PART II

Sampling Probability and Inference

The second part of the book looks into the probabilistic foundation of statistical analysis, which originates in probabilistic sampling, and introduces the reader to the arena of hypothesis testing.

Chapter 5 explores the main random and controllable source of error, sampling, as opposed to non-sampling errors, potentially very dangerous and unknown. It shows the statistical advantages of extracting samples using probabilistic rules, illustrating the main sampling techniques with an introduction to the concept of estimation associated with precision and accuracy. Non-probability techniques, which do not allow quantification of the sampling error, are also briefly reviewed. **Chapter 6** explains the principles of hypothesis testing based on probability theories and the sampling principles of previous chapter. It also explains how to compute confidence intervals and how statistics allow one to test hypotheses on one or two samples. **Chapter 7** extends the discussion to the case of more than two samples, through a class of techniques which goes under the name of analysis of variance. The principles are explained and with the aid of SPSS examples the chapter provides a quick introduction to advanced and complex designs under the broader general linear modelling approach.

CHAPTER 5

Sampling

THIS CHAPTER provides an introduction to sampling theory and the sampling process. When research is conducted through a sample survey instead of analyzing the whole target population, it is unavoidable to commit an error. The overall survey error can be split into two components:

(a) the sampling error, due to the fact that only a sub-set of the reference population is interviewed; and
(b) the non-sampling error, due to other measurement errors and survey biases not associated with the sampling process, discussed in chapters 3 and 4.

With probability samples as those described in this chapter, it becomes possible to estimate the population characteristics and the sampling error at the same time (inference of the sample characteristics to the population). This chapter explores the main sampling techniques, the estimation methods and their precision and accuracy levels depending on the sample size. Non-probability techniques, which do not allow quantification of the sampling error, are also briefly reviewed.

Section 5.1 introduces the key concepts and principles of sampling
Section 5.2 discusses technical details and lists the main types of probability sampling
Section 5.3 lists the main types of non-probability samples

THREE LEARNING OUTCOMES

This chapter enables the reader to:

- Appreciate the potential of probability sampling in consumer data collection
- Get familiar with the main principles and types of probability samples
- Become aware of the key principles of statistical inference and probability

PRELIMINARY KNOWLEDGE: For a proper understanding of this chapter, familiarity with the key probability concepts reviewed in the appendix at the end of this book is essential.

This chapter also exploits some mathematical notation. Again, a good reading of the same appendix facilitates understanding.

5.1 To sample or not to sample

It is usually unfeasible, for economic or practical reasons, to measure the characteristics of a population by collecting data on all of its members, as *censuses* aim to do. As a matter of fact, even censuses are unlikely to be a complete survey of the target population, either because it is impossible to have a complete and up-to-date list of all of the population elements or due to *non-response errors*, because of the failure to reach some of the respondents or the actual refusal to co-operate to the survey (see chapter 3).

In most situations, researchers try to obtain the desired data by surveying a sub-set, or *sample*, of the population. Hopefully, this should allow one to generalize the characteristics observed in the sample to the entire target population, inevitably accepting some margin of error which depends on a wide range of factors. However, generalization to the whole population is not always possible – or worse – it may be misleading.

The key characteristic of a sample allowing generalization is its probabilistic versus non-probabilistic nature. To appreciate the relevance of this distinction, consider the following example. A multiple retailer has the objective of estimating the average age of customers shopping through their on-line web-site, using a sample of 100 shopping visits. Three alternative sampling strategies are proposed by competing marketing research consultants:

1. (convenience sampling) The first 100 visitors are requested to state their age and the average age is computed. If a visitor returns to the web site more than once, subsequent visits are ignored.
2. (quota sampling) For ten consecutive days, 10 visitors are requested to state their age. It is known that 70% of the retailer's customers spend more than £ 50. In order to include both light and heavy shoppers in the sample, the researchers ensures that expenditure for 7 visits are below £ 50 and the remaining are above. The mean age will be a weighted average.
3. (simple random sampling) A sample of 100 customers is randomly extracted from the database of all registered users. The sampled customers are contacted by phone and asked to state their age.

The first method is the quickest and cheapest and the researcher promises to give the results in 3 days. The second method is slightly more expensive and time consuming, as it requires 10 days of work and allows a distinction between light and heavy shoppers. The third method is the most expensive, as it requires telephone calls.

However, only the latter method is probabilistic and allows inference on the population age, as the selection of the sampling units is based on random extraction.

Surveys 1 and 2 might be seriously biased. Consider the case in which daytime and weekday shoppers are younger (for example University students using on-line access in their academic premises), while older people with home on-line access just shop in the evenings or at the week-ends. Furthermore, heavy shoppers could be older than light shoppers.

In case one, let one suppose that the survey starts on Monday morning and by Tuesday lunchtime 100 visits are recorded, so that the sampling process is completed in less than 48 hours. However, the survey will exclude – for instance – all those that shop on-line over the week-end. Also, the sample will include two mornings and only one afternoon and one evening. If the week-end customers and the morning customers have different characteristics related to age, then the sample will be biased and the estimated age is likely to be lower than the actual one.

In case two, the alleged 'representativeness' of the sample is not guaranteed for similar reasons, unless the rule for extracting the visitors is stated as random. Let one suppose that the person in charge of recording the visits starts at 9 a.m. every day and (usually) by 1 p.m. has collected the age of the 3 heavy shoppers, and just 3 light shoppers. After 1 p.m.only light shoppers will be interviewed. Hence, all heavy shoppers will be interviewed in the mornings. While the proportion of heavy shoppers is respected, they're likely to be the younger ones (as they shop in the morning). Again, the estimated age will be lower than the actual one.

Of course, random selection does not exclude bad luck. Samples including the 100 youngest consumers or the 100 oldest ones are possible. However, given that the extraction is random (probabilistic), we know the likelihood of extracting those samples and we know – thanks to the normal distribution – that balanced samples are much more likely than extreme ones. In a nutshell, sampling error can be quantified in case three, but not in cases one and two.

The example introduces the first key classification of samples into two main categories – probability and non-probability samples. *Probability sampling* requires that each unit in the sampling frame is associated to a given probability of being included in the sample, which means that the probability of each potential sample is known. Prior knowledge on such probability values allows *statistical inference*, that is the generalization of sample statistics (*parameters*) to the target population, subject to a margin of uncertainty, or *sampling error*. In other words, through the probability laws it becomes possible to ascertain the extent to which the estimated characteristics of the sample reflect the true characteristic of the target population. The sampling error can be estimated and used to assess the precision and accuracy of sample estimates. While the sampling error does not cover the overall survey error as discussed in chapter 3, it still allows some control over it. A good survey plan allows one to minimize the non-sampling error without quantifying it and relying on probabilities and sampling theory opens the way to a quantitative assessment of the accuracy of sample estimates. When the sample is *non-probabilistic*, the selection of the sampling units might fall into the huge realm of subjectivity. While one may argue that expertise might lead to a better sample selection than chance, it is impossible to assess scientifically the ability to avoid the potential *biases* of a subjective (non-probability) choice, as shown in the above example.

However, it can not be ignored that the use of non-probability samples is quite common in marketing research, especially quota sampling (see section 5.3). This is a controversial point. Some authors correctly argue that in most cases the sampling error is much smaller than error from non-sampling sources (see the study by Assael and Keon, 1982) and that efforts (including the budget ones) should be rather concentrated on eliminating all potential sources of biases and containing non-responses (Lavrakas, 1996).

