Sample Surveys: Design, Methods and Applications, Vol. 29A
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00001-1

1

Introduction to Survey Sampling

Ken Brewer and Timothy G. Gregoire

1. Two alternative approaches to survey sampling inference

1.1. Laplace and his ratio estimator

At some time in the mid-1780s (the exact date is difficult to establish), the eminent mathematician Pierre Laplace started to press the ailing French government to conduct an enumeration of the population in about 700 communes scattered over the Kingdom (Bru, 1988), with a view to estimating the total population of France. He intended to use for this purpose the fact that there was already a substantially complete registration of births in all communes, of which there would then have been of the order of 10,000. He reasoned that if he also knew the populations of those sample communes, he could estimate the ratio of population to annual births, and apply that ratio to the known number of births in a given year, to arrive at what we would now describe as a ratio estimate of the total French population (Laplace, 1783[1], 1814a and 1814b). For various reasons, however, notably the ever-expanding borders of the French empire during Napoleon's early years, events militated against him obtaining a suitable total of births for the entire French population, so his estimated ratio was never used for its original purpose (Bru, 1988; Cochran, 1978; Hald, 1998; Laplace, 1814a and 1814b, p. 762). He did, however, devise an ingenious way for estimating the precision with which that ratio was measured. This was less straightforward than the manner in which it would be estimated today but, at the time, it was a very considerable contribution to the theory of survey sampling.

1.2. A prediction model frequently used in survey sampling

The method used by Laplace to estimate the precision of his estimated ratio was not dependent on the knowledge of results for the individual sample communes, which

[1] This paper is the text of an address given to the Academy on 30 October 1785, but appears to have been incorrectly dated back to 1783 while the Memoirs were being compiled. A virtually identical version of this address also appears in Laplace's *Oeuvres Complètes* 11 pp. 35–46. This version also contains three tables of vital statistics not provided in the Memoirs' version. They should, however, be treated with caution, as they contain several arithmetical inconsistencies.

would normally be required these days for survey sampling inference. The reason why it was not required there is chiefly that a particular model was invoked, namely one of drawing balls from an urn, each black ball representing a French citizen counted in Laplace's sample, and each white ball representing a birth within those sample communes in the average of the three preceding years. As it happens, there is another model frequently used in survey sampling these days, which leads to the same ratio estimator. That model is

$$Y_i = \beta X_i + U_i, \tag{1a}$$

which together with

$$E(U_i) = 0, \tag{1b}$$

$$E\left(U_i^2\right) = \sigma^2 X_i \tag{1c}$$

and

$$E(U_i U_j) = 0 \tag{1d}$$

for all $j \neq i$ can also be used for the same purpose.

Equation (1a) describes a survey variable value Y_i (for instance the population of commune i) as generated by a survey parameter, β, times an auxiliary value, X_i, (that commune's average annual births) plus a random variable, U_i. Equation (1b) stipulates that this random variable has zero mean, Eq. (1c) that its variance is proportional to the auxiliary variable (in this case, annual births), and Eq. (1d) that there is no correlation between any pair of those random variables.

Given this model, the minimum variance unbiased estimator of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}, \tag{2}$$

which in this instance is simply the ratio of black to white balls in Laplace's urn.

1.3. The prediction model approach to survey sampling inference

While, given the model of Eqns. (1), the logic behind the ratio estimator might appear to be straightforward, there are in fact two very different ways of arriving at it, one obvious and one somewhat less obvious but no less important. We will examine the obvious one first.

It is indeed obvious that there is a close relationship between births and population. To begin with, most of the small geographical areas (there are a few exceptions such as military barracks and boarding schools) have approximately equal numbers of males and females. The age distribution is not quite so stable, but with a high probability different areas within the same country are likely to have more or less the same age distribution, so the proportion of females of child-bearing age to total population is also more or less constant. So, also with a reasonable measure of assurance, one might expect the ratio of births in a given year to total population to be more or less constant, which makes the ratio estimator an attractive choice.

We may have, therefore, a notion in our minds that the number in the population in the ith commune, Y_i, is proportional to the number of births there in an average year, X_i, plus a random error, U_i. If we write that idea down in mathematical form, we arrive at a set of equations similar to (1) above, though possibly with a more general variance structure than that implied by Eqns. (1c) and (1d), and that set would enable us to predict the value of Y_i given only the value of X_i together with an estimate of the ratio β. Laplace's estimate of β was a little over 28.35.

The kind of inference that we have just used is often described as "model-based," but because it is a prediction model and because we shall meet another kind of model very shortly, it is preferable to describe it as "prediction-based," and this is the term that will be used here.

1.4. The randomization approach to survey sampling inference

As already indicated, the other modern approach to survey sampling inference is more subtle, so it will take a little longer to describe. It is convenient to use a reasonably realistic scenario to do so.

The hypothetical country of Oz (which has a great deal more in common with Australia than with Frank L. Baum's mythical Land of Oz) has a population of 20 million people geographically distributed over 10,000 postcodes. These postcodes vary greatly among themselves in population, with much larger numbers of people in a typical urban than in a typical rural postcode.

Oz has a government agency named Centrifuge, which disburses welfare payments widely over the entire country. Its beneficiaries are in various categories such as Age Pensioners, Invalid Pensioners, and University Students. One group of its beneficiaries receives what are called Discretionary Benefits. These are paid to people who do not fall into any of the regular categories but are nevertheless judged to be in need of and/or deserving of financial support.

Centrifuge staff, being human, sometimes mistakenly make payments over and above what their beneficiaries are entitled to. In the Discretionary Benefits category, it is more difficult than usual to determine when such errors (known as overpayments) have been made, so when Centrifuge wanted to arrive at a figure for the amounts of Overpayments to Discretionary Beneficiaries, it decided to do so on a sample basis. Further, since it keeps its records in postcode order, it chose to select 1000 of these at random (one tenth of the total) and to spend considerable time and effort in ensuring that the Overpayments in these sample postcodes were accurately determined. (In what follows, the number of sample postcodes, in this case 1000, will be denoted by n and the number of postcodes in total, in this case 10,000, denoted by N.)

The original intention of the Centrifuge sample designers had been to use the same kind of ratio estimator as Laplace had used in 1802, namely

$$\hat{Y} = \frac{\sum_{i=1}^{N} \delta_i Y_i}{\sum_{i=1}^{N} \delta_i X_i} \sum_{i=1}^{N} X_i, \tag{3}$$

with Y_i being the amount of overpayments in the ith postcode and X_i the corresponding postcode population. In (3), δ_i is a binary (1/0) indicator of inclusion into the sample

of size n: for any particular sample, all but n of the N elements of the population will have a value of $\delta = 0$ so that the sum of $\delta_i Y_i$ over $i = 1 \ldots N$ yields the sum of just the n values of Y_i on those elements selected into the sample.

However, when this proposal came to the attention of a certain senior Centrifuge officer who had a good mathematical education, he queried the use of this ratio estimator on the grounds that the relationship between Overpayments (in this particular category) and Population in individual postcodes was so weak that the use of the model (1) to justify it was extremely precarious. He suggested that the population figures for the selected postcodes should be ignored and that the ratio estimator should be replaced by the simpler expansion estimator, which was

$$\hat{Y} = (N/n) \sum_{i=1}^{N} \delta_i Y_i. \tag{4}$$

When this suggestion was passed on to the survey designers, they saw that it was needed to be treated seriously, but they were still convinced that there was a sufficiently strong relationship between Overpayments and Population for the ratio estimator also to be a serious contender. Before long, one of them found a proof, given in several standard sampling textbooks, that without reliance on any prediction model such as Eqns. (1), the ratio estimator was more efficient than the expansion estimator provided (a) that the sample had been selected randomly from the parent population and (b) that the correlation between the Y_i and the X_i exceeded a certain value (the exact nature of which is irrelevant for the time being). The upshot was that when the sample data became available, that requirement was calculated to be met quite comfortably, and in consequence the ratio estimator was used after all.

1.5. A comparison of these two approaches

The basic lesson to be drawn from the above scenario is that there are two radically different sources of survey sampling inference. The first is prediction on the basis of a mathematical model, of which (1), or something similar to it, is the one most commonly postulated. The other is randomized sampling, which can provide a valid inference regardless of whether the prediction model is a useful one or not. Note that a model can be useful even when it is imperfect. The famous aphorism of G.E.P. Box, "All models are wrong, but some are useful." (Box, 1979), is particularly relevant here.

There are also several other lessons that can be drawn. To begin with, models such as that of Eqns. (1) have parameters. Equation (1a) has the parameter β, and Eq. (1c) has the parameter σ^2 that describes the extent of variability in the Y_i. By contrast, the randomization-based estimator (4) involves no estimation of any parameter. All the quantities on the right hand side of (4), namely N, n, and the sample Y_i, are known, if not without error, at least without the need for any separate estimation or inference.

In consequence, we may say that estimators based on prediction inference are parametric, whereas those based on randomization inference are nonparametric. Parametric estimators tend to be more accurate than nonparametric estimators when the model on which they are based is sufficiently close to the truth as to be useful, but they are also sensitive to the possibility of model breakdown. By contrast, nonparametric estimators tend to be less efficient than parametric ones, but (since there is no model to break

down) they are essentially robust. If an estimator is supported by both parametric and nonparametric inference, it is likely to be both efficient and robust. When the correlation between the sample Y_i and the sample X_i is sufficiently large to meet the relevant condition, mentioned but not defined above in the Oz scenario, the estimator is also likely to be both efficient and robust, but when the correlation fails to meet that condition, another estimator has a better randomization-based support, so the ratio estimator is no longer robust, and the indications are that the expansion estimator, which does not rely upon the usefulness of the prediction model (1), would be preferable.

