2

Getting to Know the Complex Sample Design

2.1 Introduction

The first step in the applied analysis of survey data involves defining the research questions that will be addressed using a given set of survey data. The next step is to study and understand the sampling design that generated the sample of elements (e.g., persons, businesses) from the target population of interest, given that the actual survey data with which the reader will be working were collected from the elements in this sample. This chapter aims to help the readers understand the complex sample designs that they are likely to encounter in practice and identify the features of the designs that have important implications for correct analyses of the survey data.

Although a thorough knowledge of sampling theory and methods can benefit the survey data analyst, it is not a requirement. With a basic understanding of complex sample design features, including stratification, clustering, and weighting, readers will be able to specify the key design parameters required by today's survey data analysis software systems. Readers who are interested in a more in-depth treatment of sampling theory and methods are encouraged to review work by Hansen, Hurwitz, and Madow (1953), Kish (1965), or Cochran (1977). More recent texts that blend basic theory and applications include Levy and Lemeshow (2007) and Lohr (1999). A short monograph by Kalton (1983) provides an excellent summary of survey sample designs.

The sections in this chapter outline the key elements of complex sample designs that analysts need to understand to proceed knowledgeably and confidently to the next step in the analysis process.

2.1.1 Technical Documentation and Supplemental Literature Review

The path to understanding the complex sample design and its importance to the reader's approach to the analysis of the survey data should begin with a review of the technical documentation for the survey and the sample design. A review of the literature, including both supplemental methodological reports and papers that incorporate actual analysis of the survey data, will be quite beneficial for the reader's understanding of the data to be analyzed.

Technical documentation for the sample design, weighting, and analysis procedures should be part of the "metadata" that are distributed with a survey data set. In the real world of survey data, the quality of this technical documentation can be highly variable, but, at a minimum, the reader should expect to find a summary description of the sample, a discussion of weighting and estimation procedures, and, ideally, specific guidance on how to perform complex sample analysis of the survey data. Readers who plan to analyze a public use survey data set but find that documentation of the design is lacking or inadequate should contact the help desk for the data distribution Web site or inquire with the study team directly to obtain or clarify the basic information needed to correctly specify the sample design when analyzing the survey data.

Before diving into the statistical analysis of a survey data set, time is well spent in reviewing supplemental methodological reports or published scientific papers that used the same data. This review can identify important new information or even guide readers' choice of an analytic approach to the statistical objectives of their research.

2.2 Classification of Sample Designs

As illustrated in Figure 2.1, Hansen, Madow, and Tepping (1983) define a sampling design to include two components: the sampling plan and a method for drawing inferences from the data generated under the sampling plan. The vast majority of survey data sets will be based on sample designs that fall in Cell A of Figure 2.1—that is, designs that include a sampling plan based on probability sample methods and assume that statistical inferences concerning population characteristics and relationships will be derived using the "design-based" theory initially proposed by Neyman

	Method of Inference	
Sampling Plan	**Design-based**	**Model-based**
Probability Sample	A	B
Model-dependent Sample	C	D

FIGURE 2.1
Classification of sample designs for survey data.

(1934). Consequently, in this book we will focus almost exclusively on sample designs of this type.

2.2.1 Sampling Plans

Probability sampling plans assign each member of the population a known nonzero probability of inclusion in the sample. A probability sample plan may include features such as stratification and clustering of the population prior to selection—it does not require that the selection probability of one population element be independent of that for another. Likewise, the sample inclusion probabilities for different population elements need not be equal. In a probability sampling of students in a coeducational secondary school, it is perfectly appropriate to sample women at a higher rate than men with the requirement that weights will be needed to derive unbiased estimates for the combined population. Randomization of sample choice is always introduced in probability sampling plans.

Model-dependent sampling plans (Valliant et al., 2000) assume that the variables of interest in the research follow a known probability distribution and optimize the choice of sample elements to maximize the precision of estimation for statistics of interest. For example, a researcher interested in estimating total annual school expenditures, y, for teacher salaries may assume the following relationship of the expenditures to known school enrollments, x: $y_i = \beta x_i + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2 x_i)$. Model-dependent sampling plans may employ stratification and clustering, but strict adherence to randomized selection is not a requirement. Model-dependent sampling plans have received considerable attention in the literature on sampling theory and methods; however, they are not common in survey practice due to a number of factors: the multipurpose nature of most surveys; uncertainty over the model; and lack of high-quality ancillary data (e.g., the x variable in the previous example).