The only way to actually assess potential biases due to non-probability approaches consists of comparing their results from those obtained on the whole population, which is not a viable strategy. It is also true that quota sampling strategies such as those implemented by software for computer-assisted surveys (see chapter 3) usually guarantee minor violations of the purest probability assumptions; The example which opened this chapter could look exaggerated (more details are provided in section 5.3). However, this chapter aims to emphasize a need for coherence when using statistics. In the rest of this book, many more or less advanced statistical techniques are discussed. Almost invariably, these statistical methods are developed on the basis of probability assumptions, in most cases the normality of the data distribution. For example, the probability basis is central to the hypothesis testing, confidence intervals and ANOVA

techniques described in chapters 6 and 7. There are techniques which relax the need for data obtained through probability methods, but this is actually the point. It is necessary to know why and how sample extraction is based on probability before deciding whether to accept alternative routes and shortcuts.[1]

5.1.1 Variability, precision and accuracy: standard deviation versus standard error

To achieve a clear understanding of the inference process, it is essential to highlight the difference between various sources of *sampling error*. When relying on a sub-set of the population – the sample – it is clear that measurements are affected by an error. The size of this error depends on three factors:

1. The *sampling fraction*, which is the ratio between the sample size and the population size. Clearly, estimates become more precise as the sample size grows closer to the population size. However, a key result of sampling theory is that the gain in precision marginally decreases as the sampling fraction increases. Thus it is not economically convenient to pursue increased precision by simply increasing the sample size as shown in detail in section 5.3.
2. The *data variability* in the population. If the target variable has a large dispersion around the mean, it is more likely that the computed sample statistics are distant from the true population mean, whereas if the variability is small even a small sample could return very precise statistics. Note that the concept of *precision* refers to the degree of variability and not to the distance from the true population value (which is *accuracy*). The population variability is measured by the population variance and standard deviation (see appendix to this chapter). Obviously these population parameters are usually unknown as their knowledge would require that the mean itself is known, which would make the sampling process irrelevant. Estimates of the population variability are obtained by computing variability statistics on the sample data, such as the *sample standard deviation* and *sample variance*.
3. Finally, the success of the sampling process depends on the *precision of the sample estimators*. This is appraised by variability measures for the sample statistics and should not be confused with the sample variance and standard deviation.In fact the objective of measurement is not the data variability any more, but rather the variability of the *estimator* (the sample statistic) intended as a random variable distributed around the true population parameter across the sample space. For example, if the researcher is interested in the population mean value and a mean value is computed on sample data, the precision of such estimate can be evaluated through the *variance of the mean* or its square root – the *standard error of the mean*.

The distinction between standard deviation and standard error should be apparent if we think that a researcher could estimate the population *standard deviation* on a sample and the measure of the accuracy of the *sample standard deviation* will be provided by a statistic called *standard error of the standard deviation*.

Note that a *precise* sampling estimator is not necessarily an *accurate* one, although the two concepts are related. Accuracy measures closeness to the true population value, while precision refers to the variability of the estimator. For example, a sample mean

estimator is more accurate than another when its estimated mean is closer to the true population mean, while it is more precise if its standard error is smaller. Accuracy is discussed in section 5.2.2.

5.1.2 The key terms of sampling theory

Let one refer to a sub-set or *sample* of n observations (*sample size*), extracted from a *target population* of N elements. If we knew the true population value and the probability of extraction associated to each of the N elements of the target population (the *sampling method*), given the sample size n, we could define all potential samples, that is, all potential outcomes of the sampling process. Hence we could derive the exact distribution of any sample statistic around the true population value; the *sampling distribution* would be known exactly. To make a trivial example, consider the case where $N = 3$ and $n = 2$, where A, B and C are population units. There are three potential samples (A,B), (B,C) and (C,A). Together they constitute the sampling space.

Clearly enough, extracting all potential samples would be a useless exercise, given that the researcher could directly compute the true value of the population. So, rather than solving the above *direct problem*, the statistician is interested in methods solving the *indirect problem*, which means that:

(a) only one sample is extracted;
(b) only the *sample statistics* are known;
(c) a *sampling distribution* can be ascertained on the basis of the *sampling method*, this is the so-called specification problem; and
(d) estimates of the true values of the desired statistics within the target population are obtained from the sample statistics through *statistical inference.*

5.2 Probability sampling

Before exploring specific probabilistic sampling methods and inference, it is useful to introduce some basic notation (beyond the basic mathematical notation of the appendix to this chapter) to simplify discussion. If we define X_i as the value assumed by the variable $X$for the i-th member of the population, a *population* of N elements can be defined as:

$$P = (X_1, X_2, \ldots, X_i, \ldots, X_N)$$

whereas a sample of n elements extracted from P is identified by referring to the population subscripts

$$S = (x_{i_1}, x_{i_2}, \ldots, x_{i_j}, \ldots, x_{i_n})$$

where i_j indicates the population unit which is included as the j-th element of the sample, for example $i_1 = 7$ means that the first observation in the sample corresponds to the 7^{th} element of the population (or that $x_{1_7} = X_7$).

Sampling is needed to infer knowledge of some *parameters* of the population P, usually the population mean, the variability and possibly other distribution features (shape of the distribution, symmetry, etc.). As these are unknown, estimates are

BOX 5.1 *Sample statistics and inference in simple random sampling*

Population parameters	Sample statistics
$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$ (mean)	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean)
$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$ (standard deviation)	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$ (sample standard deviation)
$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$ (variance)	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ (sample variance)

obtained through the corresponding *sample statistics*.[2] In the rest of the discussion it will be assumed that variables in capital letters (Greek letter for statistics) refer to the population, while small letters are used for sample observations and statistics. Box 5.1 summarizes the equations for computing the basic population parameters and sample statistics, mean, standard deviation and variance. The population statistics are the usual central tendency and variability measured detailed in the appendix and discussed in chapter 4. The sample statistics are obviously quite similar (apart from the notation).

There are obvious similarities in the equations computing the population parameters (which are the usual descriptive statistics illustrated in the appendix) and the sample statistics. There is also a major difference for the variability measures, as the population standard deviation and variance have the number of elements in the denominator N, while for the sample statistics the denominator is $n-1$, which ensures that the estimator is *unbiased*. This means that the expected value of the estimator[3] is equal to the actual population parameter. This is a desirable property, together with *efficiency* and *consistency*. Efficiency refers to precision and means that the estimator is the most precise (has the lowest variance) among all unbiased estimators for the same population parameters. Consistency refers to accuracy and means that as the sample size becomes very large (tends to infinity), the probability that there is a difference between the estimator value and the actual population parameter tends to zero.

5.2.1 Basic probabilistic sampling methods

Since the only requirement for a probabilistic sampling method is that the probability of extraction of a population unit is known, it is clear that the potential range of sampling methods is extremely large. However, for many surveys it will be sufficient to employ one of the elemental methods, or a combination of two or more of them. Box 5.2 summarizes the key characteristics of the sampling methods discussed in this chapter. Those interested in the equations for estimating population statistics and their precisions are referred to the appendix to this chapter.

The methods listed in box 5.2 guarantee that extraction is probabilistic, hence inference from the sample statistics to the population. However, while inference from

a simple random sampling is quite straightforward, more complex sampling method like stratified and cluster sampling require that probabilities of extraction and sample allocation rules are taken into account to allow correct generalization of the sample statistics to the population parameters.