It could be argued, however, that the expansion estimator itself could be considered as based on the even simpler prediction model

$$Y_i = \alpha + U_i, \tag{5}$$

where the random terms U_i have zero means and zero correlations as before. In this case, the parameter to be estimated is α, and it is optimally estimated by the mean of the sample observations Y_i. However, the parametrization used here is so simple that the parametric estimator based upon it coincides with the nonparametric estimator provided by randomization inference. This coincidence appears to have occasioned some considerable confusion, especially, but not exclusively, in the early days of survey sampling.

Moreover, it is also possible to regard the randomization approach as implying its own quite different model. Suppose we had a sample in which some of the units had been selected with one chance in ten, others with one chance in two, and the remainder with certainty. (Units selected with certainty are often described as "completely enumerated.") We could then make a model of the population from which such a sample had been selected by including in it (a) the units that had been selected with one chance in ten, together with nine exact copies of each such unit, (b) the units that had been selected with one chance in two, together with a single exact copy of each such unit, and (c) the units that had been included with certainty, but in this instance without any copies. Such a model would be a "randomization model." Further, since it would be a nonparametric model, it would be intrinsically robust, even if better models could be built that did use parameters.

In summary, the distinction between parametric prediction inference and nonparametric randomization inference is quite a vital one, and it is important to bear it in mind as we consider below some of the remarkable vicissitudes that have beset the history of survey sampling from its earliest times and have still by no means come to a definitive end.

2. Historical approaches to survey sampling inference

2.1. The development of randomization-based inference

Although, as mentioned above, Laplace had made plans to use the ratio estimator as early as the mid-1780s, modern survey sampling is more usually reckoned as dating from the work of Anders Nicolai Kiaer, the first Director of the Norwegian Central Bureau of Statistics. By 1895, Kiaer, having already conducted sample surveys successfully in his own country for fifteen years or more, had found to his own satisfaction that it was

not always necessary to enumerate an entire population to obtain useful information about it. He decided that it was time to convince his peers of this fact and attempted to do so first at the session of the International Statistical Institute (ISI) that was held in Berne that year. He argued there that what he called a "partial investigation," based on a subset of the population units, could indeed provide such information, provided only that the subset had been carefully chosen to reflect the whole of that population in miniature. He described this process as his "representative method," and he was able to gain some initial support for it, notably from his Scandinavian colleagues. Unfortunately, however, his idea of representation was too subjective and lacking in probabilistic rigor to make headway against the then universally held belief that only complete enumerations, "censuses," could provide any useful information (Lie, 2002; Wright, 2001).

It was nevertheless Kiaer's determined effort to overthrow that universally held belief that emboldened Lucien March, at the ISI's Berlin meeting in 1903, to suggest that randomization might provide an objective basis for such a partial investigation (Wright, 2001). This idea was further developed by Arthur Lyon Bowley, first in a theoretical paper (Bowley, 1906) and later by a practical demonstration of its feasibility in a pioneering survey conducted in Reading, England (Bowley, 1912).

By 1925, the ISI at its Rome meeting was sufficiently convinced (largely by the report of a study that it had itself commissioned) to adopt a resolution giving acceptance to the idea of sampling. However, it was left to the discretion of the investigators whether they should use randomized or purposive sampling. With the advantage of hindsight, we may conjecture that, however vague their awareness of the fact, they were intuiting that purposive sampling was under some circumstances capable of delivering accurate estimates, but that under other circumstances, the underpinning of randomization inference would be required.

In the following year, Bowley published a substantial monograph in which he presented what was then known concerning the purposive and randomizing approaches to sample selection and also made suggestions for further developments in both of them (Bowley, 1926). These included the notion of collecting similar units into groups called "strata," including the same proportion of units from each stratum in the sample, and an attempt to make purposive sampling more rigorous by taking into account the correlations between, on the one hand, the variables of interest for the survey and, on the other, any auxiliary variables that could be helpful in the estimation process.

2.2. Neyman's establishment of a randomization orthodoxy

A few years later, Corrado Gini and Luigi Galvani selected a purposive sample of 29 out of 214 districts (circondari) from the 1921 Italian Population Census (Gini and Galvani, 1929). Their sample was chosen in such a way as to reflect almost exactly the whole-of-Italy average values for seven variables chosen for their importance, but it was shown by Jerzy Neyman (1934) that it exhibited substantial differences from those averages for other important variables.

Neyman went on to attack this study with a three pronged argument. His criticisms may be summarized as follows:

(1) Because randomization had not been used, the investigators had not been able to invoke the Central Limit Theorem. Consequently, they had been unable to use

the normality of the estimates to construct the confidence intervals that Neyman himself had recently invented and which appeared in English for the first time in his 1934 paper.

(2) On the investigators' own admission, the difficulty of achieving their "purposive" requirement (that the sample match the population closely on seven variables) had caused them to limit their attention to the 214 districts rather than to the 8354 communes into which Italy had also been divided. In consequence, their 15% sample consisted of only 29 districts (instead of perhaps 1200 or 1300 communes). Neyman further showed that a considerably more accurate set of estimates could have been expected had the sample consisted of this larger number of smaller units. Regardless of whether the decision to use districts had required the use of purposive sampling, or whether the causation was the other way round, it was evident that purposive sampling and samples consisting of far too few units went hand in hand.

(3) The population model used by the investigators was demonstrably unrealistic and inappropriate. Models by their very nature were always liable to represent the actual situation inadequately. Randomization obviated the need for population modeling.[2] With randomization-based inference, the statistical properties of an estimator are reckoned with respect to the distribution of its estimates from all samples that might possibly be drawn using the design under consideration. The same estimator under different designs will admit to differing statistical properties. For example, an estimator that is unbiased under an equal probability design (see Section 3 of this chapter for an elucidation of various designs that are in common use) may well be biased under an unequal probability design.

In the event, the ideas that Neyman had presented in this paper, though relevant for their time and well presented, caught on only gradually over the course of the next decade. W. Edwards Deming heard Neyman in London in 1936 and soon arranged for him to lecture, and his approach to be taught, to U.S. government statisticians. A crucial event in its acceptance was the use in the 1940 U.S. Population and Housing Census of a one-in-twenty sample designed by Deming, along with Morris Hansen and others, to obtain answers to additional questions. Once accepted, however, Neyman's arguments swept all other considerations aside for at least two decades.

Those twenty odd years were a time of great progress. In the terms introduced by Kuhn (1996), finite population sampling had found a universally accepted "paradigm" (or "disciplinary matrix") in randomization-based inference, and an unusually fruitful period of normal science had ensued. Several influential sampling textbooks were published, including most importantly those by Hansen et al. (1953) and by Cochran (1953, 1963). Other advances included the use of self-weighting, multistage, unequal probability samples by Hansen and Hurwitz at the U.S. Bureau of the Census, Mahalanobis's invention of interpenetrating samples to simplify the estimation of variance for complex survey designs and to measure and control the incidence of nonsampling errors, and the beginnings of what later came to be described as "model-assisted survey sampling."

[2] The model of Eqns. (1) above had not been published at the time of Neyman's presentation. It is believed first to have appeared in Fairfield Smith (1938) in the context of a survey of agricultural crops. Another early example of its use is in Jessen (1942).

A lone challenge to this orthodoxy was voiced by Godambe (1955) with his proof of the nonexistence of any uniformly best randomization-based estimator of the population mean, but few others working in this excitingly innovative field seemed to be concerned by this result.

2.3. Model-assisted or model-based? The controversy over prediction inference

It therefore came as a considerable shock to the finite population sampling establishment when Royall (1970b) issued his highly readable call to arms for the reinstatement of purposive sampling and prediction-based inference. To read this paper was to read Neyman (1934) being stood on its head. The identical issues were being considered, but the opposite conclusions were being drawn.

By 1973, Royall had abandoned the most extreme of his recommendations. This was that the best sample to select would be the one that was optimal in terms of a model closely resembling Eqns. (1). (That sample would typically have consisted of the largest n units in the population, asking for trouble if the parameter β had not in fact been constant over the entire range of sizes of the population units.) In Royall and Herson (1973a and 1973b), the authors suggested instead that the sample should be chosen to be "balanced", in other words that the moments of the sample X_i should be as close as possible to the corresponding moments of the entire population. (This was very similar to the much earlier notion that samples should be chosen purposively to resemble the population in miniature, and the samples of Gini and Galvani (1929) had been chosen in much that same way!)

With that exception, Royall's original stand remained unshaken. The business of a sampling statistician was to make a model of the relevant population, design a sample to estimate its parameters, and make all inferences regarding that population in terms of those parameter estimates. The randomization-based concept of defining the variance of an estimator in terms of the variability of its estimates over all possible samples was to be discarded in favor of the prediction variance, which was sample-specific and based on averaging over all possible realizations of the chosen prediction model.

Sampling statisticians had at no stage been slow to take sides in this debate. Now the battle lines were drawn. The heat of the argument appears to have been exacerbated by language blocks; for instance, the words "expectation" and "variance" carried one set of connotations for randomization-based inference and quite another for prediction-based inference. Assertions made on one side would therefore have appeared as unintelligible nonsense by the other.

A major establishment counterattack was launched with Hansen et al. (1983). A small (and by most standards undetectable) divergence from Royall's model was shown nevertheless to be capable of distorting the sample inferences substantially. The obvious answer would surely have been "But this distortion would not have occurred if the sample had been drawn in a balanced fashion. Haven't you read Royall and Herson (1973a and b)?" Strangely, it does not seem to have been presented at the time.

