (for the difference) with a desired width (2*L*), regardless of what the actual difference is.

This approach can be generalised to any sort of sample-size estimation, provided that the structure of *c* can be determined. This is based on the design of the study. For example, computing the sample size required to evaluate a 2-way interaction between 2 dichotomous variables is equivalent to evaluating mean values in each of 4 possible groups (formed by the possible combinations of the 2 variables). Consequently:

$$c = \sqrt{4/n}$$

and the sample size required in each of the 4 groups will be:

$$n = \frac{4Z_{\alpha}^2 \sigma^2}{L^2}$$

This leads to the useful guideline that a study in which you want to evaluate interactions among dichotomous variables needs to be 4 times as large as is required to estimate main effects.

Power calculation by simulation
An approach to power calculation that is applicable to almost any analytical situation is one that is based on simulation (Feivesen, 2002). In general, you simulate a large number of datasets that are representative of the type that you are going to analyse, and then compute the proportion of times that the main factor you are interested in has a P-value less than, or equal to, the level you have set for significance (*eg* 0.05). This approach can be applied to multivariable regression-type models as well as simpler unconditional analyses.

There are 2 scenarios for generating the simulated datasets. In the first (and simplest) approach, you might want to evaluate the power of a study which you have already conducted. For example, let's assume that you evaluated the impact of rainwater cistern on the number of days with diarrhea (log transformed) for people in the Brazil study. A regression with family size (dichotomised) and cistern as the predictors (but ignoring the clustering of observations within households) provides an estimate of effect of β=0.054 (P=0.358) for large families (compared with <6 members). You would like to know what the power of this size study was to detect a difference of 0.054 units in log(diarrhea days).

The steps involved in determining the power by simulation are:

1. For each observation in the dataset, compute the predicted value based on the coefficients from the model and the particular X values (large family and rainwater cistern) for the observation.
2. Generate a random value for the outcome with a mean at the predicted value and a standard deviation equal to that observed in the data.
3. Reanalyse the data and note the P-value for the coefficient for the rainwater cistern (β_2) effect.
4. Repeat steps 1–3 many times (*eg* 1,000) and determine the proportion of datasets in which the P-value for the rainwater cistern effect is $\leq$0.05. This is an estimate of the power of the study to detect a true effect corresponding to β_2 =0.054.

Note This post-hoc power calculation has been presented because it is the simplest example of the use of simulation methods for sample-size calculation. In general, post-hoc power calculations are not useful (Hoenig and Heisey, 2001; Smith and Bates, 1992).

The second scenario arises if you want to compute sample sizes prior to conducting a study, the process is similar except that you start by creating a hypothetical dataset based on an expected final model. This means that you will need to specify the distributions of the X variables, the size of the dataset, the hierarchical structure of the data (if it is hierarchical in nature; see Chapters 20–22) and all of the relevant variance estimates. An example of the determination of the power of a future study, but based on some existing data (for covariate effects), is shown in Example 2.7.

2.12 SAMPLING TO DETECT DISEASE

Sampling to detect the presence (or confirm the absence) of disease is a much more common event in veterinary medicine (where documenting freedom from a disease is a common requirement for trade of livestock) than human medicine, but still warrants some attention. It is fundamentally different than sampling to estimate a parameter such as the prevalence of

Example 2.7 Power calculation by simulation
data = brazil_smpl

You have carried out a study to evaluate the effects of rainwater cistern and a dichotomised version of family size (<6 vs 6+) on the number of days with diarrhea in a month (log transformed). You carry out a regression analysis on these data to evaluate the effects of the two predictors. The important results from that regression analysis are:

- the coefficient for large family was 0.054 suggesting that people in large families have more diarrhea days, but the P-value was 0.358, so you have relatively little confidence that the estimate was really different from 0.
- the standard error of prediction for days with diarrhea was 0.605 (this represents the standard deviation of predicted results—see Chapter 14).