1. **Simple Random Sampling** has the desirable characteristic of being based on a straightforward random extraction of units from the population. It does not require any information on the population units but on their list or *sampling frame*. When this is available, the researcher only issue is to guarantee perfect

BOX 5.2 *Probabilistic sampling methods*

	Description	Notation
Simple Random Sampling	Each population unit has the same probability of being in the sample. It follows that all potential samples of a given size n have the same probability of being extracted.	N population size; n sample size; $\pi_i = n/N$ probability of extraction for each population unit i
Systematic sampling	The population (sampling frame) units are sorted from 1 to N according to some characteristic, only the first sampling unit is extracted randomly, while other units are extracted systematically following a step k. The *sampling step* k is given by the ratio N/n and must be an integer. The starting unit r is selected randomly among the first k unit of the sample. Then, this method requires the systematic extraction of every k-th unit thereafter, i.e. units $r+k$, $r+2k$, $r+3k$, …, $r+(n-1)k$. Note that the sorting order should be related to the target characteristic, while cyclical sorting should be avoided.	$k = N/n$ sampling step r starting unit $\pi_i = n/N = 1/k$ probability of extraction for each population unit i
Stratified Sampling	The population is subdivided into S complementary and exhaustive groups (*strata*) with respect to some relevant population characteristic, so that units are expected to be relatively similar (homogeneous) within each stratum and different (heterogeneous) across different strata, with respect to the variable being studied. Random extraction (as in simple random sampling) occurs within each stratum, so that adequate representativeness is ensured for all the identified population sub-groups. The sample size allocation across strata can be either proportional to the stratum size or follow some alternative allocation criterion aimed to increase precision, such as *Neyman allocation*	S number of strata N_j number of population units in stratum j n_j sample size within stratum j $\pi_{ij} = n_j/N_j$ probability of extraction for each population unit i within stratum j $N = \sum_{j=1}^{S} N_j$ $n = \sum_{j=1}^{S} n_j$

(Continued)

BOX 5.2 *Cont'd*

	Description	Notation
Cluster Sampling (Area Sampling)	When a population is naturally or artificially subdivided into G complementary and exhaustive groups *(clusters)* and reasonable homogeneity across these clusters and heterogeneity within each of them can be assumed, this method is based on the random selection of clusters. In one-stage cluster sampling, all the population units belonging to the sampled clusters are included in the sample. In two-stages cluster sampling, some form of probability sampling is applied within each of the sampled clusters. This method is especially useful (and economically convenient) when the population is subdivided in geographic areas (*Area sampling*), so that a sample of areas (for example regions) is extracted in the first stage, and simple random sampling is performed within each area.	G number of clusters g number of sampled clusters N_k number of population units in cluster k n_k sample size within cluster k **One-stage cluster sampling** ($n_k = N_k$) $\pi_{ik} = g/G$ probability of extraction for each population unit i within cluster k **Two-stage cluster sampling** ($n_k<N_k$) $\pi_{ik}=(g/G)(n_k/N_k)$ probability of extraction for each population unit i within cluster k $N = \sum_{j=1}^{G} N_j \quad n = \sum_{j=1}^{G} n_j$

randomness. Sample statistics such as mean values, standard deviations and variances, can be computed by applying equations listed in box 5.1. However, some limitations of this basic probability method should be pointed out. First, ignoring any potentially available additional information on the population units reduces the precision of estimates. Among the class of probability methods, simple random sampling is the least *efficient* method, as for a given sample size its sample statistics are those with the highest standard errors (that is, the lowest precision).

2. **Systematic Sampling** has very similar characteristics to simple random sampling and insofar as the units are listed randomly in the sampling frame, it only represents an extraction method for simple random sampling. This method becomes more interesting when the sorting order of the population units in the sampling frame is expected to be related to the target characteristic. The stricter is this relation, the higher is the gain in efficiency. Ideally, the sampling frame contains at least one known variable which is related to the target one and can be used to sort the population units. Let one refer to the example of section 5.1. Suppose that the database of registered users includes information on the time of registration. As explained, one may expect that the age of the customers is related to the time of the day and week-day they shop, so it is also likely that it is related to the registration information. If customers are sorted according to the time of registration, systematic sampling will guarantee that customers registering throughout the day are included in the sampling, hence increasing the probability that all potential ages will be represented in the sample with a gain in representativeness and precision. However, sorting out the population units also involves some risks, as if the sorting order is cyclical the consequence could be a serious bias. Let's say that the customers are sorted out by weekday

and time of registration, so that the sampling frame starts with customers registering on a Monday morning, then goes on with other Monday's registrants until the evening, then switches to Tuesday morning customers and so on. An inappropriate sampling step could lead to the inclusion in the sample of all morning registrants jumping all those who registered in afternoons or evenings. Clearly if the assumption that age is correlated to the time of registration is true, the application of systematic sampling would lead to the extraction of a seriously biased sample. In synthesis, systematic sampling should be preferred to simple random sampling when a sorting variable related to the target variable is available, taking care to avoid any circuity in the sorting strategy.

A desirable feature of systematic sampling is that it can also be used without requiring a sampling frame, or more precisely, the sampling frame is built while doing systematic sampling. For example, this is the case when only one in every ten customers at a supermarket till is stopped for interview. The sample is built without knowing the list of customers on that day and the actual population size N is known only when the sampling process is terminated.

3. **Stratified sampling** is potentially the most efficient among elemental sampling strategies. When the population can be easily divided into sub-groups (*strata*), random selection of sampling units within each of the sub-groups can lead to major gains in precision. The rationale behind this sampling process is that the target characteristics shows less variability within each stratum, as it is related to the stratification variable(s) and varies with it. Thus, by extracting the sample units from different strata, representativeness is increased. It is essential to distinguish stratified sampling from the non-probability quota sampling discussed in the example of section 5.1 and in section 5.4. Stratified sampling requires the actual subdivision of the sampling frame into subpopulations, so that random extraction is ensured independently within each of the strata. While this leads to an increase in costs it safeguards inference. Instead, quota sampling does not require any sampling frame and only presumes knowledge of the proportional allocation of the sub-populations. Even with a random quota sampling, the sampling units are extracted from the overall population subject to the rule to maintain the prescribed percentages of units coming from the identified sub-populations. Referring to the example, in order to implement a stratified sampling design, the researcher should:

   (1) associate the (average) amount spent to each of the customers in the database;
   (2) sub-divide the sampling frame into two sub-groups, one with those customers who spend more than £ 50, the other with those than spend less;
   (3) decide on the size of each of the strata (for example 70 light shoppers and 30 heavy shoppers if the population proportions are maintained);
   (4) extract randomly and independently from the two sub-populations; and
   (5) compute the sample mean through a weighted average.

   Besides increasing representativeness, stratified random sampling is particularly useful if separate estimates are needed for each sub-domain, for instance estimating the average on-line shopper age separately for heavy and light shoppers.

4. **Cluster sampling** is one of the most employed elemental methods and is often a component of more complex methods. Its most desirable feature is that it does not necessarily require a list of population units, and it is hence applicable with

BOX 5.3 *Stratum sample size in stratified sampling*

An issue that emerges from the choice of stratifying a sample is whether the allocation of the sample size into the strata should be proportional to the allocation of the target population. This has major consequences on the precision of the estimates and the representativeness of subdomain estimates. While sometimes it may be convenient to assume the same sample size for each of the subdomains (for example, when the objective is comparing means across sub-domains), in most situations the choice is between *proportional* and *optimal* allocation. Proportional allocation simply requires that the sample size allocation corresponds to the distribution of population units across the sub-domains:

$$n_i = N_i \cdot \frac{n}{N}$$

In this case the sample becomes self-weighting, as the selected units can be regarded as a unique sample and population parameters can be estimated without the need for applying weights. Proportional allocation greatly simplifies the estimation process, , provided that the response rate is the same across the sub-domains. However, non-response rates may greatly differ for some stratification variables (for example income), so that the computational gain disappear. Maintaining proportionality does not necessarily lead to the highest degree of precision, especially when the variability of the target variable is very different across strata. In such case, the allocation process should be based on information on the variability of the sub-domain, so that larger sample sizes are allocated to the strata with higher variability and smaller sample sizes to those with lower variation:

$$n_i = n \cdot \frac{N_i \sigma_{X_i}}{\sum_{i=1}^{S} N_i \sigma_{X_i}}$$

where σ_{Xi} is the standard deviation of the target variable within the i stratum. However, it is unlikely that such information is available to the researcher prior to the survey and it is usually derived from previous or pilot studies. The optimal allocation described above (also known as Neyman's allocation) ignores any additional information on sampling cost, which may differ across strata. Nevertheless, it is valid whenever the unit sampling cost is the same for all observation from any stratum. In some circumstances, sampling costs are different across strata. For instance, it may be more expensive to reach individuals in rural areas. More general equations can be obtained to maximize precision assuming that total cost $C$is a constraint (the most frequent case) and C_i is the unit sampling cost within stratum i (see Levy and Lemeshow, 1999, p. 161):

$$n_i = C \cdot \frac{N_i \sigma_{X_i} / \sqrt{C_i}}{\sum_{i=1}^{S} \left(N_i \sigma_{X_i} / \sqrt{C_i} \right)}$$