Much later, a third position was also offered, the one held by the present authors, namely that since there were merits in both approaches, and that it was possible to combine them, the two should be used together. For the purposes of this Handbook volume, it is necessary to consider all three positions as dispassionately as possible. Much can be gained by asking the question as to whether Neyman (1934) or Royall

(1970b) provided the more credible interpretation of the facts, both as they existed in 1934 or 1970 and also at the present day (2009).

2.4. A closer look at Neyman's criticisms of Gini and Galvani

The proposition will be presented here that Neyman's criticisms and prescriptions were appropriate for his time, but that they have been overtaken by events. Consider first his contention that without randomization, it was impossible to use confidence intervals to measure the accuracy of the sample estimates.

This argument was received coolly enough at the time. In moving the vote of thanks to Neyman at the time of the paper's presentation, Bowley wondered aloud whether confidence intervals were a "confidence trick." He asked "Does [a confidence interval] really lead us to what we need—the chance that within the universe which we are sampling the proportion is within these certain limits? I think it does not. I think we are in the position of knowing that *either* an improbable event had occurred *or* the proportion in the population is within these limits. . . The statement of the theory is not convincing, and until I am convinced I am doubtful of its validity."

In his reply, Neyman pointed out that Bowley's question in the first quoted sentence above "contain[ed] the statement of the problem in the form of Bayes" and that in consequence its solution "*must* depend upon the probability law *a priori*." He added "In so far as we keep to the old form of the problem, any further progress is impossible." He thus concluded that there was a need to stop asking Bowley's "Bayesian" question and instead adopt the stance that the "*either. . .or*" statement contained in his second quoted sentence "form[ed] a basis for the practical work of a statistician concerned with problems of estimation." There can be little doubt but that Neyman's suggestion was a useful prescription for the time, and the enormous amount of valuable work that has since been done using Neyman and Pearson's confidence intervals is witness to this.

However, the fact remains that confidence intervals are not easy to understand. A confidence interval is in fact a sample-specific range of potentially true values of the parameter being estimated, which has been constructed so as to have a particular property. This property is that, over a large number of sample observations, the proportion of times that the true parameter value falls inside that range (constructed for each sample separately) is equal to a predetermined value known as the confidence level. This confidence level is conventionally written as $(1 - \alpha)$, where α is small compared with unity. Conventional choices for α are 0.05, 0.01, and sometimes 0.001. Thus, if many samples of size n are drawn independently from a normal distribution and the relevant confidence interval for $\alpha = 0.05$ is calculated for each sample, the proportion of times that the true parameter value will lie within any given sample's own confidence interval will, before that sample is selected, be 0.95, or 95%.

It is not the case, however, that the probability of this true parameter value lying within the confidence interval as calculated for any individual sample of size n will be 95%. The confidence interval calculated for any individual sample of size n will, in general, be wider or narrower than average and might be centered well away from the true parameter value, especially if n is small. It is also sometimes possible to recognize when a sample is atypical and, hence, make the informed guess that in this particular case, the probability of the true value lying in a particular 95% confidence interval differs substantially from 0.95.

If, however, an agreement is made beforehand that a long succession of wagers is to be made on the basis that (say) Fred will give Harry $1 every time the true value lies inside any random sample's properly calculated 95% confidence interval, and Harry will give Fred $19 each time it does not; then at the end of that long sequence, those two gamblers would be back close to where they started. In those circumstances, the 95% confidence interval would also be identical with the 95% Bayesian credibility interval that would be obtained with a flat prior distribution over the entire real line ranging from minus infinity to plus infinity. In that instance, Bowley's "Bayesian question" could be given an unequivocally affirmative answer.

The result of one type of classical hypothesis test is also closely related to the confidence interval. Hypothesis tests are seldom applied to data obtained from household or establishment surveys, but they are frequently used in other survey sampling contexts.

The type of classical test contemplated here is often used in medical trials. The hypothesis to be tested is that a newly devised medical treatment is superior to an existing standard treatment, for which the effectiveness is known without appreciable error. In this situation, there can never be any reason to imagine that the two treatments are identically effective so that event can unquestionably be accorded the probability zero. The probability that the alternative treatment is the better one can then legitimately be estimated by the proportion of the area under the likelihood function that corresponds to values greater than the standard treatment's effectiveness. Moreover, if that standard effectiveness happens to be lower than that at the lower end of the one-sided 95% confidence interval, it can reasonably be claimed that the new treatment is superior to the standard one "with 95% confidence."

However, in that situation, the investigators might well wish to go further and quote the proportion of the area corresponding to all values less than standard treatment's effectiveness (Fisher's p-statistic). If, for instance, that proportion were 0.015, they might wish to claim that the new treatment was superior "with 98.5% confidence." To do so might invite the objection that the language used was inappropriate because Neyman's α was an arbitrarily chosen fixed value, whereas Fisher's p was a realization of a random variable, but the close similarity between the two situations would be undeniable. For further discussions of this distinction, see Hubbard and Bayarri (2003) and Berger (2003).

The situation would have been entirely different, however, had the investigation been directed to the question as to whether an additional parameter was required for a given regression model to be realistic. Such questions often arise in contexts such as biodiversity surveys and sociological studies. It is then necessary to accord the null hypothesis value itself (which is usually but not always zero) a nonzero probability. It is becoming increasingly well recognized that in these circumstances, the face value of Fisher's p can give a grossly misleading estimate of the probability that an additional parameter is needed. A relatively new concept, the "false discovery rate" (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Efron et al., 2001; Sorić, 1989), can be used to provide useful insights. To summarize the findings in these papers very briefly, those false discovery rates observed empirically have, more often than not, been found to exceed the corresponding p-statistic by a considerable order of magnitude.

It is also relevant to mention that the populations met with in finite population sampling, and especially those encountered in establishment surveys, are often far removed

from obeying a normal distribution, and that with the smaller samples often selected from them, the assumption of normality for the consequent estimators is unlikely even to produce accurate confidence intervals!

Nevertheless, and despite the misgivings presented above, it is still the case that randomization does provide a useful basis for the estimation of a sample variance. The criterion of minimizing that variance is also a useful one for determining optimum estimators. However, we should not expect randomization alone to provide anything further.

Neyman's second contention was that purposive sampling and samples consisting of fewer than an adequate number of units went hand in hand. This was undoubtedly the case in the 1930s, but a similar kind of matching of sample to population (Royall and his co-authors use the expression "balanced sampling") can now be undertaken quite rapidly using third-generation computers, provided only that the matching is not made on too many variables simultaneously. Brewer (1999a) presents a case that it might be preferable to choose a sample randomly and use calibrated estimators to compensate for any lack of balance, rather than to go to the trouble of selecting balanced samples. However, those who prefer to use balanced sampling can now select randomly from among many balanced or nearly balanced samples using the "cube method" (Deville and Tillé, 2004). This paper also contains several references to earlier methods for selecting balanced samples.

Neyman's third contention was basically that population models were not to be trusted. It is difficult here to improve on the earlier quote from George Box that "All models are wrong, but some models are useful." Equations (1) above provide a very simple model that has been in use since 1938. It relates a variable of interest in a sample survey to an auxiliary variable, all the population values of which are conveniently known.

In its simplest form, the relationship between these variables is assumed to be basically proportional but with a random term modifying that proportional relationship for each population unit. (Admittedly, in some instances, it is convenient to add an intercept term, or to have more than one regressor variable, and/or an additional equation to model the variance of that equation's random term, but nevertheless that simple model can be adequate in a remarkably wide set of circumstances.)

As previously mentioned, such models have been used quite frequently in survey sampling. However, it is one thing to use a prediction model to improve on an existing randomization-based estimator (as was done in the Oz scenario above) and it is quite another thing actually to base one's sampling inference on that model. The former, or "model-assisted" approach to survey sampling inference, is clearly distinguished from prediction-based inference proper in the following quotation, taken from the Preface to the encyclopedic book, *Model Assisted Survey Sampling* by Särndal et al. (1992, also available in paperbook 2003):

> Statistical modeling has strongly influenced survey sampling theory in recent years. In this book, sampling theory is assisted by modeling. It becomes simple to explain how the auxiliary information in a given survey will lead to a particular estimation technique. The teaching of sampling and the style of presentation in journal articles have changed a great deal by this new emphasis. Readers of this book will become familiar with this new style.

> We use the randomization theory or design-based point of view. This is the traditional mode of inference in surveys, ever since the sampling breakthroughs in the 1930s and 1940s. The reasoning is familiar to survey statisticians in government and elsewhere.

As this quotation indicates, using a prediction model to form an estimator as Royall proposed, without regard to any justification in terms of randomization theory, is quite a different approach. It is often described as "model-based," or pejoratively as "model-dependent," but it appears preferable to use the expression, "prediction-based."

A seminal paper attacking the use of a prediction model for such purposes was that by Hansen et al. (1983), which has already been mentioned; but there can be no serious doubt attached to the proposition that this model provides a reasonable first approximation to many real situations. Once again, Neyman's contention has been overtaken by events.