Though not included in Figure 2.1, **quota sampling**, **convenience sampling**, **snowball sampling**, and "**peer nomination**" are nonprobability methods for selecting sample members (Kish, 1965). Practitioners who employ these methods base their choices on assumptions concerning the "representativeness" of the selection process and often analyze the survey data using inferential procedures that are appropriate for probability sample plans. However, these sampling plans do not adhere to the fundamental principles of either probability sample plans or a rigorous probability model-based approach. Is it impossible to draw correct inferences from nonprobability sample data? No, because by chance an arbitrary sample could produce reasonable results; however, the survey analyst is left with no theoretical basis to measure the variability and bias associated with his or her inferences. Survey data analysts who plan to work with data collected under nonprobability sampling plans should carefully evaluate and report the potential selection biases and other survey errors that could affect their final results. Since there is no true statistical basis of support for the analysis of data collected under these non-probability designs, they will not be addressed further in this book.

2.2.2 Inference from Survey Data

In the Hansen et al. (1983) framework for sample designs, a sampling plan is paired with an approach for deriving inferences from the data that are collected under the chosen plan. **Statistical inferences** from sample survey data may be "design-based" or "model-based." The natural design pairings of sampling plans and methods of inference are the diagonal cells of Figure 2.1 (A and D); however, the hybrid approach of combining probability sample plans with model-based approaches to inference is not uncommon in survey research. Both approaches to statistical inference use probability models to establish the correct form of confidence intervals or hypothesis tests for the intended statistical inference. Under the "design-based" or "randomization-based" approach formalized by Neyman (1934), the inferences are derived based on the distribution of all possible samples that could have been chosen under the sample design. This approach is sometimes labeled "nonparametric" or "distribution free" because it relies only on the known probability that a given sample was chosen and not on the specific probability distribution for the underlying variables of interest, for example, $y \sim N(\beta_0 + \beta_1 x, \sigma^2_{y \cdot x})$ (see Chapter 3).

As the label implies, model-based inferences are based on a probability distribution for the random variable of interest—not the distribution of probability for the sample selection. Estimators, standard errors, confidence intervals, and hypothesis tests for the parameters of the distribution (e.g., means, regression coefficients, and variances) are typically derived using the method of maximum likelihood or possibly Bayesian models (see Little, 2003).

2.3 Target Populations and Survey Populations

The next step in becoming acquainted with the sample design and its implications for the reader's analysis is to verify the survey designers' intended study population—who, when, and where. Probability sample surveys are designed to describe a **target population**. The target populations for survey designs are **finite populations** that may range from as few as a 100 **population elements** for a survey of special groups to millions and even billions for national population surveys. Regardless of the actual size, each discrete population element ($i = 1, \ldots, N$) could, in theory, be counted in a census or sampled for survey observation.

In contrast to the target population, the **survey population** is defined as the population that is truly eligible for sampling under the survey design

(Groves et al., 2004). In survey sampling practice, there are geographical, political, social, and temporal factors that restrict our ability to identify and access individual elements in the complete target population and the de facto **coverage** of the survey is limited to the survey population. Examples of geographic restrictions on the survey population could include persons living in remote, sparsely populated areas such as islands, deserts, or wilderness areas. Rebellions, civil strife, and governmental restrictions on travel can limit access to populations living in the affected areas. Homelessness, institutionalization, military service, nomadic occupations, physical and mental conditions, and language barriers are social and personal factors that can affect the coverage of households and individuals in the target population. The timing of the survey can also affect the coverage of the target population. The target population definition for a survey assumes that the data are collected as a "snapshot" in time when in fact the data collection may span many months.

The target population for the National Comorbidity Survey Replication (NCS-R) is defined to be age 18 and older adults living in the households in the United States as of July 1, 2002. Here is the exact definition of the survey population for the NCS-R:

> The survey population for the NCS-R included all U.S. adults aged 18+ years residing in households located in the coterminous 48 states. Institutionalized persons including individuals in prisons, jails, nursing homes, and long-term medical or dependent care facilities were excluded from the survey population. Military personnel living in civilian housing were eligible for the study, but due to security restrictions residents of housing located on a military base or military reservation were excluded. Adults who were not able to conduct the NCS-R interview in English were not eligible for the survey. (Heeringa et al., 2004)

Note that among the list of exclusions in this definition, the NCS-R survey population excludes residents of Alaska and Hawaii, institutionalized persons, and non-English speakers. Furthermore, the survey observations were collected over a window of time that spanned several years (February 2001 to April 2003). For populations that remain stable and relatively unchanged during the survey period, the time lapse required to collect the data may not lead to bias for target population estimates. However, if the population is mobile or experiences seasonal effects in terms of the survey variables of interest, considerable change can occur during the window of time that the survey population is being observed.

