for survey data analysts to augment the design-based approach. In some cases, both approaches produce the same results; but different results occur in other cases. The model-based approach may not be useful in descriptive data analysis but can be useful in inferential analysis. We will introduce the model-based perspective where appropriate and provide references for further treatment of the topics. Proper conduct of model-based analysis would require knowledge of general statistical models and perhaps some consultation from survey statisticians. Sections of the book relevant to this alternative approach and related topics are marked with asterisks.

Since the publication of the first edition of this book, the software situation for the analysis of complex survey data has improved considerably. User-friendly programs are now readily available, and many commonly used statistical methods are now incorporated in the packages, including logistic regression and survival analysis. These programs will be introduced with illustrations in this edition. These programs are perhaps more open to misuse than other standard software. The topics and issues discussed in this book will provide some guidelines for avoiding pitfalls in survey data analysis.

In our presentation, we assume some familiarity with such sampling designs as simple random sampling, systematic sampling, stratified random sampling, and simple two-stage cluster sampling. A good presentation of these designs may be found in Kalton (1983) and Lohr (1999). We also assume general understanding of standard statistical methods and one of the standard statistical program packages, such as SAS or Stata.

2. SAMPLE DESIGN AND SURVEY DATA

Our consideration of survey data focuses on sample designs that satisfy two basic requirements. First, we are concerned only with probability sampling in which each element of a population has a known (nonzero) probability of being included in the sample. This is the basis for applying statistical theory in the derivation of the properties of the survey estimators for a given design. Second, if a sample is to be drawn from a population, it is necessary to be able to construct a sampling frame that lists suitable sampling units that encompass all elements of the population. If it is not feasible or is impractical to list all population elements, some clusters of elements can be used as sampling units. For example, it is impractical to construct a list of all households in the United States, but we can select the sample in several stages. In the first stage, counties are randomly sampled; in the second stage, census tracts within the selected counties are sampled; in the third stage, street blocks are sampled within the selected tracts. Then, in the final stage of selection, a list of households is needed only for the selected blocks. This multistage design satisfies the requirement that all population elements have a known nonzero probability of being selected.

Types of Sampling

The simplest sample design is *simple random sampling*, which requires that each element has an equal probability of being included in the sample and that the list of all population elements is available. Selection of a sample element can be carried out with or without replacement. Simple random sampling with replacement (SRSWR) is of special interest because it simplifies statistical inference by eliminating any relation (covariance) between the selected elements through the replacement process. In this scheme, however, an element can appear more than once in the sample. In practice, simple random sampling is carried out without replacement (SRSWOR), because there is no need to collect the information more than once from an element. Additionally, SRSWOR gives a smaller sampling variance than SRSWR. However, these two sampling methods are practically the same in a large survey in which a small fraction of population elements are sampled. We will use the term SRS for SRSWOR throughout this book unless otherwise specified.

The SRS design is modified further to accommodate other theoretical and practical considerations. The common practical designs include systematic sampling, stratified random sampling, multistage cluster sampling, PPS sampling (probability proportional to size), and other controlled selection procedures. These more practical designs deviate from SRS in two important ways. First, the inclusion probabilities for the elements (also the joint inclusion probabilities for sets for the elements) may be unequal. Second, the sampling unit can be different from the population element of interest. These departures complicate the usual methods of estimation and variance calculation and, if proper methods of analysis are not used, can lead to a bias in estimation and statistical tests. We will consider these departures in detail, using several specific sampling designs, and examine their implications for survey analysis.