2.5. Other recent developments in sample survey inference

A similarly detailed assessment of the now classic papers written by Royall and his colleagues in the 1970s and early 1980s is less necessary, since there have been fewer changes since they were written, but it is worth providing a short summary of some of them. Royall (1970b) has already been mentioned as having turned Neyman (1934) on its head. Royall (1971) took the same arguments a stage further. In Royall and Herson (1973a and 1973b), there is an implicit admission that selecting the sample that minimized the prediction-based variance (prediction variance) was not a viable strategy. The suggestion offered there is to select balanced samples instead: ones that reflect the moments of the parent population. In this recommendation, it recalls the early twentieth-century preoccupation with finding a sample that resembled the population in miniature but, as has been indicated above, this does not necessarily count against it.

Royall (1976) provides a useful and entertaining introduction to prediction-based inference, written at a time when the early criticisms of it had been fully taken into account. Joint papers by Royall and Eberhardt (1975) and Royall and Cumberland (1978, 1981a and 1981b) deal with various aspects of prediction variance estimation, whereas Cumberland and Royall (1981) offer a prediction-based consideration of unequal probability sampling. The book by Valliant et al. (2000) provides a comprehensive account of survey sampling from the prediction-based viewpoint up to that date, and that by Bolfarine and Zacks (1992) presents a Bayesian perspective on it.

Significant contributions have also been made by other authors. Bardsley and Chambers (1984) offered ridge regression as an alternative to pure calibration when the number of regressor variables was substantial. Chambers and Dunstan (1986) and Chambers et al. (1992) considered the estimation of distribution functions from a prediction-based standpoint. Chambers et al. (1993) and Chambers and Kokic (1993) deal specifically with questions of robustness against model breakdown. A more considerable bibliography of important papers relating to prediction-inference can be found in Valliant et al. (2000).

The randomization-based literature over recent years has been far too extensive to reference in the same detail, and in any case comparatively little of it deals with the question of sampling inference. However, two publications already mentioned above

are of especial importance. These are the polemical paper by Hansen et al. (1983) and the highly influential text-book by Särndal et al. (1992), which sets out explicitly to indicate what can be achieved by using model-assisted methods of sample estimation without the explicit use of prediction-based inference. Other recent papers of particular interest in this field include Deville and Särndal (1992) and Deville et al. (1993).

Publications advocating or even mentioning the use of both forms of inference simultaneously are few in number. Brewer (1994) would seem to be the earliest to appear in print. It was written in anticipation of and to improve upon Brewer (1995), which faithfully records what the author was advocating at the First International Conference on Establishment Surveys in 1993, but was subsequently found not to be as efficient or even as workable as the alternative provided in Brewer (1994). A few years later, Brewer (1999a) compared stratified balanced with stratified random sampling and Brewer (1999b) provided a detailed description of how the two inferences could be used simultaneously in unequal probability sampling; also Brewer's (2002) textbook has provided yet further details on this topic, including some unsought spin-offs that follow from their simultaneous use, and an extension to multistage sampling.

All three views are still held. The establishment view is that model-assisted randomization-based inference has worked well for several decades, and there is insufficient reason to change. The prediction-based approach continues to be presented by others as the only one that can consistently be held by a well-educated statistician. And a few say "Why not use both?" Only time and experience are likely to resolve the issue, but in the meantime, all three views need to be clearly understood.

3. Some common sampling strategies

3.1. Some ground-clearing definitions

So far, we have only been broadly considering the options that the sampling statistician has when making inferences from the sample to the population from which it was drawn. It is now time to consider the specifics, and for that we will need to use certain definitions.

A *sample design* is a procedure for selecting a sample from a population in a specific fashion. These are some examples:

- simple random sampling with and without replacement;
- random sampling with unequal probabilities, again with and without replacement;
- systematic sampling with equal or unequal probabilities;
- stratified sampling, in which the population units are first classified into groups or "strata" having certain properties in common;
- two-phase sampling, in which a large sample is drawn at the first phase and a subsample from that large sample at the second phase;
- multistage sampling, usually in the context of area sampling, in which a sample of (necessarily large) first-stage units is selected first, samples within those first-stage sample units at the second stage, and so on for possibly third and fourth stages; and
- permanent random number sampling, in which each population unit is assigned a number, and the sample at any time is defined in terms of the ranges of those permanent random numbers that are to be in sample at that time.

This list is not exhaustive, and any given sample may have more than one of those characteristics. For instance, a sample could be of three stages, with stratification and unequal probability sampling at the first stage, unstratified unequal probability sampling at the second stage, and systematic random sampling with equal probabilities at the third stage. Subsequently, subsamples could be drawn from that sample, converting it into a multiphase multistage sample design.

A *sample estimate* is a statistic produced using sample data that can give users an indication as to the value of a population quantity. Special attention will be paid in this section to estimates of population total and population mean because these loom so large in the responsibilities of national statistical offices, but there are many sample surveys that have more ambitious objectives and may be set up so as to estimate small domain totals, regression and/or correlation coefficients, measures of dispersion, or even conceivably coefficients of heteroskedasticity (measures of the extent to which the variance of the U_i can itself vary with the size of the auxiliary variable X_i).

A *sample estimator* is a prescription, usually a mathematical formula, indicating how estimates of population quantities are to be obtained from the sample survey data.

An *estimation procedure* is a specification as to what sample estimators are to be used in a given sample survey.

A *sample strategy* is a combination of a sample design and an estimation procedure. Given a specific sample strategy, it is possible to work out what estimates can be produced and how accurately those estimates can be made.

One consequence of the fact that two quite disparate inferential approaches can be used to form survey estimators is that considerable care needs to be taken in the choice of notation. In statistical practice generally, random variables are represented by uppercase symbols and fixed numbers by lowercase symbols, but between the two approaches, an observed value automatically changes its status. Specifically, in both approaches, a sample value can be represented as the product of a population value and the inclusion indicator, δ, which was introduced in (3). However, in the prediction-based approach, the population value is a random variable and the inclusion indicator is a fixed number, whereas in the randomization-based approach, it is the inclusion indicator that is the random variable while the population value is a fixed number. There is no ideal way to resolve this notational problem, but we shall continue to denote population values by, say, Y_i or X_i and sample values by $\delta_i Y_i$ or $\delta_i X_i$, as we did in Eq. (3).

3.2. Equal probability sampling with the expansion estimator

In what follows, the sample strategies will first be presented in the context of randomization-based inference, then that of the nearest equivalent in prediction-based inference, and finally, wherever appropriate, there will be a note as to how they can be combined.

3.2.1. Simple random sampling with replacement using the expansion estimator

From a randomization-based standpoint, simple random sampling with replacement (srswr) is the simplest of all selection procedures. It is appropriate for use where (a) the

population consists of units whose sizes are not themselves known, but are known not to differ too greatly amongst themselves, and (b) it has no geographical or hierarchical structure that might be useful for stratification or area sampling purposes. Examples are populations of easily accessible individuals or households, administrative records relating to individuals, households, or family businesses; and franchise holders in a large franchise.

The number of population units is assumed known, say N, and a sample is selected by drawing a single unit from this population, completely at random, n times. Each time a unit is drawn, its identity is recorded, and the unit so drawn is returned to the population so that it stands exactly the same chance of being selected at any subsequent draw as it did at the first draw. At the end of the n draws, the ith population unit appears in the sample ν_i times, where ν_i is a number between 0 and n, and the sum of the ν_i over the population is n.

The typical survey variable value on the ith population unit may be denoted by Y_i. The population total of the Y_i may be written Y. A randomization-unbiased estimator of Y is the *expansion estimator*, namely $\hat{Y} = (N/n)\sum_{i=1}^{N} \nu_i Y_i$. (To form the corresponding randomization-unbiased estimator of the population mean, $\bar{Y} = Y/N$, replace the expression N/n in this paragraph by $1/n$.)

The randomization variance of the estimator $\hat{Y}$ is $V(\hat{Y}) = (N^2/n)S_{wr}^2$, where $S_{wr}^2 = N^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. $V(\hat{Y})$ is in turn estimated randomization-unbiasedly by $(N^2/n)\hat{S}_{wr}^2$, where $\hat{S}_{wr}^2 = N^{-1}\sum_{i=1}^{N}\nu_i(Y_i - \bar{Y})^2$. (To form the corresponding expressions for the population mean, replace the expression N^2/n throughout this paragraph by $1/n$. Since these changes from population total to population mean are fairly obvious, they will not be repeated for other sampling strategies.) Full derivations of these formulae will be found in most sampling textbooks.

There is no simple prediction-based counterpart to srswr. From the point of view of prediction-based inference, multiple appearances of a population unit add no information additional to that provided by the first appearance. Even from the randomization standpoint, srswr is seldom called for, as simple random sampling without replacement (or srswor) is more efficient. Simple random sampling with replacement is considered here purely on account of its extremely simple randomization variance and variance estimator, and because (by comparison with it) both the extra efficiency of srswor and the extra complications involved in its use can be readily appreciated.

3.2.2. Simple random sampling without replacement using the expansion estimator

This sample design is identical with srswr, except that instead of allowing selected population units to be selected again at later draws, units already selected are given no subsequent probabilities of selection. In consequence, the units not yet selected have higher conditional probabilities of being selected at later draws. Because the expected number of distinct units included in sample is always n (the maximum possible number under srswr), the srswor estimators of population total and mean have smaller variances than their srswr counterparts. A randomization-unbiased estimator of Y is again $\hat{Y} = (N/n)\sum_{i=1}^{N}\nu_i Y_i$, but since under srswor the ν_i take only the values 0 and 1, it will be convenient hereafter to use a different symbol, δ_i, in its place.