As the survey data analyst, the reader will also be able to restrict his or her analysis to **subpopulations** of the survey population represented in the survey data set, but the analysis can use only the available data and cannot directly reconstruct coverage of the unrepresented segments of the target population. Therefore, it is important to carefully review the definition of the survey population and assess the implications of any exclusions for the inferences to be drawn from the analysis.

2.4 Simple Random Sampling: A Simple Model for Design-Based Inference

Simple random sampling with replacement (SRSWR) is the most basic of sampling plans, followed closely in theoretical simplicity by simple random sampling without replacement (SRSWOR, or the short form, SRS, in this text). Survey data analysts are unlikely to encounter true SRS designs in survey practice. Occasionally, SRS may be used to select samples of small localized populations or samples of records from databases or file systems, but this is rare. Even in cases where SRS is practicable, survey statisticians will aim to introduce simple stratification to improve the efficiency of sample estimates (see the next sections). Furthermore, if an SRS is selected but weighting is required to compensate for nonresponse or to apply poststratification adjustments (see Section 2.7), the survey data now include complex features that cannot be ignored in estimation and inference.

2.4.1 Relevance of SRS to Complex Sample Survey Data Analysis

So why is SRS even relevant for the material in this book? There are several reasons:

1. SRS designs produce samples that most closely approximate the assumptions (**i.i.d.**—observations are *independent* and *identical in distribution*) defining the theoretical basis for the estimation and inference procedures found in standard analysis programs in the major statistical software systems. Survey analysts who use the standard programs in Stata, SAS, and SPSS are essentially defaulting to the assumption that their survey data were collected under SRS. In general, the SRS assumption results in underestimation of variances of survey estimates of descriptive statistics and model parameters. Confidence intervals based on computed variances that assume independence of observations will be biased (generally too narrow), and design-based inferences will be affected accordingly. Likewise, test statistics (t, χ^2, F) computed in complex sample survey data analysis using standard programs will tend to be biased upward and overstate the significance of tests of effects.
2. The theoretical simplicity of SRS designs provides a basic framework for design-based estimation and inference on which to build a bridge to the more complicated approaches for complex samples.

3. SRS provides a comparative benchmark that can be used to evaluate the relative efficiency of the more complex designs that are common in survey practice.

Let's examine the second reason more closely, using SRS as a theoretical framework for design-based estimation and inference. In Section 2.5, we will turn to SRS as a benchmark for the efficiency of complex sample designs.

2.4.2 SRS Fundamentals: A Framework for Design-Based Inference

Many students of statistics were introduced to simple random sample designs through the example of an urn containing a population of blue and red balls. To estimate the proportion of blue balls in the urn, the instructor described a sequence of random draws of $i = 1, \ldots, n$ balls from the N balls in the urn. If a drawn ball was returned to the urn before the next draw was made, the sampling was "with replacement" (SRSWR). If a selected ball was not returned to the urn until all n random selections were completed, the sampling was "without replacement" (SRSWOR).

In each case, the SRSWR or SRSWOR sampling procedure assigned each population element an equal probability of sample selection, $f = n/N$. Furthermore, the overall probability that a specific ball was selected to the sample was independent of the probability of selection for any of the remaining $N - 1$ balls in the urn. Obviously, in survey practice random sampling is not typically performed by drawing balls from an urn. Instead, survey sampling uses devices such as tables of random numbers or computerized random number generators to select sample elements from the population.

Let's assume that the objective of the sample design was to estimate the mean of a characteristic, y, in the population:

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{2.1}$$

Under simple random sampling, an unbiased estimate of the population mean is the sample mean:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.2}$$

The important point to note here is that there is a true population parameter of interest, $\bar{Y}$, and the estimate of the parameter, $\bar{y}$, which can be derived from the sample data. The sample estimate $\bar{y}$ is subject to sampling variability, denoted as $Var(\bar{y})$, from one sample to the next. Another measure of the sampling variability in sample estimates is termed the standard error, or $SE(\bar{y}) = \sqrt{Var(\bar{y})}$. For simple random samples (SRS) selected from large populations, across all possible samples of size n, the standard error for the estimated population proportion is calculated as follows:

$$\begin{aligned} SE(\bar{y}) &= \sqrt{Var(\bar{y})} = \sqrt{(1-n/N)\cdot\frac{S^2}{n}} \\ &\approx \sqrt{\frac{S^2}{n}} \text{ if } N \text{ is large} \end{aligned} \tag{2.3}$$

where
$S^2 = \Sigma_{i=1}^{N}(Y_i - \bar{Y})^2/(N-1)$;
n = SRS sample size; and
N = the population size.