Systematic sampling is commonly used as an alternative to SRS because of its simplicity. It selects every k-th element after a random start (between 1 and k). Its procedural tasks are simple, and the process can easily be checked, whereas it is difficult to verify SRS by examining the results. It is often used in the final stage of multistage sampling when the fieldworker is instructed to select a predetermined proportion of units from the listing of dwellings in a street block. The systematic sampling procedure assigns each element in a population the same probability of being selected. This ensures that the sample mean will be an unbiased estimate of the population mean when the number of elements in the population (N) is equal to k times the number of elements in the sample (n). If N is not exactly nk, then the equal probability is not guaranteed, although this problem can be ignored when N is large. In that case, we can use the circular systematic sampling scheme. In this scheme, the random starting point is selected between 1 and N (any element can be the starting point), and every k-th element is selected assuming that the frame is circular (the end of the list is connected to the beginning of the list). Systematic sampling can give an unrealistic estimate, however, when the elements in the frame are listed in a cyclical manner with respect to survey variables and the selection interval coincides with the listing cycle. For example, if one selects every 40th patient coming to a clinic and the average daily patient load is about 40, then the resulting systematic sample would contain only those who came to the clinic at a particular time of the day. Such a sample may not be representative of the clinic patients.

Moreover, even when the listing is randomly ordered, unlike SRS, different sets of elements may have unequal inclusion probabilities. For example, the probability of including both the i-th and the $(i + k)$-th element is $1/k$ in a systematic sample, whereas the probability of including both the i-th and the $(i + k + 1)$-th is zero. This complicates the variance calculation. Another way of viewing systematic sampling is that it is equivalent to selecting one cluster from k systematically formed clusters of n elements each. The sampling variance (between clusters) cannot be estimated from the one selected cluster. Thus, variance estimation from a systematic sample requires special strategies.

A modification to overcome these problems with systematic sampling is the so-called *repeated systematic sampling* (Levy & Lemeshow, 1999, pp. 101–110). Instead of taking a systematic sample in one pass through the list, several smaller systematic samples are selected, going down the list several times with a new starting point in each pass. This procedure not only guards against possible periodicity in the frame but also allows variance estimation directly from the data. The variance of an estimate from all subsamples can be estimated from the variability of the separate estimates from each subsample. This idea of replicated sampling offers a strategy for estimating variance for complex surveys, which will be discussed further in Chapter 4.

Stratified random sampling classifies the population elements into strata and samples separately from each stratum. It is used for several reasons: (a) The sampling variance can be reduced if strata are internally homogeneous, (b) separate estimates can be obtained for strata, (c) administration of fieldwork can be organized using strata, and (d) different sampling needs can be accommodated in separate strata. Allocation of the sample across the strata is proportionate when the sampling fraction is uniform across the strata or disproportionate when, for instance, a higher sampling fraction is applied to a smaller stratum to select a sufficient number of subjects for comparative studies. In general, the estimation process for a stratified random sample is more complicated than in SRS. It is generally described as a two-step process. The first step is the calculation of the statistics—for example, the mean and its variance—separately within each stratum. These estimates are then combined based on weights reflecting the proportion of the population in each stratum. As will be discussed later, it also can be described as a one-step process using weighted statistics. The estimation simplifies in the case of proportionate stratified sampling, but the strata must be taken into account in the variance estimation.

The formulation of the strata requires that information on the stratification variable(s) be available in the sampling frame. When such information is not available, stratification cannot be incorporated in the design. But stratification can be done after data are collected to improve the precision of the estimates. The so-called poststratification is used to make the sample more representative of the population by adjusting the demographic compositions of the sample to the known population compositions. Typically, such demographic variables as age, sex, race, and education are used in poststratification in order to take advantage of the population census data. This adjustment requires the use of weights and different strategies for variance estimation because the stratum sample size is a random variable in the poststratified design (determined after the data are collected).

Cluster sampling is often a practical approach to surveys because it samples by groups (clusters) of elements rather than by individual elements. It simplifies the task of constructing sampling frames, and it reduces the survey costs. Often, a hierarchy of geographical clusters is used, as described earlier. In multistage cluster sampling, the sampling units are groups of elements except for the last stage of sampling. When the numbers of elements in the clusters are equal, the estimation process is equivalent to SRS. However, simple random sampling of unequal-sized clusters leads to the elements in the smaller clusters being more likely to be in the sample than those in the larger clusters. Additionally, the clusters are often stratified to accomplish certain survey objectives and field procedures, for instance, the oversampling of predominantly minority population clusters. The use of disproportionate stratification and unequal-sized clusters complicates the estimation process.