The randomization variance of the estimator $\hat{Y}$ is $V(\hat{Y}) = (N-n)(N/n)S^2$, where $S^2 = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$. The variance estimator $V(\hat{Y})$ is in turn estimated

randomization-unbiasedly by $(N-n)(N/n)\hat{S}^2$, where $\hat{S}^2 = (n-1)^{-1}\sum_{i=1}^{N}\delta_i (Y_i - \hat{\bar{Y}})^2$. The substitution of the factor N^2 (in the srswr formulae for the variance and the unbiased variance estimator) by the factor $N(N-n)$ (in the corresponding srswor formulae) is indicative of the extent to which the use of sampling without replacement reduces the variance.

Note, however, that the sampling fraction, n/N, is not particularly influential in reducing the variance, even for srswor, unless n/N is an appreciable fraction of unity. An estimate of a proportion obtained from an srswor sample of 3000 people in, say, Wales, is not appreciably any more accurate than the corresponding estimate obtained from a sample of 3000 people in the United States; and this is despite the proportion of Welsh people in the first sample being about 1 in 1000 and the proportion of Americans in the second being only 1 in 100,000. For thin samples like these, such variances are to all intents and purposes inversely proportional to the sample size, and the percentage standard errors are inversely proportional to the square root of the sample size. Full derivations of these formulae will be again be found in most sampling textbooks.

Since srswor is both more efficient and more convenient than srswr, it will be assumed, from this point on, that sampling is without replacement unless otherwise specified. One important variant on srswor, which also results in sampling without replacement, is systematic sampling with equal probabilities, and this is the next sampling design that will be considered.

3.2.3. Systematic sampling with equal probabilities, using the expansion estimator

Systematic sampling, by definition, is the selection of sample units from a comprehensive list using a constant skip interval between neighboring selections. If, for instance, the skip interval is 10, then one possible systematic sample from a population of 104 would consist of the second unit in order, then the 12th, the 22nd, etc. up to and including the 102nd unit in order. This sample would be selected if the starting point (usually chosen randomly as a number between 1 and the skip interval) was chosen to be 2. The sample size would then be 11 units with probability 0.4 and 10 units with probability 0.6, and the expected sample size would be 10.4, or more generally the population size divided by the skip interval.

There are two important subcases of such systematic selection. The first is where the population is deliberately randomized in order prior to selection. The only substantial difference between this kind of systematic selection and srswor is that in the latter case, the sample size is fixed, whereas in the former it is a random variable. Even from the strictest possible randomization standpoint, however, it is possible to consider the selection procedure as conditioned on the selection of the particular random start (in this case 2), in which case the sample size would be fixed at 10 and the srswor theory would then hold without any modification. This conditional randomization theory is used very commonly, and from a model-assisted point of view it is totally acceptable.

That is emphatically not true, however, for the second subcase, where the population is not deliberately randomized in order prior to selection. Randomization theory in that subcase is not appropriate and it could be quite dangerous to apply it. In an extreme case, the 104 units could be soldiers, and every 10th one from the 3rd onwards could be a sergeant, the remainder being privates. In that case, the sample selected above

would consist entirely of privates, and if the random start had been three rather than two, the sample would have been entirely one of sergeants. This, however, is a rare and easily detectable situation within this nonrandomized subcase. A more likely situation would be one where the population had been ordered according to some informative characteristic, such as age. In that instance, the sample would in one sense be a highly desirable one, reflecting the age distribution of the population better than by chance. That would be the kind of sample that the early pioneers of survey sampling would have been seeking with their purposive sampling, one that reflected in miniature the properties of the population as a whole.

From the randomization standpoint, however, that sample would have had two defects, one obvious and one rather more subtle. Consider a sample survey aimed at estimating the level of health in the population of 104 persons as a whole. The obvious defect would be that although the obvious estimate based on the systematic sample would reflect that level considerably more accurately than one based on a random sample would have done, the randomization-based estimate of its variance would not provide an appropriate measure of its accuracy.

The more subtle defect is that the randomization-based estimate of its variance would in fact tend to overestimate even what the variance would have been if a randomized sample had been selected. So the systematic sample would tend to reduce the actual variance but slightly inflate the estimated variance! (This last point is indeed a subtle one, and most readers should not worry if they are not able to work out why this should be. It has to do with the fact that the average squared distance between sample units is slightly greater for a systematic sample than it is for a purely random sample.)

In summary, then, systematic sampling is temptingly easy to use and in most cases will yield a better estimate than a purely randomized sample of the same size, but the estimated variance would not reflect this betterment, and in some instances a systematic sample could produce a radically unsuitable and misleading sample. To be on the safe side, therefore, it would be advisable to randomize the order of the population units before selection and to use the srswor theory to analyze the sample.

3.2.4. Simple prediction inference using the expansion estimator

Simple random sampling without replacement does have a prediction-based counterpart. The appropriate prediction model is the special case of Eqns. (1) in which all the X_i take the value unity. The prediction variances of the U_i in (1c) are in this instance all the same, at σ^2. Because this very simple model is being taken as an accurate refection of reality, it would not matter, in theory, how the sample was selected. It could (to take the extreme case) be a "convenience sample" consisting of all the people in the relevant defined category whom the survey investigator happened to know personally, but of course, in practice, the use of such a "convenience sample" would make the assumptions underlying the equality of the X_i very hard to accept. It would be much more convincing if the sample were chosen randomly from a carefully compiled list, which would then be an srswor sample, and it is not surprising that the formulae relevant to this form of prediction sampling inference should be virtually identical to those for randomization sampling srswor.

The minimum-variance prediction-unbiased estimator of Y under the simple prediction model described in the previous paragraph is identical with the randomization-unbiased estimator under srswor, namely $\hat{Y} = (N/n) \sum_{i=1}^{N} \delta_i Y_i$. Further,

the prediction variance of $\hat{Y}$ is $V(\hat{Y}) = (N-n)(N/n)\sigma^2$. A prediction-unbiased estimator of $V(\hat{Y})$ is $\hat{V}(\hat{Y}) = (N-n)(N/n)\hat{\sigma}^2$, where $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^{N}(\delta_i Y_i - \hat{\bar{Y}})^2$ where $\hat{\bar{Y}} = \hat{Y}/N$. Note that although the prediction variance is typically sample-specific, in this instance it is the same for all samples. However, the estimated prediction variance does, as always, vary from sample to sample.

3.3. Equal probability sampling with the ratio estimator

So far, we have been using estimators that depend only on the sample observations Y_i themselves. More often than not, however, the sampling statistician has at hand relevant auxiliary information regarding most of the units in the population. We have already noted that Laplace, back at the turn of the 19th century, had access (at least in principle) to annual birth registration figures that were approximately proportional to the population figures that he was attempting to estimate. To take a typical modern example, the population for a Survey of Retail Establishments (shops) would typically consist mainly of shops that had already been in existence at the time of the most recent complete Census of Retail Establishments, and the principal information collected at that Census would have been the sales figures for the previous calendar or financial year. Current sales would, for most establishments and for a reasonable period, remain approximately proportional to those Census sales figures.

Returning to the model of Eqns. (1), we may equate the Y_i with the current sales of the sample establishments, the X_i with the Census sales of the sample and nonsample establishments, and the X with the total Census sales over all sample and nonsample establishments combined. It may be remembered that "Centrifuge's" ratio estimators worked well both when the model of Eqns. (1) was a useful one and also in the weaker situation when there was a comparatively modest correlation between the Y_i and the X_i. In a similar fashion, the corresponding ratio estimator for this Survey of Retail Establishments tends to outperform the corresponding expansion estimator, at least until it is time to conduct the next Census of Retail Establishments, which would typically be some time in the next 5–10 years.

It was stated above that the population for a Census of Retail Establishments would typically consist mainly of shops that had already been in existence at the time of the most recent complete Census. Such shops would make up the "Main Subuniverse" for the survey. In practice, there would usually be a substantial minority of shops of which the existence would be known, but which had not been in business at the time of that Census, and for these there would be a separate "New Business Subuniverse," which for want of a suitable auxiliary variable would need to be estimated using an expansion estimator, and in times of rapid growth there might even be an "Unlisted New Business Provision" to allow for the sales of shops that were so new that their existence was merely inferred on the basis of previous experience. Nevertheless, even then, the main core of the estimate of survey period sales would still be the sales of shops in the Main Subuniverse, these sales would be based on Ratio Estimation, and the relevant Ratio Estimator would be the product of the $\hat{\beta}$ of Eq. (2) and the Total of Census Sales X.

The modern way of estimating the variance of that ratio estimator depends on whether the relevant variance to be estimated is the randomization variance, which is based on

the variability of the estimates over all possible samples, or whether it is the prediction variance, which is sample specific. (For a discussion of the difference between the randomization and prediction approaches to inference, the reader may wish to refer back to Sections 1.3 and 1.4.)

The most common practice at present is to estimate the randomization-variance, and for that the procedure is as follows: denote the population total of the Y_i by Y, its expansion estimator by $\hat{Y}$, and its ratio estimator by $\hat{Y}_R$. Then the randomization variance of $\hat{Y}_R$ is approximated by

$$V(\hat{Y}_R) \approx V(\hat{Y}) + \beta^2 V(\hat{X}) - 2\beta C(\hat{Y}, \hat{X}), \tag{6}$$

where β is the same parameter as in Eq. (1a), $V(\hat{Y})$ is the randomization variance of the expansion estimator of Y, $V(\hat{X})$ is the variance of the corresponding expansion estimator of X, based on the same sample size, and $C(\hat{Y}, \hat{X})$ is the covariance between those two estimators.