In a rare and ideal situation where the budget is not a problem and the objective is simply to maximize precision, for a given sample size n, the allocation is obtained as follows:

$$n_i = n \cdot \frac{N_i \sigma_{X_i} / \sqrt{C_i}}{\sum_{i=1}^{S} \left(N_i \sigma_{X_i} \sqrt{C_i} \right)}$$

More complex optimizations take into account multivariate sampling designs, which means optimizing allocation on a set variables rather than only one. These have been the subject of research (see Solomon and Zacks, 1970 or Bethel, 1989). When there is some flexibility on the budget and precision targets, complex optimization algorithms help choosing the best balance after weighting the relevance of costs versus precision through a system of weights. See for example Gentle (2006, ch. 10 and references therein) or Narula and Wellington (2002).

more convenient survey methods (as mall intercepts) and to geographical (area) sampling. In many situations, while it is impossible or economically unfeasible to obtain a list of each individual sampling unit, it is relatively easy to define groups (*clusters*) of sampling units.

To avoid confusion with stratified sampling, it is useful to establish the main difference between the two strategies. While stratified sampling aims to maximize homogeneity within each stratum, cluster sampling is most effective when heterogeneity within each cluster is high and the clusters are relatively similar.

As an example, consider a survey targeted to measure some characteristics of cinema audience for a specific movie shown in London. *A priori*, it is not possible to distinguish between those that will go and see that specific film. However, the researcher has information on the list of cinemas that will show the movie. A simple random sampling extraction can be carried out from the list to select a small number of cinemas, then an exhaustive survey of the audience on a given week can be carried out. Cluster sampling can be articulated in more stages. For instance, if the above study had to be targeted on the whole UK population, in a first stage a number of cities could be extracted and in a second stage a number of cinemas could be extracted within each city.

While the feasibility and convenience of this method is obvious, the results are not necessarily satisfactory in terms of precision. Cluster sampling produce relatively large standard errors especially when the sampling units within each cluster are homogeneous with respect to some characteristic. For example, given differences in ticket prices and cinema locations around the city, it is likely that the audience is relatively homogeneous in terms of socio-economic status. This has two consequences – first, it is necessary that the number of selected clusters is high enough to avoid a selection bias and secondly if there is low variability within a cluster, an exhaustive sampling (interviewing the whole audience) is likely to be excessive and unnecessarily expensive. Hence, the trade-off to be considered is between the benefits from limiting the survey to a small number of clusters and the additional costs due to the excessive sampling rate. In cases where a sampling frame can actually be obtained for the selected clusters, a convenient strategy is to apply simple random sampling or systematic sampling within the clusters.

5.2.2 Precision and accuracy of estimators

Given the many sampling options discussed in the previous sections, researchers need to face many alternative choices, not to mention the frequent use of mixed and complex methods, as discussed later. In most case, the decision is made easy by budget constraints. There are many trade-offs to be considered. For instance, random sampling is the least efficient method among the probability ones, as it requires a larger sample size to achieve the same precisions level of , say, stratified sampling. However, stratification is a costly process in itself, as the population units need to be classified according to the stratification variable.

As mentioned, efficiency is measured through the *standard error*, as shown in the appendix. In general, the standard error increases with higher population variances and decreases with higher sample sizes. However, as sample size increases the relative gain in efficiency becomes smaller. This has some interesting and not intuitive implications.

As shown below, it is not convenient to go above a certain sample size, because while the precision gain decreases the additional cost does not.

Standard error also increases with population size. However, the required sampling size to attain a given level of precision does not increase proportionally with population size. Hence, as shown in the example of box 5.4, the target sample size does not vary much between a relatively small population and a very large one.

A concept that might be useful when assessing the performance of estimators is *relative sampling accuracy* (or *relative sampling error*), that is a measure of accuracy for the estimated mean (or proportion) at a given level of confidence. If one considers simple random sampling, relative accuracy can be computed by assuming a normal distribution for the mean estimate. The *confidence level* α (further discussed in chapter 6) is the probability that the relative difference between the estimated sample mean and the population mean is larger than the relative accuracy level r:

$$\Pr\left(\left|\frac{\overline{x}-\overline{X}}{\overline{X}}\right| < r\right) = 1-\alpha$$

where $\overline{X}$ is the population mean and $1-\alpha$ is the level of confidence.[4] The above equation means that the probability that the difference between the sample mean and the population mean (in percentage terms) is smaller than the fixed threshold r is equal to the level of confidence. Higher levels of confidence (lower levels of α) imply less accuracy (a higher r), which means that there is a trade-off between the confidence level and the target accuracy level. Thus one first fixes α at a relatively small level, usually at 0.05 or 0.01, which means a confidence level of 95% or 99%, respectively. Then it becomes possible to determine the relative accuracy level $r$as a function of the population and sample size of the standard error and to a coefficient which depends on the value chosen for α.

The exact equation to compute the relative accuracy level is (Cochran, 1977):

$$r = \pm\frac{t_{\alpha/2}S_x}{\sqrt{n}\overline{X}}\sqrt{1-\frac{n}{N}}$$

where $t_{\alpha/2}$ is a fixed coefficient which depends from the chosen confidence level α and from the size of the sample.The meaning of $t_{\alpha/2}$ will be made clearer with the discussion of confidence intervals and hypothesis testing in chapter 6; here it may suffice to notice that the value of $t_{\alpha/2}$ (which is fixed and tabulated) becomes larger as the level of confidence increases (that is α becomes smaller). Thus, if one wants higher confidence on the estimate, then a larger r must be accepted, which means accepting a lower accuracy level. The above equation also shows that when the sampling fraction n/N becomes larger (that is the sample becomes larger compared to the population size), then the level of accuracy is higher. Relative accuracy is directly related to the standard deviation.

When a proportion of cases p, rather than a mean is concerned, the equation becomes.

$$r = \pm t_{\alpha/2}\sqrt{\frac{1-p}{p\cdot n}\left(1-\frac{n}{N}\right)}$$

which can be interpreted exactly as the equation for the mean.

5.2.3 Deciding the sample size

When it is not a straightforward consequence of budget constraints, sample size is determined by several factors. It may depend on the precision level, especially when the research findings are expected to influence major decisions and there is little error tolerance. While box 5.4 shows that a size of 500 could do for almost any population size, another driving factor is the need for statistics on sub-groups of the population, which leads to an increase in the overall sample size. Also, if the survey topic is liable to be affected by non-response issues, it might be advisable to have a larger initial sample size, although non-response should be treated with care as discussed in chapters 3 and 4. Sometimes, the real constraint is time. Some sampling strategies can be implemented quickly, other are less immediate, especially when it becomes necessary to add information to the sampling frame.

Since the precision of estimators depends on the sampling design, the determination of sample size also requires different equations according to the chosen sampling method. Furthermore, since the sampling accuracy is based on a probability design and distributional assumption, it is necessary to set *confidence levels*, (introduced in

BOX 5.4 *SRS: Relative precision of a mean estimator, population and sample sizes*

The following table provides an useful comparison of relative precisions with varying sample and population sizes, for a fixed population mean and standard deviation. Consider simple random sampling to measure average monthly household expenditure on pizza. Let's fix the 'true' population mean at \$ 20 and the population standard deviation at 9 and the confidence level at $\alpha = 0.05$.