One method to draw a self-weighting sample of elements in one-stage cluster sampling of unequal size clusters is to sample clusters with probability proportional to the size of clusters (PPS *sampling*). However, this requires that the true size of clusters be known. Because the true sizes usually are unknown at the time of the survey, the selection probability is instead made proportional to the estimated size (PPES *sampling*). For example, the number of beds can be used as a measure of size in a survey of hospital discharges with hospitals as the clusters. One important consequence of PPES sampling is that the expected sample size will vary from one primary sampling unit (PSU) to another. In other words, the sample size is not fixed but varies from sample to sample. Therefore, the sample size, the denominator in the calculation of a sample mean, is a random variable, and, hence, the sample mean becomes a ratio of two random variables. This type of variable, a ratio variable, requires special strategies for variance estimation.

The Nature of Survey Data

If we are to infer from sample to population, the sample selection process is an integral part of the inference process, and the survey data must contain information on important dimensions of the selection process. Considering the departures from SRS in most social surveys, we need to view the survey data not only as records of measurements, but also as having different representation and structural arrangements.

Sample weights are used to reflect the differing probabilities of selection of the sample elements. The development of sample weights requires keeping track of selection probabilities separately in each stratum and at each stage of sampling. In addition, it can involve correcting for differential response rates within classes of the sample and adjusting the sample distribution by demographic variables to known population distributions (poststratification adjustment). Moreover, different sample weights may be needed for different units of analysis. For instance, in a community survey it may be necessary to develop person weights for an analysis of individual data and household weights for an analysis of household data.

We may feel secure in the exclusion of the weights when one of the following self-weighting designs is used. True PPS sampling in a one-stage cluster sampling will produce a self-weighting sample of elements, as in the SRS design. The self-weighting can also be accomplished in a two-stage design when true PPS sampling is used in the first stage and a fixed number of elements is selected within each selected PSU. The same result will follow if simple random sampling is used in the first stage and a fixed proportion of the elements is selected in the second stage (see Kalton, 1983, chaps. 5 and 6). In practice, however, the self-weighting feature is destroyed by nonresponse and possible errors in the sampling frame(s). This unintended self-selection process can introduce bias, but it is seldom possible to assess the bias from an examination of the sample data. Two methods employed in an attempt to reduce the bias are poststratification and nonresponse adjustments. Poststratification involves assigning weights to bring the sample proportion in demographic subgroups into agreement with the population proportion in the subgroups. Nonresponse adjustment inflates the weights for those who participate in the survey to account for the nonrespondents with similar characteristics. Because of the nonresponse and poststratification adjustments by weighting, the use of weights is almost unavoidable even when a self-weighting design is used.

The sample design affects the estimation of standard errors and, hence, must also be incorporated into the analysis. A close examination of the familiar formulas for standard errors found in statistics textbooks and incorporated into most computer program packages shows that they are based on the SRSWR design. These formulas are relatively simple because the covariance between elements is zero, as a result of the assumed independent selection of elements. It is not immediately evident how the formulas should be modified to adjust for other complex sampling designs.

To better understand the need for adjustment to the variance formulas, let us examine the variance formula for several sample designs. We first consider variance for a sample mean from the SRSWOR design. The familiar variance formula for a sample mean, $\bar{y}$ (selecting a sample of n elements from a population of N elements by SRSWR where the population mean is $\bar{Y}$), in elementary statistics textbooks is σ^2/n, where $\sigma^2 = \sum(Y_i - \bar{Y})^2/N$. This formula needs to be modified for the SRSWOR design because the selection of an element is no longer independent of the selection of another element. Because of the condition of not allowing duplicate selection, there is a negative covariance $[-\sigma^2/(N-1)]$ between i-th and j-th sample elements. Incorporating $n(n-1)$ times the covariance, the variance of the sample mean for SRSWOR is $\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$, which is smaller than that from SRSWR by the factor of $(N-n)/(N-1)$. Substituting the unbiased estimator of σ^2 of $[(N-1)s^2/N]$, the estimator for the variance of the sample mean from SRSWOR is