The approximate randomization-variance of $\hat{Y}_R$ can therefore be estimated by

$$\hat{V}(\hat{Y}_R) = \hat{V}(\hat{Y}) + \hat{\beta}^2 \hat{V}(\hat{X}) - 2\hat{\beta}\hat{C}(\hat{Y}, \hat{X}), \tag{7}$$

where $\hat{V}(\hat{Y})$ is the randomization-unbiased estimator of $V(\hat{Y})$, given in Subsection 3.2.2, $\hat{V}(\hat{X})$ is the corresponding expression in the X-variable, $\hat{C}(\hat{Y}, \hat{X})$ is the corresponding expression for the randomization-unbiased estimator of covariance between them, namely $(N-n)(N/n)\sum_{i=1}^{N} \delta_i (Y_i - \bar{Y})(X_i - \bar{X})$, and $\hat{\beta}$ is the sample estimator of β, as given in Eq. (2).

3.4. Simple balanced sampling with the expansion estimator

An alternative to simple random sampling is simple balanced sampling, which has already been referred to in Section 2.3. When the sample has been selected in such a way as to be balanced on the auxiliary variables X_i, in the way described in that Section, the expansion estimator is comparable in accuracy to that Section's ratio estimator itself. This is because the expansion estimator based on the balanced sample is then "calibrated" on those X_i. That is to say, the expansion estimate of the total X is necessarily without error; it is exactly equal to X. It is easy to see that in the situation described in the previous subsection, $\hat{Y}_R$ was similarly "calibrated" on the X_i, that is, $\hat{X}_R$ would have been exactly equal to X.

It is a matter of some contention as to whether it is preferable to use simple random sampling and the ratio estimator or simple balanced sampling and the expansion estimator. The choice is basically between a simple selection procedure and a relatively complex estimator on the one hand and a simple estimator with a relatively complex selection procedure on the other. The choice is considered at length in Brewer (1999a). It depends crucially on the prior choice of sampling inference. Those who hold exclusively to randomization for this purpose would necessarily prefer the ratio estimation option. It is only those who are prepared to accept prediction inference, either as an alternative or exclusively, for whom the choice between the two strategies described above would be a matter of taste.

For a further discussion of balanced sampling, see Sections 2.3 and 2.4.

3.5. Stratified random sampling with equal inclusion probabilities within strata

If any kind of supplementary information is available that enables population units to be grouped together in such a way that they are reasonably similar within their groups and reasonably different from group to group, it will usually pay to treat these groups as separate subpopulations, or *strata*, and obtain estimates from each stratum separately. Examples of such groups include males and females, different descriptions of retail outlets (grocers, butchers, other food and drink, clothing, footwear, hardware, etc.), industries of nonretail businesses, dwellings in urban and in rural areas, or in metropolitan and nonmetropolitan areas.

It takes a great deal of similarity to obtain a poorer estimate by stratification, and the resulting increase in variance is almost always trivial, so the default rule is "Use all the relevant information that you have. When in doubt, stratify." There are, however, several exceptions to this rule.

The first is that if there are many such groups, and all the differences between all possible pairs of groups are known to be small, there is little to gain by stratification, and the business of dealing with lots of little strata might itself amount to an appreciable increase in effort. However, this is an extreme situation, so in most cases, it is safer to stick with the default rule. (In any case, do not worry. Experience will gradually give you the feel as to when to stratify and when not to do so.)

The remaining exceptions all relate to stratification by size. Size is an awkward criterion to stratify on because the boundaries between size strata are so obviously arbitrary. If stratification by size has already been decided upon, one useful rule of thumb is that size boundaries such as "under 10,000," "10,000–19,999," "20,000–49,999," "50,000–99,999," "100,000–199,999," and "over 200,000" (with appropriate adjustments to take account of the scale in which the units are measured) are difficult to improve on appreciably. Moreover, there is unlikely to be much gain in forming more than about six size strata.

Another useful rule of thumb is that each stratum should be of about the same order of magnitude in its total measure of size. This rule can be particularly helpful in choosing the boundary between the lowest two and that between the highest two strata. Dalenius (1957) does give formulae that enable optimum boundaries between size strata to be determined, but they are not recommended for general use, partly because they are complicated to apply and partly because rules of thumb and common sense will get sufficiently close to a very flat optimum. A more modern approach may be found in Lavallée and Hidiroglou (1988).

Finally, there is one situation where it might very well pay not to stratify by size at all, and that is where PRN sampling is being used. This situation will be seen later (in Section 3.9).

3.5.1. Neyman and optimal allocations of sample units to strata

Another important feature of stratification is that once the strata themselves have been defined, there are some simple rules for allocating the sample size efficiently among them. One is "Neyman allocation," which is another piece of sampling methodology recommended by Neyman in his famous 1934 paper that has already been mentioned several times. The other, usually known as "Optimum allocation," is similar to Neyman

allocation but also allows for the possibility that the cost of observing the value of a sample unit can differ from stratum to stratum.

Neyman allocation minimizes the variance of a sample estimate subject to a given total sample size.[3] Basically, the allocation of sample units to a stratum h should be proportional to $N_h S_h$, where N_h is the number of population units in the hth stratum and S_h is the relevant population standard deviation in that stratum.[4]

Optimum allocation is not very different. It minimizes the variance of a sample estimate subject to a given total cost and consequently allocates units in a stratum to sample proportionally to $N_h S_h / \sqrt{C_h}$, where C_h is the cost of obtaining the value Y_i for a single sample unit in the hth stratum. Since, however, it is typically more difficult to gather data from small businesses than from large ones, the effect of using Optimal rather than Neyman allocation for business surveys is to concentrate the sample toward the larger units.

Strangely, Optimum allocation seems seldom to have been used in survey practice. This is partly, perhaps, because it complicates the sample design, partly because (for any given level of accuracy) it results in the selection of a larger sample, and partly because it is not often known how much more expensive it is to collect data from smaller businesses.

3.5.2. Stratification with ratio estimation

Since the effect of stratification is effectively to divide the population into a number of subpopulations, each of which can be sampled from and estimated for separately, it is theoretically possible to choose a different selection procedure and a different estimator for each stratum. However, the arguments for using a particular selection procedure and a particular estimator are usually much the same for each stratum, so this complication seldom arises.

A more important question that does frequently arise is whether or not there is any point in combining strata for estimation purposes. This leads to the distinction between "stratum-by-stratum estimation" (also known as "separate stratum estimation") and "across-stratum estimation" (also known as "combined stratum estimation"), which will be the principal topic of this subsection.

The more straightforward of these two options is stratum-by-stratum estimation, in which each stratum is regarded as a separate subpopulation, to which the observations in other strata are irrelevant. The problem with this approach, however, is that in the randomization approach the ratio estimator is biased, and the importance of that bias, relative to the corresponding standard error, can be large when the sample size is small. It is customary in some statistical offices to set a minimum (say six) to the sample size for any stratum, but even for samples of six, it is possible for the randomization bias

[3] We are indebted to Gad Nathan for his discovery that Tschuprow (or Chuprov) had actually published the same result in 1923, but his result was buried in a heap of less useful mathematics. Also, it was Neyman who brought it into prominence, and he would presumably have devised it independently of Tschuprow in any case.

[4] A special allowance has then to be made for those population units that need to be completely enumerated, and the question as to what is the relevant population standard deviation cannot be answered fully at this point, but readers already familiar with the basics of stratification are referred forward to Subsection 3.5.2.

to be appreciable, so the assumption is made that the estimation of the parameter β in Eq. (1a) should be carried out over all size strata combined. That is to say, the value of β is estimated as the ratio of the sum over the strata of the expansion estimates of the survey variable y to the sum over the strata of the expansion estimates of the auxiliary variable x. This is termed the across-stratum ratio estimator of β, and the product of this with the known sum over all sampled size strata of the auxiliary variable X is termed the across-stratum estimator of the total Y of the survey variable y.

This across-stratum ratio estimator, being based on a larger effective sample size than that of any individual stratum, has a smaller randomization bias than the stratum-by-stratum ratio estimator, but because the ratio of y to x is being estimated over all size strata instead of separately for each, there is the strong probability that the randomization variance of the across-stratum ratio estimator will be greater than that of the stratum-by-stratum ratio estimator. Certainly, the estimators of variance yield larger estimates for the former than the latter. So there is a trade-off between unestimated (but undoubtedly real) randomization bias, and estimated randomization variance.

When looked at from the prediction approach, however, the conclusion is quite different. If the prediction models used for the individual size strata have different parameters β_h, say, where h is a stratum indicator, then it is the across-stratum ratio estimator that is now biased (since it is estimating a nonexistent common parameter β) while the stratum-by-stratum ratio estimator (since it relies on small sample sizes for each) may have the larger prediction variance. If however, the prediction models for the different size strata have the same parameter β in common, the stratum-by-stratum ratio estimator is manifestly imprecise, since it is not using all the relevant data for its inferences, and even the across-stratum ratio estimator, while prediction-unbiased, is not using the prediction-optimal weights to estimate the common parameter β.

It therefore appears that looked at from either approach, the choice between these two estimators is suboptimal, and if viewed from both approaches simultaneously, it would usually appear to be inconclusive. The underlying fact is that stratification by size is at best a suboptimal solution to the need for probabilities of inclusion in sample to increase with the size of the population unit. We shall see later (Section 3.9) that a more logical approach would be to avoid using size as an axis of stratification entirely and to use unequal probabilities of inclusion in sample instead. While this does involve certain complications, they are nothing that high-speed computers cannot cope with, whereas the complications brought about by frequent transitions from one size stratum to another within the framework of PRN sampling are distinctly less tractable.