	Population size						
Sample size	100	1,000	2,000	5,000	100,000	1,000,000	100,000,000
30	11.75%	16.28%	16.53%	16.78%	16.77%	16.78%	16.78%
50	6.39%	12.14%	12.46%	12.65%	12.78%	12.78%	12.78%
100	0.00%	8.04%	8.48%	8.75%	8.92%	8.93%	8.93%
200		5.02%	5.65%	6.02%	6.26%	6.27%	6.27%
500		1.98%	2.97%	3.56%	3.93%	3.95%	3.95%
1000		0.00%	1.40%	2.23%	2.76%	2.79%	2.79%
2000			0.00%	1.18%	1.93%	1.97%	1.97%

With a sample of 30 units on a population of 100 units, we commit an 11.75% relative error (in excess or in deficiency), i.e. about \$ 2.35. With the same sample and a population of 1000 units, the error only rises to 16.28% (\$ 3.26). And even with a population of 100 millions of units, the error is basically unchanged at 16.78% (£ 3.36). These are however quite big relative errors. Let's say that we need to stay under 5%. With a sample size of 200, we achieve a 5.02% error on a population of 1,000, while for all other population sizes, the error is slightly higher.
With a sample size of 500, we commit a 1.98% error on a population of 1,000 (but we have sampled half of the population units!), an error 2.97% with 2,000 units, then we manage to stay under 4% with any population size between 5,000 and 100,000,000. For any sample size, there is virtually no difference whether the population is made by 100 thousands units or 100 millions units. In marketing research, for estimating sample means, a sample of 500 is usually more than acceptable, whatever the population size.

previous section and further explained in section 6.1), which should not be confused with precision measures. The latter can be controlled scientifically on the basis of the method and the assumptions. But a high level of precision does not rule out very bad luck, as extracting samples randomly allows (with a lower probability) for extreme samples which provide inaccurate statistics. The confidence level specifies the risk level that the researcher is willing to take, as it is the probability value associated with a *confidence interval*, which is the range of values which will include the population value with a probability equal to the confidence level (see chapter 6). Clearly, the width of the confidence interval and the confidence level are positively related. If one wants a very precise estimate (a small confidence interval), the lower is the probability to find the true population value within that bracket (a lower confidence). In order to get precise estimates with a high confidence level, it is necessary to increase sample size. As discussed for accuracy, the confidence level is a probability measure which ranges between 0 and 1, usually denoted with $(1-\alpha)$, where α is the *significance level*, commonly fixed to 0.05, which means confidence level of 0.95 (or 95%), or to 0.01 (99%) when risk aversion towards errors is higher.

Another non-trivial issue faced when determining the sample size is the knowledge of one or more population parameters, generally the population variance. It is very unlikely that such value is known beforehand, unless measured in previous studies. The common practice is to get some preliminary estimates through a pilot study, through a mathematical model (see for example Deming, 1990, ch. 14) or even make some arbitrary (and conservative) assumption. Note that even in the situation where the population variance is underestimated (and hence the sample size is lower than required), the sample statistics will still be unbiased although less efficient than expected.

In synthesis, the target sample size increases with the desired accuracy level and is higher for larger population variances and for higher confidence levels. In simple random sampling, the equation to determine the sample size for a given level of accuracy is the following:

$$n = \left(\frac{t_{\alpha/2}\sigma}{r\mu} \right)^2$$

As before, $t_{\alpha/2}$ is a fixed value (the *t*statistic) which depends on the chosen confidence level $1-\alpha$. If the size is larger than 50, the fixed value $t_{\alpha/2}$can be replaced with the one taken from normal approximation ($z_{\alpha/2}$), σ is the known *standard deviation* of the population (usually substituted by an estimate *s*), *r*is the *relative accuracy* level defined above and μ is the *population mean*. Since the latter is also usually unknown (very likely, since most of the time it is the objective of the survey), it is also estimated through a pilot study or a conservative guess can be used.

Equations for determining sample size as a function of population parameters and target accuracy vary with the sampling method. For example, with stratified sampling the sample size for a pre-determined accuracy *r* is given by:

$$n = \left[\left(\sqrt{N} \frac{r\mu}{z_{\alpha/2} \sum_{j=1}^{S} \sqrt{N_j} s_j} \right)^2 + \frac{1}{N} \right]^{-1}$$

5.2.4 Post-stratification

A typical obstacle to the implementation of stratified sampling is the unavailability of a sampling frame for each of the identified strata, which implies the knowledge of the stratification variable(s) for all the population units. In such a circumstance it may be useful to proceed through simple random sampling and exploit the stratified estimator once the sample has been extracted, which increases efficiency. All that is required is the knowledge of the stratum sizes in the population and that such post-stratum sizes are sufficiently large. The advantage of post-stratifications is two-fold:

- It allows to correct the potential bias due to insufficient coverage of the survey (incomplete sampling frame); and
- It allows to correct the bias due to missing responses, provided that the post-stratification variable is related both to the target variable and to the cause of non-response

Post-stratification is carried out by extracting a simple random sample of size n, and then units are classified into strata. Instead of the usual SRS mean, a post-stratified estimator is computed by weighting the means of the sub-groups by the size of each sub-group. The procedure is identical to the one of stratified sampling and the only difference is that the allocation into strata is made ex-post. The gain in precision is related to the sample size in each stratum and (inversely) to the difference between the sample weights and the population ones (the complex equation for the unbiased estimator for the standard error of the mean is provided in the appendix to this chapter). The standard error for the post-stratified mean estimator is larger than the stratified sampling one, because additional variability is given by the fact that the sample stratum sizes are themselves the outcome of a random process.

5.2.5 Sampling solutions to non-sampling errors

A review of the non-sampling sources of errors is provided in chapter 3, but it may be useful to discuss here some methods to reduce the biases of an actual sample which is smaller than the planned one or less representative than expected. This is generally due to two sources of errors – *non-response errors*, when some of the sampling units cannot be reached, are unable or unwilling to answer to the survey questions and *incomplete sampling frames*, when the sampling frame does not fully cover the target population (*coverage error*). In estimating a sample mean, the size of these errors is

(a) directly proportional to the discrepancy between the mean value for the actually sampled units and the (usually unknown) mean value for those units that could not be sampled and
(b) inversely proportional to the proportion of non-respondents on the total sample (or to the proportion of those that are not included the sampling frame on the total population size).

The post-stratification method is one solution to non-response errors, especially useful when a specific stratum is under-represented in the sample. For example, a telephone survey might be biased by an under-representation of those people who spend less time at home, like commuters and those who often travel for work as

compared to pensioners and housewives. The relevance of this bias can be high if this characteristic is related to the target variable. Still considering the example of box 5.5, if the target variable is the level of satisfaction with railways, this group is likely to be very important. If the proportion of commuters on the target population is known (or estimated in a separate survey), post-stratification constitutes a *non-response* correction as it gives the appropriate weight to the under-represented subgroup.

Similarly, post-stratification may help in situations where the problem lies in the sampling frame. If the phone book is used as a sampling frame, household in some rural areas where not all households have a phone line will be under-represented. Another solution is the use of dual frame surveys, specifically two parallel sampling frames (for example white pages and electoral register) and consequently two different survey methods (personal interviews for those sampled through the electoral register). Other methods (see Thompson, 1997) consist in extracting a further random sample out of the list of non-respondents (to get representative estimates for the group of non-respondents), in using a *regression estimator* or a *ratio estimator* or in weighting sample units through same demographic or socio-economic variables measured in the population. This latter method is similar to post-stratification, but not identical as unlike post-stratification it does not require homogeneity within each stratum and heterogeneity across strata.