$$\hat{V}(\bar{y}) = \frac{s^2}{n}(1-f),$$

where $s^2 = \sum(x_i - \bar{x})^2/(n-1)$ and $f = n/N$. Both $(N-n)/(N-1)$ and $(1-f)$ are called the finite population correction (FPC) factor. In a large population, the covariance will be very small because the sampling fraction is small. Therefore, SRSWR and SRSWOR designs will produce practically the same variance, and these two procedures can be considered equivalent for all practical purposes.

Stratified sampling is often presented as a more efficient design because it gives, if used appropriately, a smaller variance than that given by a comparable SRS. Because the covariances between strata are zero, the variance of the sample estimate is derived from the within-stratum variances, which are combined based on the stratum sample sizes and the stratum weights. The value of a stratified sample variance depends on the distribution of the strata sample sizes. An optimal (or Neyman) allocation produces a sampling variance less than or equal to that based on SRS except in extremely rare situations. For other disproportionate allocations, the sampling variance may turn out to be larger than that based on SRS when the finite population correction factor (FPC) within strata cannot be ignored. Therefore, it cannot be assumed that stratification will always reduce sampling variance compared to SRS.

The *cluster sampling* design usually leads to a larger sampling variance than that from SRS. This is because the elements within naturally formed clusters are often similar, which then yields a positive covariance between elements within the cluster. The homogeneity within clusters is measured by the intraclass correlation coefficient (ICC)—the correlation between all possible pairs of elements within clusters. If clusters were randomly formed (i.e., if each cluster were a random sample of elements), the ICC would be zero. In many natural clusters, the ICC is positive and, hence, the sampling variance will be larger than that for the SRS design.

It is difficult to generalize regarding the relative size of the sampling variance in a complex design because the combined effects of stratification and clustering, as well as that of the sample weights, must be assessed. Therefore, all observations in survey data must be viewed as products of a specific sample design that contains sample weights and structural arrangements. In addition to the sample weights, strata and cluster identification (at least PSUs) should be included in sample survey data. Reasons for these requirements will become clearer later.

One complication in the variance calculation for a complex survey stems from the use of weights. Because the sum of weights in the denominator of any weighted estimator is not fixed but varies from sample to sample, the estimator becomes a ratio of two random variables. In general, a ratio estimator is biased, but the bias is negligible if the variation in the weights is relatively small or the sample size is large (Cochran, 1977, chap. 6). Thus, the problem of bias in the ratio estimator is not an issue in large social surveys. Because of this bias, however, it is appropriate to use the mean square error—the sum of the variance plus the square of the bias—rather than the variance. However, because the bias often is negligible, we will use the term ‘‘variance’’ even if we are referring to mean square error in this book.

A Different View of Survey Data*

So far, the nature of survey data is described from the design-based perspective—that is, sample data are observations sampled from a finite population using a particular sample selection design. The sampling design specifies the probability of selection of each potential sample, and a proper estimator is chosen to reflect the design. As mentioned in the introduction, the model-based perspective offers an alternative view of sample survey data. Observations in the finite population are viewed as realizations of a random variable generated from some model (a random variable that followed some probability distribution). The assumed probability model supplies the link between units in the sample and units not in the sample. In the model-based approach, the sample data are used to predict the unobserved values, and thus inferences may be thought of as prediction problems (Royall, 1970, 1973).

These two points of view may not make a difference in SRS, where we can reasonably assume that sample observations were independent and identically distributed from a normal distribution with mean μ and variance σ. From the model point of view, the population total is the sum of observations in the sample and the sum of observations that are not in the sample; that is, $Y=\sum_{i\in S} y_i+\sum_{i\notin S} y_i$. Based on the assumption of common mean, the estimate of population total can be made as $\hat{Y}=n\bar{y}+(N-n)\bar{y}=N\bar{y}$, where $\bar{y}$ is the best unbiased predictor of the unobserved observations under the model. It turns out to be the same as the expansion estimator in the design-based approach, namely, $\hat{Y}=(N/n)\sum_{i=1}^{n} y_i=N\bar{y}$, where (N/n) is the sample weight (inverse of selection probability in SRS). Both approaches lead to the same variance estimate (Lohr, 1999, sec. 2.8).