3.6. Sampling with probabilities proportional to size with replacement

As we have just seen, there are now serious arguments for using Unequal Probability Sampling within the context of surveys (chiefly establishment surveys) for which the norm has long been stratification by size and equal inclusion probabilities within strata. However, the genesis of unequal probability sampling, dating from Hansen and Hurwitz (1943), occurred in the very different context of area sampling for household surveys. The objective of Hansen and Hurwitz was to establish a master sample for the conduct of household surveys within the continental United States. It was unreasonable

to contemplate the construction of a framework that included every household in the United States.[5]

Because of this difficulty, Hansen and Hurwitz instead constructed a multistaged framework. They started by dividing the United States into geographical strata, each containing roughly the same number of households. Within each stratum, each household was to have the same probability of inclusion in sample and to make this possible the selection was carried out in stages. The first stage of selection was of Primary Sampling Units (PSUs), which were relatively large geographical and administrative areas. These were sometimes counties, sometimes amalgamations of small counties, and sometimes major portions of large counties.

The important fact was that it was relatively easy to make a complete list of the PSUs within each stratum. However, it was not easy to construct a complete list of PSUs that were of more or less equal size in terms of numbers of households (or dwellings or individuals, whatever was the most accessible measure of size). Some were appreciably larger than others, but the intention remained that in the final sample, each household in the stratum would have the same probability of inclusion as every other household. So Hansen and Hurwitz decided that they would assign each PSU in a given stratum a measure of size; that the sum of those measures of size would be the product of the sample interval (or "spacing interval" or "skip interval") i and the number of PSUs to be selected from that stratum, say n, which number was to be chosen beforehand. Then, a random number r would be chosen between one and the sample interval, and the PSUs selected would be those containing the size measures numbered $r, r + i, r + 2i \ldots r + (n - 1)i$ (see Table 1).

Clearly, the larger the size of a PSU, the larger would be its probability of inclusion in sample. To ensure that the larger probability of selection at the first stage did not translate into a larger probability of inclusion of households at the final stage, Hansen and Hurwitz then required that the product of the probabilities of inclusion at all subsequent stages was to be inversely proportional to the probability of selection at the first stage. So at the final stage of selection (Hansen and Hurwitz contemplated up to three such stages), the population units were individual households and each had the same eventual probability of inclusion in sample as every other household in the stratum.

To ease the estimation of variance, both overall and at each stage, Hansen and Hurwitz allowed it to proceed as though selection had been with replacement at each stage. Since the inclusion probabilities, even at each stage, were comparatively small, this was a reasonable approximation. One of the great simplifications was that the overall variance, the components from all stages combined, could be estimated as though there had been only a single stage of selection. Before the introduction of computers, this was a brilliant simplification, and even today the exact estimation of variance when sampling is without replacement still involves certain complications, considered in Section 3.7.

[5] Conceptually, it might be easier to think of this as a list of every dwelling. In fact, the two would have been identical since the definition of a dwelling was whatever a household was occupying, which might for instance be a share of a private house. A household in turn was defined as a group of people sharing meals on a regular basis.

Table 1
Example of PSU selection with randomized listing

Sample fraction 1/147		Number of sample PSUs 2		Cluster size 32.8	
PSU No.	No. of Dwellings	No. of Clusters	Cumulated Clusters	Selection Number	Within-PSU Sample Fraction
1	1550	47	47		
10	639	20	67		
7	728	22	89		
5	1055	32	121	103	1/32
9	732	22	143		
2	911	28	171		
6	553	17	188		
3	1153	35	223		
4	1457	44	267	250	1/44
8	873	27	294		
Total	9651	294			

Note: The number of clusters in PSU number 10 has been rounded up from 19.48 to 20 in order for the total number of clusters to be divisible by 147. Note also that the selection number 103 lies in the interval between 90 and 121 while the selection number 250 lies in the interval between 224 and 267.

3.7. Sampling with unequal probabilities without replacement

The transition from sampling with replacement to sampling without replacement was reasonably simple for simple random sampling but that was far from the case for sampling with unequal probabilities. The first into the field were Horvitz and Thompson (1952). Their estimator is appropriately named after them as the Horvitz-Thompson Estimator or HTE. It is simply the sum over the sample of the ratios of each unit's survey variable value (y_i for the ith unit) to its probability of inclusion in sample (π_i). The authors showed that this estimator was randomization unbiased. They also produced a formula for its variance and a (usually unbiased) estimator of that variance. These last two formulae were functions of the "second-order inclusion probabilities," that is, the probabilities of inclusion in sample of all possible pairs of population units. If the number of units in the population is denoted by N, then the number of possible pairs is $N(N-1)/2$, so the variance formula involved a summation over $N(N-1)/2$ terms, and even the variance estimation formula required a sum over $n(n-1)/2$ pairs of sample units.

Papers by Sen (1953) and by Yates and Grundy (1953) soon followed. Both of these made use of the fact that when the selection procedure ensured a sample of predetermined size (n units), the variance was both minimized in itself and capable of being estimated much more accurately than when the sample size was not fixed. Both papers arrived at the same formulae for the fixed-sample-size variance and for an estimator of that variance that was randomization unbiased, provided that the joint inclusion probabilities, π_{ij}, for all possible pairs of units were greater than zero. However, this Sen–Yates–Grundy variance estimator still depended on the $n(n-1)/2$ values of the π_{ij} so that the variance could not be estimated randomization-unbiasedly without evaluating this large number of joint inclusion probabilities.

Many without-replacement selection schemes have been devised in attempts to minimize these problems. One of the earliest and simplest was randomized systematic sampling, or "RANSYS," originally described by Goodman and Kish (1950). It involved randomizing the population units and selecting systematically with a skip interval that was constant in terms of the size measures. After 1953, dozens of other methods followed in rapid succession. For descriptions of these early methods, see Brewer and Hanif (1982) and Chaudhury and Vos (1988). However, it seemed to be generally true that if the sample was easy to select, then the inclusion probabilities were difficult to evaluate, and the converse also holds.

Poisson sampling (Hájek, 1964) is one such method that deserves a special mention. Although in its original specification, it did not ensure samples of fixed size, it did have other interesting properties. To select a Poisson sample, each population in turn is subjected to a Bernoulli trial, with the probability of "success" (inclusion in sample) being π_i, and the selection procedure continues until the last population unit has been subjected to its trial. The achieved sample sizes are, however, highly variable, and consequently, Poisson sampling in its original form was not an immediately popular choice. However, several modified versions were later formulated; several of these and also the original version are still in current use.

One of the most important of these modified versions was Conditional Poisson Sampling or CPS, also found in Hájek (1964) and discussed in detail by Chen et al. (1994). For CPS, Poisson samples with a particular expected sample size are repeatedly selected, but only to be immediately rejected once it is certain that the eventual sample will not have exactly that expected sample size. One notable feature of CPS is that it has the maximum entropy attainable for any population of units having a given set of first-order inclusion probabilities π_i.[6] Several fast algorithms for using CPS are now available, in which the second-order inclusion probabilities are also computed exactly. See Tillé (2006).

In the meantime, however, another path of investigation had also been pioneered by Hájek (1964). He was concerned that the estimation of variance for the HTE was unduly complicated by the fact that both the Sen–Yates–Grundy formula for the randomization variance and their estimator of that variance required knowledge of the second-order inclusion probabilities. In this instance, Hájek (and eventually others) approximated the fixed sample size variance of the HTE by an expression that depended only on the first-order inclusion probabilities. However, initially these approximations were taken to be specific to particular selection procedures. For instance, Hájek's 1964 approximation was originally taken to be specific to CPS.

In time, however, it was noted that very different selection procedures could have almost identical values of the π_{ij}. The first two for which this was noticed were RANSYS, for which the π_{ij} had been approximated by Hartley and Rao (1962), and the Rao–Sampford selection procedure (J.N.K. Rao, 1965; Sampford, 1967), for which

[6] Entropy is a measure of unpredictability or randomness. If a population is deliberately arranged in order of size and a sample is selected from it systematically, that sample will have low entropy. If however (as with RANSYS) the units are arranged in random order before selection, the sample will have high entropy, only a few percentage points smaller than that of CPS itself. While low entropy sample designs may have very high or very low randomization variances, high entropy designs with the same set of first-order inclusion probabilities all have more or less the same randomization variance. For a discussion of the role of entropy in survey sampling, see Chen et al. (1994).

they had been approximated by Asok and Sukhatme (1976). These were radically different selection procedures, but the two sets of approximations to the π_{ij} were identical to order n^3/N^3. Although both procedures produced fixed size samples, and the population units had inclusion probabilities that were exactly proportional to their given measures of size, it appeared that the only other thing that the two selection procedures had in common was that they both involved a large measure of randomization. Entropy, defined as $\sum_{k=1}^{M}[P_k - log(P_k)]$, where P_k is the probability of selecting the kth out of the M possible samples, is a measure of the randomness of the selection. It therefore appeared plausible that all high-entropy sampling procedures would have much the same sets of π_{ij}, and hence much the same randomization variance. If so, it followed that approximate variance formulae produced on the basis of any of these methods would be valid approximations for them all, and that useful estimators of these approximate variances would be likely also to be useful estimators of the variances of the HTE for all high-entropy selection procedures.

Whether this is the case or not is currently a matter of some contention, but Preston and Henderson (2007) provide evidence to the effect that the several randomization variance estimators provided along these lines are all reasonably similar in precision and smallness of bias, all at least as efficient as the Sen–Yates–Grundy variance estimator (as measured by their randomization mean squared errors MSEs), and all a great deal less cumbersome to use.