5.2.6 Complex probabilistic sampling methods

The probability methods discussed in this chapter can be combined and developed into more complex sampling strategies, aimed at increasing efficiency or reducing cost through more practical targeting.

A commonly used approach in sampling household is *two-stage sampling*. Most of the household budget surveys in Europe are based on this method. Two-stage methods imply the use of two different sampling units, where the second-stage sampling units are a sub-set of the first-stage ones. Typically, a sample of cities or municipalities is extracted in the first stage, while in the second stage the actual sample of households is extracted out of the first-stage units. Any probability design can be applied within each stage. For example, municipalities can be stratified according to their populations in the first stage, to ensure that the sample will include small and rural towns as well as large cities, while in the second stage one could apply *area sampling*, a particular type of cluster sampling where:

(1) each sampled municipality is subdivided into blocks on a map through geographical co-ordinates;
(2) blocks are extracted through simple random sampling; and
(3) all households in a block are interviewed.

Clearly, if the survey method is personal interview, this sampling strategy minimizes costs as the interviewers will be able to cover many sampling units in a small area.

As complex methods are subject to a number of adjustments and options, the sample statistics can become very complex themselves. Box 5.6 below brings real examples of complex sampling design in consumer research, while box 5.7 illustrates how samples can be extracted from sampling frames in SPSS and SAS.

BOX 5.5 *Sample size determination: an example*

Consider a survey whose target is to estimate the level of satisfaction of travellers on the train service between London and Reading, on a population of 13,500 season ticket holders. The target variable is measured on a 1–7 Likert scale, the required precision level is ±10% ($r = 0.1$) and the required confidence level is 95% ($\alpha = 0.05$). First, we need an estimate (or a conservative guess) of the population mean and standard deviation. If we want a conservative estimate, we may choose the population average and standard deviation that maximize the ratio (σ/μ). The average μ must be included between 1 and 7 and the *maximum* standard deviation depends on the distribution we expect from the variable (Deming, 1990, ch. 14). If we assume a **normal distribution** around the mean (where the mean is 4), we can estimate that the standard deviation is about 1.15, in fact:

- within the normal distribution that 99% of values are between $4 - 2.6\sigma$ and $4 + 2.6\sigma$
- 100% of values are between 1 and 7
- With a good degree of approximation we can assume that $4 - 2.6\sigma = 1$ and $4 + 2.6\sigma = 7$, which means that σ **= 3/2.6 = 1.15**
- An alternative rule of thumb from statistical literature tells us that when the distribution is normal, then σ **= max–min/6 = 6/6 = 1**

Other distribution are possible, though. The **binomial distribution** assumes that the data are concentrated in the two extreme values and the maximum variability is given in the case where the respondents are exactly split in two halves, that is half of the respondents say 1 and the other half say 7. The average is still 4 and the standard deviation is given by σ **= max–min/2 = 6/2 = 3**
The **rectangular distribution** assumes that the values are equally distributed across the various options (1/6 of the respondents say 1, 1/6 say 2, 1/6 say 3, etc.). Again the average is 4, but the standard error is given by σ **= max–min*0.29 = 6*0.29 = 1.74**
There are other distribution possible, but all have a standard error smaller than in the case of the **binomial distribution**. To summarize: if we assume a normal distribution, then σ **= 1.15**. If we assume a binomial distribution, then σ **= 3**. The last assumption we need is related to the level of confidence α. This is different from the precision level and depends on the fact that we could still get very peculiar samples. The lower is α, the less likely is that our results are biased by the extraction of very unlikely samples. Usually we set $\alpha = 0.05$.
Now we have all elements to estimate our sample size. The most conservative estimate will be based on the binomial distribution.One remaining issues regards the choice between the *t* (student) distribution and the *z* one. Initially, let's be very conservative and let's assume that the sample is going to include around 20 units. Hence, with $\alpha = 0.05$, we use the statistic $t_{0.025}(20) = 2.09$
The sample size for simple random sampling is computed as follows:

$$n = \left(\frac{t_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{2.09 \cdot 3}{0.1 \cdot 4}\right)^2 = 246$$

Since the sample size is much higher than 50, we might have used the normal approximation. In such case $z_{0.025} = 1.96$:

$$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{1.96 \cdot 3}{0.1 \cdot 4}\right)^2 = 216$$

Hence, in the **most conservative** case, a sample size of 216 will guarantee the desired precision.

(Continued)

BOX 5.5 *Cont'd*

Such sample size varies if we use alternative assumption on the data distribution of the target variable (not to be confused with the distribution used in the equation, which depends on the sample size). Examples are:

Rectangular (uniform) distribution of the target variable

$$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{1.96 \cdot 1.74}{0.1 \cdot 4}\right)^2 = 73$$

Normal distribution of the target variable

$$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{1.96 \cdot 1.15}{0.1 \cdot 4}\right)^2 = 32$$

In the above case the sample size is low enough to suggest the use of the *t* distribution in the sample size equation, even if this has a minor impact:

$$n = \left(\frac{t_{\alpha/2}(30)\sigma}{r\mu}\right)^2 = \left(\frac{2.04 \cdot 1.15}{0.1 \cdot 4}\right)^2 = 34$$

We could also get different sample size by setting a higher confidence level. For a 99% confidence level ($\alpha = 0.01$)these are the results

Binomial distribution $$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{2.58 \cdot 3}{0.1 \cdot 4}\right)^2 = 374$$

Rectangular distribution $$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{2.58 \cdot 1.74}{0.1 \cdot 4}\right)^2 = 126$$

Normal distribution $$n = \left(\frac{z_{\alpha/2}\sigma}{r\mu}\right)^2 = \left(\frac{2.58 \cdot 1.15}{0.1 \cdot 4}\right)^2 = 55$$

According to different assumptions, we derived desired sampling sizes ranging from 34 to 374. So, what is the choice?
The first question would be: is it safe to assume a normal distribution of the target variable in the population? While this is not an extremely strong assumption for some variable observed in nature (height, weight, etc.), for attitudinal studies it might be risky. It could be that the population follows an uniform distribution or even that it is exactly split into the two extremes (think about a perfectly bipolar political system). So, if we want to stay on 'safe grounds', we assume a binomial distribution. For the remaining choice, the level of confidence, the researcher has the freedom (arbitrariness) to choose among the alternatives, usually based on the available budget. Usually a level of confidence of 95% is considered to be acceptable. In our case, a reasonable proposal could be to build a sample of 216 season ticket holders.

BOX 5.6 *Examples of complex sampling designs*

Expenditure and Food Survey (United Kingdom)	*Method*: Two-stage stratified random sample with clustering *Interview methods*: Face-to-face and diaries *Sample size*: 7,048 *Basic survey unit*: household *Description*: The EFS household sample is drawn from the Small Users file of the Postcode Address File - the Post Office's list of addresses. Postal sectors (ward size) are the primary sample unit, 672 postal sectors are randomly selected during the year after being arranged in strata defined by Government Office Regions (sub-divided into metropolitan and non-metropolitan areas) and two 1991 Census variables – socioeconomic group and ownership of cars.
Consumer Expenditure Survey (United States)	*Method*: Two-stage stratified random sampling *Interview methods*: Face-to-face and diaries *Sample size*: approximately 7,600 per quarter (interviews) and about 7,700 per year (diaries) *Basic survey unit*: household *Description*: The selection of households begins with the definition and selection of primary sampling units (PSUs), which consist of counties (or parts thereof), groups of counties, or independent cities. The sample of PSUs used for the survey consists of 105 areas, classified into four categories: 31 'A' PSUs, which are Metropolitan Statistical Areas (MSAs) with a population of 1.5 million or greater, 46 'B' PSUs, which are MSAs with a population less than 1.5 million, 10 'C' PSUs, which are nonmetropolitan areas used in the BLS Consumer Price Index program (CPI), 18 'D' PSUs, which are nonmetropolitan areas not used in the CPI. Within these PSUs, the sampling frame is generated from the census file.
Chinese Rural Household Survey	*Method*: Three-stage sampling method with systematic PPS (systematic Probability Proportional to size) in the first two stages *Interview method*: diaries filled with the interviewer help *Sample size*: about 67,000 *Basic survey unit*: household *Description*: At the first stage, systematic PPS drawing of 18 to 30 counties (PSUs) from each province (about 850 counties). At the second stage, administrative villages (about 20,000) are drawn from counties as SSUs with a systematic PPS method. At the third stage, households are drawn from each selected village with equal probability simple random sampling and up to 10 households are allocated to each SSUs