If a different model were adopted, however, the variance estimates might differ. For example, in the case of ratio[1] and regression estimation under SRS, the assumed model is $Y_i=\beta x_i+\varepsilon_i$, where Y_i is for a random variable and x_i is an auxiliary variable for which the population total is known. Under this model, the linear estimate of the population total will be $\hat{Y}=\sum_{i\in S} y_i+\sum_{i\notin S} y_i=n\bar{y}+\hat{\beta}\sum_{i\notin S} x_i$. The first part is from the sample, and the second part is the prediction for the unobserved units based on the assumed model. If we take $\hat{\beta}$ as the sample ratio of $\bar{y}/\bar{x}$, then we have $\hat{Y}=n\bar{y}+\frac{\bar{y}}{\bar{x}}\sum_{i\notin S} x_i=\frac{\bar{y}}{\bar{x}}(n\bar{x}+\sum_{i\notin S} x_i)=\frac{\bar{y}}{\bar{x}}X$, where X is the population total of x_i. This is simply the ratio estimate of Y. If we take $\hat{\beta}$ as the estimated regression coefficient, then we have regression estimation. Although the ratio estimate is known to be slightly biased from the design-based viewpoint, it is unbiased from the model-based reasoning if the model is correct.

But the estimate of variance by the model-based approach is slightly different from the estimate by the design-based approach. The design-based estimate of variance of the estimated population total is $\hat{V}_D(\hat{Y}) = (1 - \frac{n}{N})\left(\frac{N^2}{n}\right)\frac{\sum[y_i - (\bar{y}/\bar{x})x_i]^2}{n-1}$. The model-based estimator is $\hat{V}_M(\hat{Y}) = (1 - \frac{x}{X})\left(\frac{X^2}{x}\right)\frac{\sum[\{y_i - (\bar{y}/\bar{x})\}/\sqrt{x_i}]^2}{n-1}$, where x is the sample total and X is the population total of the auxiliary variable (see Lohr, 1999, sec. 3.4).

The ratio estimate model is valid when (a) the relation between y_i and x_i is a straight line through the origin and (b) the variance of y_i about this line is proportional to x_i. It is known that the ratio estimate is inferior to the expansion estimate (without the auxiliary variable) when the correlation between y_i and x_i is less than one-half the ratio of coefficient of variation of x_i over the coefficient of variation of y_i (Cochran, 1977, chap. 6). Therefore, the use of ratio estimation in survey analysis would require checking the model assumptions. In practice, when the data set includes a large number of variables, ratio estimation would be cumbersome to select different auxiliary variables for different estimates.

To apply the model-based approach to a real problem, we must first be able to produce an adequate model. If the model is wrong, the model-based estimators will be biased. When using model-based inference in sampling, one needs to check the assumptions of the model by examining the data carefully. Checking the assumptions may be difficult in many circumstances. The adequacy of a model is to some extent a matter of judgment, and a model adequate for one analysis may not be adequate for another analysis or another survey.

3. COMPLEXITY OF ANALYZING SURVEY DATA

Two essential aspects of survey data analysis are adjusting for the differential representation of sample observations and assessing the loss or gain in precision resulting from the complexity of the sample selection design. This chapter introduces the concept of weight and discusses the effect of sample selection design on variance estimation. To illustrate the versatility of weighting in survey analysis, we present two examples of developing and adjusting sample weights.

Adjusting for Differential Representation: The Weight

Two types of sample weights are commonly encountered in the analysis of survey data: (a) the expansion weight, which is the reciprocal of the selection probability, and (b) the relative weight, which is obtained by scaling down the expansion weight to reflect the sample size. This section reviews these two types of weights in detail for several sample designs.