Assume that you would like to know the power of a comparable study (same size, same distribution of covariates) to detect a 0.07 unit increase in log(diarrhea days). The simulation process to answer this question is as follows.

You generate 1,000 datasets with randomly generated log(diarrhea days) values. For each person in each dataset, the log(diarrhea days) value is drawn from a normal distribution with the following characteristics:

- it has a mean value that corresponds to the predicted value from the real data that you started with (*ie* based on the rainwater cistern and family size variables) except that the effect of family size is now set to 0.07
- it has a standard deviation of 0.605

You analyse each of these new datasets and determine the proportion that gave a P-value for the family size coefficient that was >0.05. It turns out that the power would be 0.230 (23.0%). Consequently, if the true effect of family size is to increase log(diarrhea days) by 0.07 units, a study based on 337 small families and 159 large families will have a 23% chance of finding a significant effect of family size if the effect of rainwater cistern is controlled for. This estimate is very close to the estimate of 22.5% based on a simple comparison of 2 groups which ignores rainwater cistern status (computations not shown).

disease. If you want to be absolutely certain that a disease is not present in a population, then the only option is to test the entire population (and even this only works if the test you have is perfect). As this is rarely feasible, we rely on the fact that most diseases, if present in a population, will exist at or above some minimal prevalence. For example, we might think that if a contagious disease was present in a population, it would be very unlikely that less than 1% of the population would be infected. Based on this, you can compute a sample size required to be reasonably confident that you would detect the disease if the prevalence was 1% or higher.

If you are sampling from a finite population (*eg* <1,000 individuals), then the formula to determine the required sample size is (Cannon, 2001):

$$n = \left(1 - (\alpha)^{1/D}\right)\left(N - \frac{D-1}{2}\right) \qquad \textit{Eq 2.17}$$

where:

- n=required sample size
- α=1–confidence level (usually=0.05)
- D=estimated minimum number of diseased individuals in the group (population

size*minimum expected prevalence)
- N=population size.

If you are sampling from a large (infinite) population, then the following approximate formula can be used:

$$n = 1\text{n}\,\alpha / 1\text{n}\,q \qquad \textit{Eq 2.18}$$

where n=the required sample size, α is usually set to 0.05 or 0.01, q=(1–minimum expected prevalence).

If you take the required sample and get no positive results (assuming that you set α to 0.05), then you can say that you are 95% confident that the prevalence of the disease in the population is below the minimal threshold which you specified for the disease in question. Thus, you accept this as sufficient evidence of the absence of the disease. Example 2.8 shows the calculation of the required sample size to determine freedom from norovirus infection in a nursing home.

A much more complete discussion of issues related to sampling to determine freedom from disease has been published by Cameron and Baldock (1998a; 1998b). Bayesian procedures for sample size calculations for determination of freedom from disease which take into account the fact that the disease tends to cluster (in households or in regions) have been developed, but are beyond the scope of this text (Branscum *et al*, 2006).

Example 2.8 Sample size for freedom from disease
data = hypothetical

Assume that you want to demonstrate the elimination of norovirus from a 100-person nursing home following an outbreak with confirmed cases. You believe that if norovirus is present, a minimum of 10% of residents would still be positive (based on PCR testing) because of the highly contagious nature of the disease.

$$N = 100 \quad \alpha = 0.05 \quad D = 10$$
$$n = \left(1 - (\alpha)^{1/D}\right)\left(N - \frac{D-1}{2}\right)$$
$$= \left(1 - (0.05)^{1/10}\right)\left(100 - \frac{10-1}{2}\right)$$
$$= (0.259)(95.5)$$
$$= 24.7 \simeq 25$$

If you test 25 randomly selected residents and all test results are negative, you can state that you are 95% confident that the prevalence of norovirus infection in the home is <10%. As you don't believe that the disease would exist at a prevalence <10%, you are confident that it is not present. **Note** This assumes the test is 100% sensitive and specific. See Chapter 5 for a discussion of test characteristics. (If you use the large population formula (Eq 2.18), you get a sample size estimate of 28.4.)