BOX 5.7 *Sampling with SPSS and SAS*

SPSS allows selection of random samples, including complex designs. Once you have organized your sampling frame into an SPSS workfile, Simple Random Sampling extraction is straightforward, simply click on Data\Select Cases\Random Sample, then set the sampling size clicking on SAMPLE and draw the sample by clicking on OK. Non-sampled units can be filtered out or deleted. For more complex sampling designs, SPSS has an add-on module (SPSS Complex Samples™) which let you control multi-stage sampling methods and choose design variables, techniques (stratified and cluster sampling plus sampling a number of options on SRS), sample sizes, with room for personalization and an output which includes inclusion probabilities and sample weight for each of the cases. To create complex sample, click on ANALYZE/COMPLEX SAMPLES/SELECT A SAMPLE, then after naming the file corresponding to the sampling procedure, a 'wizard' box will guide you through the steps of sampling.

In SAS, the SAS\STAT component has procedures for the extraction of samples and statistical inference (chapters 11, 61, 62 and 63 in SAS/STAT User's Guide version 8). The procedure SURVEYSELECT allows to extract probability-based samples, while the procedure SURVEYMEANS compute sample statistics taking into account the sample design. It is also possible to estimate sample-based regression relationships with SURVEYREG. Besides simple random sampling, systematic sampling and sampling proportional to size (PSS) as the one described for the China rural household survey, multi-stage sampling is allowed by these procedures, as well as stratified and cluster sampling and sampling methods with unequal inclusion probabilities.

5.3 Non-probability sampling

The previous two sections have shown how the probability foundation of sampling extraction allows an estimation of population parameters together with evaluations of precision and accuracy. While this is not possible for samples that are not extracted according to probability rules, non-probability sampling is a common practice, especially quota sampling. It should also be noted that non-probability sampling is not necessarily biasing or uninformative, in some circumstances – for example when there is no sampling frame – it may be the only viable solution. The key limit is that generally techniques for statistical inference cannot be used to generalize sample results to the population, although there is research showing that under some conditions statistical model can be applied to non-randomly extracted samples (Smith, 1984). Given the frequent use of non-probability samples, it may be helpful to review the characteristics, potential advantages and limits of the most common non-probability samples.

The extreme of non-probability sampling is the so-called **convenience sampling**, where units that are easier to be interviewed are selected by the researcher. There are many examples of studies where the sample representativeness is not important and the aim is to show a methodology, where academics interview students in the class where they are teaching. Clearly this is the cheapest possible method (other than making up the data). As with any other more elegant non-probability sample, inference is not possible. But – even worse – by definition these samples are affected by a *selection bias*. In sampling, selection bias consists in assigning to some units a higher probability of being selected, without acknowledging this within a probabilistic sampling process. If the units with higher inclusion probabilities have specific characteristics that differ from the rest of the population – as it is often the case – and these characteristics are related to the target variable(s), sample measurement will suffer from a significant bias.

There are many sources of selection bias and some depend on the interview method rather than the sampling strategy. Consider this example of a selection bias due to convenience sampling. A researcher wants to measure consumer willingness to pay for a fish meal and decides to interview people in fish restaurants, which will allow him to achieve easily a large sample size. While the researcher will be able to select many fish lovers, it is also true that the sample will miss all those people who like fish but consider existing restaurant fish meals too expensive and prefer to cook it at home. The selection bias will lead to a higher willingness to pay than the actual one. The sample cannot be considered representative of consumers in general, but only of the set of selected sample units.

Even if the researcher leaves behind convenience and tries to select units without any prior criteria, as in *haphazard sampling*, without a proper sampling frame and an extraction method based on probability inference is not valid.

In other circumstances, the use of a prior non-probability criterion for selection of sample units is explicitly acknowledge. For example, in **judgmental sampling**, selection is based on judgment of researchers who exploit their experience to draw a sample that they consider representative of the target population. The subjective element is now apparent.

Quota sampling is possibly the most controversial (and certainly the most adopted) of non-probability technique. It is often mistaken for stratified sampling, but it generally does not guarantee the efficiency and representativeness of its counterpart. As shown in box 5.1, quota sampling does not follow probabilistic rules for extraction. Where there are one or more classification variables for which the percentage of population units is known, quota sampling only requires that the same proportions apply to the sample. It is certainly cheaper than stratified sampling, given that only these percentages are required as compared to the need for a sampling frame for each stratum. The main difference from judgmental sampling is that judgment is based on the choice of variables defining quotas. Sampling error cannot be quantified in quota sampling and there is no way to assess sample representativeness. Furthermore, this sampling method is exposed to *selection biases* as extraction within each quota is still based on haphazard methods if not on convenience or judgment. There are some relevant advantages in applying quota sampling, which may explain why it is common practice for marketing research companies. Quota sampling is often implemented in computer packages for CATI, CAPI and CAWI interviews (see chapter 3), where units are extracted randomly but retained in the sample only if they are within the quota. While this alters the probability design, it offers some control on biases from non-response errors, since non-respondents are substituted with units from the same quota. The point in favour of quota sampling is that a probability sample with a large portion of non-responses is likely to be worse than quota sampling dealing with non-responses. Obviously, probability sampling with appropriate methods for dealing with non-responses as those mentioned in section 5.2.5 are preferable.

If the objective of a research is to exploit a sample to draw conclusion on the population or generalize the results, non-probability sampling methods are clearly inadequate. For confirmatory studies where probability sampling is possible, non-probability methods should be always avoided. However, in many circumstances they are accepted, especially when the aim of research is *qualitative* or simply preliminary to *quantitative* research (like piloting a questionnaire). Qualitative research is exploratory rather than confirmatory and provides an unstructured and flexible understanding of the problem and its reliance on the basis of small samples. It can be a valuable tool when operating in contexts where a proper sampling process is unfeasible. A typical

situation is when the target population is a very rare population, small and difficult to be singled out. In this circumstance, a useful non-probability technique is **snowball sampling**. As indicated by its name, snowball sampling starts with the selection of a first small sample (possibly randomly). Then, to increase sample size, respondents are asked to identify others who belong to the population of interests, so that the referrals have demographic and psychographic characteristics similar to the referrers. Suppose, for example, that the objective is to interview those people who climbed Mount Everest in a given year (or, say, the readers of this book) and no records are available. After selecting a first small sample, it is very likely that those that have been selected will be able to indicate others who accomplished to the task.