In addition, they can be divided into two families, the members of each family having both a noticeable similarity in structure and a detectable difference in entropy level from the members of the other family. The first family includes those estimators provided by Hájek (1964), by Deville (1993, 1999, 2000; see also Chen et al., 1994) initially for CPS, and by Rosén for Pareto πps (Rosén, 1997a, 1997b). The second family, described in Brewer and Donadio (2003), is based on the π_{ij} values associated with RANSYS and with the Rao–Sampford selection procedure. These two procedures have slightly smaller entropies and slightly higher randomization variance than CPS, but both Preston and Henderson (2007) and Henderson (2006) indicate that the Hájek-Deville family of estimators should be used for CPS, Pareto πps and similar selection procedures—thus probably including Tillé (1996)—while the Brewer-Donadio family estimators would be appropriate for use with RANSYS and Rao-Sampford.

It is also possible to use replication methods, such as the jackknife and the bootstrap, to estimate the HTE's randomization variance. The same Preston and Henderson paper provides evidence that a particular version of the bootstrap can provide adequate, though somewhat less accurate, estimates of that variance than can be obtained using the two families just described.

Finally, it is of interest that the "anticipated variance" of the HTE (that is to say the randomization expectation of its prediction variance, or equivalently the prediction expectation of its randomization variance; see Isaki and Fuller, 1982) is a simple function of the π_i and independent of the π_{ij}. Hence, for any population that obeys the model of Eqns. (1), both the randomization variance and the anticipated variance of the HTE can be estimated without any reference to the π_{ij}.

3.8. The generalized regression estimator

Up to this point, it has been assumed that only a single auxiliary variable has been available for improving the estimation of the mean or total of a survey variable. It

has also been assumed that the appropriate way to use that auxiliary variable was by using Eq. (1a), which implies a ratio relationship between those two variables. More generally, the survey variable could depend on a constant term as well, or on more than a single auxiliary variable, or both. However, that relationship is seldom likely to be well represented by a model that implies the relevance of ordinary least squares (OLS).

One case where OLS might be appropriate is where the survey variable is Expenditure and the auxiliary variable is Income. The relationship between Income and Expenditure (the Consumption Function) is well known to involve an approximately linear dependence with a large positive intercept on the Expenditure axis. But OLS assumes homoskedasticity (the variance of Expenditure remaining constant as Income increases) while it is more than likely that the variance of Expenditure increases with Income, and in fact the data from the majority of sample surveys do indicate the existence of a measure of heteroskedaticity. This in itself is enough to make the use of OLS questionable. Eq. (1c) allows for the variance of the survey variable to increase linearly with the auxiliary variable, and in fact it is common for this variance to increase somewhat faster than this, and occasionally as fast as the square of the auxiliary variable.

A commonly used estimator of total in these more general circumstances is the generalized regression estimator or GREG (Cassel et al., 1976), which may be written as follows:

$$\hat{Y}_{\mathrm{GREG}} = \hat{Y}_{\mathrm{HTE}} + \sum_{k=1}^{p} (X_k - \hat{X}_{\mathrm{HTE}k})\hat{\beta}_k, \tag{8}$$

or alternatively as

$$\hat{Y}_{\mathrm{GREG}} = \sum_{k=1}^{p} X_k \hat{\beta}_k + \left(\hat{Y}_{\mathrm{HTE}} - \sum_{k=1}^{p} X_{\mathrm{HTE}k} \hat{\beta}_k \right). \tag{9}$$

In these two equations, $\hat{Y}_{\mathrm{HTE}}$ is the HTE of the survey variable, $\hat{X}_{\mathrm{HTE}k}$ is the HTE of the kth auxiliary variable and $\hat{\beta}_k$ is an estimator of the regression coefficient of the survey variable on the kth auxiliary variable, where the regression is on p auxiliary variables simultaneously. One of those auxiliary variables may be a constant term, in which case there is an intercept estimated in the equation. (In that original paper, $\hat{\beta}_k$ was a generalized least squares estimator, but this was not a necessary choice. For instance, Brewer (1999b) defined $\hat{\beta}_k$ in such a way as to ensure that the GREG was simultaneously interpretable in the randomization and prediction approaches to sampling inference, and also showed that this could be achieved with only trivial increments to its randomization and prediction variances).

In the second of these two equations, the first term on the right-hand side is a prediction estimator of the survey variable total, but one that ignores the extent to which the HTE of the survey variable total differs from the sum of the p products of the individual auxiliary variable HTEs with their corresponding regression estimates. Särndal et al. (1992) noted that the first term (the prediction estimator) had a randomization variance that was of a lower order of magnitude than the corresponding variance of the second term and therefore suggested that the randomization variance of the GREG estimator be estimated by estimating only that of the second term. It is true that as the sample size increases, the randomization variance of the prediction estimator becomes small with

respect to that of the second term, but when the sample size is small, this can lead to a substantial underestimate of the GREG's randomization variance.

This is not an easy problem to solve wholly within the randomization approach, and in Chapter 8 of Brewer (2002, p. 136), there is a recommendation to estimate the anticipated variance as a substitute. (The anticipated variance is the randomization expectation of the prediction variance.). This is obviously not a fully satisfactory solution, except in the special case considered by Brewer, where the GREG had been devised to be simultaneously a randomization estimator and a prediction estimator, so more work on it seems to be called for. Another alternative would be to estimate the GREG's randomization variance using a replication method such as the jackknife or the bootstrap, but again this alternative appears to need further study. For more information regarding the GREG, see Särndal et al. (1992).

3.9. Permanent random number (PRN) sampling

One of the important but less obvious objectives of survey sampling is to be able to control intelligently the manner in which the sample for a repeating survey is allowed to change over time. It is appropriate for a large sample unit that is contributing substantially to the estimate of total to remain in sample for fairly long periods, but it is not so appropriate for small population units to do the same, so it is sensible to rotate the sample around the population in such a way that the larger the unit is, the longer it remains in sample. One of the ways of doing this is to assign each unit a PRN, say between zero and unity, and define the sample as consisting of those population units that occupy certain regions of that PRN space. Units in a large-size stratum might initially be in sample if they had PRNs between zero and 0.2 for the initial survey, between 0.02 and 0.22 for the second, 0.04 and 0.24 for the third, and so on. In this way, each unit would remain in sample for up to 10 occasions but then be "rested" for the next 40. Those in a small-size stratum would remain occupy a smaller region of the PRN space, say initially between zero and 0.04, but the sample PRN space would be rotated just as fast so that units would remain in sample for no more than two occasions before being "rested."

From the data supplier's point of view, however, it is particularly inappropriate to be removed from the sample and then included again shortly afterwards. This can easily happen, however, if a population unit changes its size stratum, particularly if the change is upward. Consequently, it is inconvenient to use PRN sampling and size stratification together. Moreover, as has already been indicated in Section 3.5, stratification by size is a suboptimal way of satisfying the requirement that the larger the unit, the greater should be its probability of inclusion in sample.

Hence, when attempting to control and rotate samples using the PRN technique, it becomes highly desirable, if not indeed necessary, to find a better solution than stratification by size. Brewer (2002) (Chapter 13, pp. 260–265), provides a suggestion as to how this could be done. It involves the use of a selection procedure known as Pareto πps sampling, which is due to Rosén (1997a, 1997b). This is a particular form of what is known as *order sampling*, and is very similar in its π_{ij} values to CPS sampling, so it is a high-entropy sample selection procedure. It is, however, somewhat complicated to describe and therefore inappropriate to pursue further in this introductory chapter. Those who wish to pursue the possibility of using PRN sampling without stratification by size are referred to those two papers by Rosén and to Chapter 13 of Brewer (2002).

4. Conclusion

From the very early days of survey sampling, there have been sharp disagreements as to the relative importance of the randomization and prediction approaches to survey sampling inference. These disagreements are less severe now than they were in the 1970s and 1980s but to some extent they have persisted into the 21st century. What is incontrovertible, however, is that prediction inference is parametric and randomization nonparametric. Hence the prediction approach is appropriate to the extent that the prediction models are useful, whereas the randomization approach provides a robust alternative where they are not useful. It would therefore seem that ideally both should be used together, but there are many who sincerely believe the one or the other to be irrelevant. The dialogue therefore continues.

Both the randomization and the prediction approaches offer a wide range of manners in which the sample can or should be selected, and an equally wide range of manners in which the survey values (usually, but not exclusively consisting of population totals, population means, and ratios between them) can be estimated. The choices among them depend to a large extent on the natures of the populations (in particular, whether they consist of individuals and households, of establishments and enterprizes, or of some other units entirely) but also on the experience and the views of the survey investigators. However, there are some questions that frequently need to be asked, and these are the ones that have been focussed on in this chapter. They include, "What are the units that constitute the population?" "Into what groups or strata do they naturally fall?" "What characteristics of the population need to be estimated?" "How large a sample is appropriate?" (or alternatively, "How precise are the estimates required to be?") "How should the sample units be selected?" and "How should the population characteristics be estimated?"

In addition, there are many questions that need to be answered that fall outside the scope of the discipline of survey sampling. A few of them would be as follows: "What information are we seeking, and for what reasons?" "What authority, if any, do we have to ask for this information?" "In what format should it be collected?" "What organizational structure is required?" "What training needs to be given and to whom?" and not least, "How will it all be paid for?"

So those questions that specifically relate to survey sampling always need to be considered in this wider framework. The aim of this Chapter will have been achieved if the person who has read it has emerged with some feeling for the way in which the discipline of survey sampling can be used to fit within this wider framework.