REFERENCES

Branscum AJ, Johnson WO, Gardner IA. Sample size calculations for disease freedom and prevalence estimation surveys. Stat Med. 2006;25(15):2658-74.

Cameron AR, Baldock FC. Two-stage sampling in surveys to substantiate freedom from disease. Prev Vet Med. 1998a;34(1):19-30.

Cameron AR, Baldock FC. A new probability formula for surveys to substantiate freedom from disease. Prev Vet Med. 1998b;34(1):1-17.

Cannon RM. Sense and sensitivity—designing surveys based on an imperfect test. Prev Vet Med. 2001;49(3-4):141-63.

Dargartz DA, Hill GW. Analysis of survey data. Prev Vet Med. 1996;28:225-37.

Feivesen AH. Power by simulation. The Stata Journal. 2002;2:107-24.

Fleiss JL, Levin B, Paik MCB. Statistical methods for rates and proportions. 3rd Ed. New York: John Wiley and Sons; 2003.

Fosgate GT. A cluster-adjusted sample size algorithm for proportions was developed using a beta-binomial model. J Clin Epidemiol. 2007;60(3):250-5.

Heeringa SG, West BT, Berglund PA. Applied Survey Data Analysis. Boca Raton, FL: Chapman & Hall/CRC; 2010.

Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician. 2001;55:19-24.

Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med. 1998;17(14):1623-34.

Kreuter F, Vallian R. A survey on survey statistics: what is done and can be done in Stata. The Stata Journal. 2007;7:1-21.

Levy PS, Lemeshow S. Sampling of Populations. Methods and Applications 3rd Ed. New York: John Wiley & sons; 1999.

Prattley DJ, Cannon RM, Wilesmith JW, Morris RS, Stevenson MA. A model (BSurvE) for estimating the prevalence of bovine spongiform encephalopathy in a national herd. Prev Vet Med. 2007a;80(4):330-43.

Prattley DJ, Morris RS, Cannon RM, Wilesmith JW, Stevenson MA. A model (BSurvE) for evaluating national surveillance programs for bovine spongiform encephalopathy. Prev Vet Med. 2007b;81(4):225-35.

Robinson WT, Risser JM, McGoy S, Becker AB, Rehman H, Jefferson M, *et al*. Recruiting injection drug users: a three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. J Urban Health. 2006 Nov;83(6 Suppl):i29-38.

Rothman KJ, Greenland S, Lash TL. Modern Epidemiology, 3rd Ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

Salman MD. Animal Disease Surveillance. Salman MD, editor. Ames, Iowa: Iowa State Press;

2003.

Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology. 1992;3(5):449-52.

Smith PG, Day NE. The design of case-control studies: the influence of confounding and interaction effects. Int J Epidemiol. 1984;13(3):356-65.

Stärk KDC, Regula G, Hernandez J, Knopf L, Fuchs K, Morris RS, *et al*. Concepts for risk-based surveillance in the field of veterinary medicine and veterinary public health: review of current approaches. BMC Health Serv Res. 2006;6 20.

Thurmond MC. Conceptual foundations for infectious disease surveillance. J Vet Diagn Invest. 2003;15(6):501-14.

Wells SJ, Ebel ED, Williams MS, Scott AE, Wagner BA, Marshall KL. Use of epidemiologic information in targeted surveillance for population inference. Prev Vet Med. 2009;89(1-2):43-50.

Williams MS, Ebel ED, Wells SJ. Population inferences from targeted sampling with uncertain epidemiologic information. Prev Vet Med. 2009a;89(1-2):25-33.

Williams MS, Ebel ED, Wells SJ. Poisson sampling: a sampling strategy for concurrently establishing freedom from disease and estimating population characteristics. Prev Vet Med. 2009b;89(1-2):34-42.