Summing up

Sampling techniques allow one to estimate statistics on large target populations by running a survey on a smaller sub-set of subjects. When subjects are extracted randomly to be included in a sample and the probabilities of extraction are known, the sample is said to be probabilistic. The advantage of probability samples compared to those extracted through non-probability rules is the possibility of estimating the sampling error, which is the portion of error in the estimates which is due to the fact that only a sub-set of the population is surveyed. This does not exhaust the total error committed by running a survey because non-sampling errors like non-response errors also need to be taken into account (see chapters 3 and 4). The basic form of probability sampling is simple random sampling, where all subjects in the target population have the same probability of being inextracted. Other forms of probability sampling include: systematic sampling, where subjects are extracted at a systematic pace and cluster sampling, where the population is first subdivided into cluster that are similar between each other and then a sub-set of cluster is extracted; stratified sampling, where the population is first subdivided into strata that contains homogeneous subjects but are quite different from those in other strata. These techniques can be combined in more complex sampling strategies, as in the multi-stage techniques adopted for official household budget survey. Once the sample has been extracted it is possible to generalize (infer) the sample characteristics like mean and variance to the whole population by exploiting knowledge of the probability distribution (see chapter 6). Estimates of the parameters are accompanied by estimates of their precision, as generally measured by the standard error. Accuracy (that is departure from the true population mean) can also be assessed with some degree of confidence. When planning a sample survey, there is a trade-off between sample size (cost) and precision. The latter also depends on the variability and dimension of the target population. Statistical rules allows to determine the sample size depending on targeted accuracy and vice versa.

Non-probability samples, including the frequently employed quota sampling, are not based on statistical rules and depend on subjective and convenience choices. They can be useful in some circumstances where probability sampling is impossible or unnecessary, otherwise probability alternatives should be chosen.

Appendix

Estimators

	Simple random sampling and systematic sampling	Stratified sampling	One-stage cluster sampling with unknown N
Mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x}_i = \frac{\sum_{i=1}^{n_i} x_{ji}}{n_i}$ (stratum); $\bar{x}_{ST} = \frac{\sum_{j=1}^{S} \bar{x}_j N_j}{N}$	$\bar{x}_j = \frac{\sum_{i=1}^{N_j} x_{ji}}{N_j}$ (cluster); $\bar{x}_{CL} = \frac{G \cdot \sum_{j=1}^{g} \sum_{i=1}^{N_j} x_{ji}}{g \cdot \sum_{j=1}^{g} N_j}$
Variance	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$	$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$ (stratum); $s_{ST}^2 = \frac{1}{N-1} \left[\sum_{j=1}^{S} (N_j - 1) s_j^2 + \sum_{j=1}^{S} (\bar{x}_j - \bar{x}_{ST})^2 N_j \right]$	$s_w^2 = \frac{1}{n-1} \sum_{j=1}^{g} \sum_{i=1}^{N_j} (x_{ji} - \bar{x}_j)^2$ (within) $s_b^2 = \frac{1}{g-1} \sum_{j=1}^{g} (\bar{x}_j - \bar{x}_{CL})^2$(between) $s_{CL}^2 = \frac{G(N-G)s_w^2 + (G-1)Ns_b^2}{G(N-1)}$
Proportion	$p = y/n$	$p_j = y_j/n_j$ (stratum); $p_{ST} = \sum_{j=1}^{S} p_j N_j / N$	$p_j = y_j/N_j$ (cluster); $p_{CL} = \frac{\sum_{j=1}^{g} y_j}{n}$

Precision estimates

	Simple random sampling and systematic sampling	Stratified sampling	One-stage cluster sampling with unknown N
Standard error of the mean	$s_{\bar{x}} = s\sqrt{\frac{N-n}{Nn}}$	$s_{\bar{x}_{ST}} = \frac{1}{N}\sqrt{\sum_{i=1}^{S} s_i^2 N_i^2 \left(\frac{N_i - n_i}{N_i}\right)}$	$s_{\bar{x}_{CL}} = \sqrt{\frac{(G-g)g}{G(g-1)}}\sqrt{\sum_{i=1}^{g}\left(\bar{x}_i - \bar{x}_{CL}\right)^2\left(\frac{N_i}{n}\right)^2}$
Standard error of the proportion	$s_p = \sqrt{p(1-p)}\sqrt{\frac{N-n}{N(n-1)}}$	$s_{p_{ST}} = \frac{1}{N}\sqrt{\sum_{i=1}^{S} \frac{p_i(1-p_i)N_i^2}{n_i - 1}\left(\frac{N_i - n_i}{N_i}\right)}$	$s_{p_{CL}} = \frac{1}{N}\sqrt{\frac{G(G-g)}{g(g-1)}\sum_{i=1}^{g}\left(y_i - p_{CL}N_i\right)^2}$

Post stratification

Mean	$\bar{x}_{PS} = \frac{\sum_{i=1}^{S} \bar{x}_i N_i}{N}$
Standard error of the mean	$s_{\bar{x}_{PS}} = \sqrt{\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{S}\left(\frac{N_i}{N}\right)s_i^2 + \left(\frac{N-n}{n^2N}\right)\sum_{i=1}^{S} s_i^2\left(\frac{N-N_i}{N-1}\right)}$

EXERCISES

1) Open the EFS data-set in SPSS
   a. Compute the mean of the variable cc3111 (clocks, watches and jewellery) on the whole data-set
   b. Now, extract a sample of 60 units and compute the sample means using:
      i. Simple random sampling
      ii. Stratified sampling (using INCRANGES - income quartiles as stratification variable)
      iii. Systematic sampling
   c. What is the sampling error? Which one is the most accurate method? And the most precise?
   d. Sort the sample by INCANON and perform systematic sampling again – how do results change? Why?
2) Open the Trust data-set in SPSS
   a. Compute the descriptive statistics of the variable q4kilos on the whole sample (N = 500)
   b. Suppose you need to extract a simple random sample of 50 elements, what is the estimated relative accuracy at a 95% confidence level? (HINT: $t_{\alpha/2} = t_{0.025} = 1.68$)
   c. If the target accuracy is ±15%, what is the sample size?
3) Open the EFS data-set in SPSS
   a. Compute the mean of the variable p609 (recreation expenditure)
   b. If the distribution of age in the population is the following:
      i. Less than 30 = 20%
      ii. 30–55 = 40%
      iii. More than 55 = 40%

Post-stratify (using the AGEBIN variable and the proportion above) and compute the mean again. Does it differ? Why?

Further readings and web-links

- **Is random sampling necessary** – This chapter has explained the advantages of random sampling, however scientists are still debating on the extent of advantages from sampling. A good discussion is provided in Johnstone (1989).
- **Good quota sampling** – Many studies have looked into the problems of quota sampling compared to random sampling (see e.g. Moser and Stuart, 1953). King (1985) looks at the combination of quota sampling with probabilistic sampling, with a varying weight of the two methods depending on the budget constraints. See also the review by Marsh and Scarbrough (1990).
- **Area sampling** – When a sampling frame of the population is not available, a good way to ensure probabilistic extraction is the application of area sampling. See Hansen and Hauser (1945) and Baker (1991).
- **Failure of non-probability sampling and the effect of non-response** – There is an infamous case of survey failure, the prediction of the 1936

US elections (Landon vs. Roosevelt). To know more about this story, see the original results article on the 'History Matters' web-site (historymatters.gmu.edu/d/5168/) or the actual results from the elections (search on wikipedia.org for 'United States presidential election, 1936') and read the paper by Squire (1988).

Hints for more advanced studies

- ☞ If you are not convinced about the desirability of probabilistic sampling methods, how can one test the potential biases of non-probability sampling? Try and build your population data, extract several non-random and random samples and compare the performances.
- ☞ What are the sampling algorithms available in the most popular software for CATI interviews?
- ☞ This chapter discussed inference on means and variability. However, most of the times one may want to infer complex relationships from the sample by using multivariate techniques like regression analysis, cluster analysis, etc… What are the implications for sampling?

Notes

1. See the 'Further Readings' section at the end of this chapter for a more advanced discussion of the topic of random sampling vs.non-probability sampling.
2. For a review of the essential univariate statistics see the appendix.
3. The expected value of an estimator is the average across all possible values of the statistic and may be interpreted as the average value of the estimator across the sampling space (again, see the appendix for an interpretation of 'expected values').
4. A more thorough discussion of the concepts of confidence and significance can be found in chapter 6.