Since we observe only a single sample and not all possible samples of size n from the population of N, the true $SE(\bar{y})$ must be estimated from the data in our chosen sample:

$$\begin{aligned} se(\bar{y}) &= \sqrt{var(\bar{y})} = \sqrt{(1-n/N)\cdot\frac{s^2}{n}} \\ &\approx \sqrt{\frac{s^2}{n}} \text{ if } N \text{ is large} \end{aligned} \tag{2.4}$$

where
$s^2 = \Sigma_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$;
n = SRS sample size; and
N = the population size.

The term $(1 - n/N)$ in the expressions for $SE(\bar{y})$ and $se(\bar{y})$ is the **finite population correction (fpc)**. It applies only where selection of population elements is without replacement (see Theory Box 2.1) and is generally assumed to be equal to 1 in practice if $f = n/N < 0.05$.

If the sample size, n, is large, then under Neyman's (1934) method of design-based inference, a 95% confidence interval for the true population mean, $\bar{Y}$, can be constructed as follows:

$$\bar{y} \pm t_{0.975,n-1}\cdot se(\bar{y}) \tag{2.5}$$

THEORY BOX 2.1 THE FINITE POPULATION CORRECTION (fpc)

The fpc reflects the expected reduction in the sampling variance of a survey statistic due to sampling without replacement (WOR). For an SRS sample design, the fpc factor arises from the algebraic derivation of the expected sampling variance of a survey statistic over all possible WOR samples of size n that could be selected from the population of N elements (see Cochran, 1977).

In most practical survey sampling situations, the population size, N, is very large, and the ratio n/N is so close to zero that the fpc ~ 1.0. As a result, the fpc can be safely ignored in the estimation of the standard error of the sample estimate. Since complex samples may also employ sampling without replacement at one or more stages of selection, in theory, variance estimation for these designs should also include finite population corrections. Where applicable, software systems such as Stata and SUDAAN provide users with the flexibility to input population size information and incorporate the fpc values in variance estimation for complex sample survey data. Again, in most survey designs, the size of the population at each stage of sampling is so large that the fpc factors can be safely ignored.

2.4.3 An Example of Design-Based Inference under SRS

To illustrate the simple steps in design-based inference from a simple random sample, Table 2.1 presents a hypothetical sample data set of n = 32 observations from a very large national adult population (because the sampling fraction, n/N, is small, the fpc will be ignored). Each subject was asked to rate his or her view of the strength of the national economy (y) on a 0–100 scale, with 0 representing the weakest possible rating and 100 the strongest possible rating. The sample observations are drawn from a large population with population mean $\bar{Y} = 40$ and population variance $S_y^2 \cong 12.80^2 = 164$. The individual case identifiers for the sample observations are provided in Column (1) of Table 2.1. For the time being, we can ignore the columns labeled Stratum, Cluster, and Case Weight.

If we assume that the sample was selected by SRS, the sample estimates of the mean, its standard error, and the 95% confidence interval would be calculated as follows:

$$\bar{y} = \sum_{i=1}^{n} y_i / n = \sum_{i=1}^{32} y_i / 32 = 40.77$$

$$se(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{\sum_{i=1}^{32} (y_i - \bar{y})^2 / \ [n \cdot (n-1)]} = 2.41$$

$$95\%\ CI = \bar{y} \pm t_{.975,31} \cdot se(\bar{y}) = (35.87, 45.68)$$

TABLE 2.1

Sample Data Set for Sampling Plan Comparisons

Case No. (1)	Stratum (2)	Cluster (3)	Economy Rating score, y_i (4)	Case Weight, w_i (5)
1	1	1	52.8	1
2	1	1	32.5	2
3	1	1	56.6	2
4	1	1	47.0	1
5	1	2	37.3	1
6	1	2	57.0	1
7	1	2	54.2	2
8	1	2	71.5	2
9	2	3	27.7	1
10	2	3	42.3	2
11	2	3	32.2	2
12	2	3	35.4	1
13	2	4	48.8	1
14	2	4	66.8	1
15	2	4	55.8	2
16	2	4	37.5	2
17	3	5	49.4	2
18	3	5	14.9	1
19	3	5	37.3	1
20	3	5	41.0	2
21	3	6	45.9	2
22	3	6	39.9	2
23	3	6	33.5	1
24	3	6	54.9	1
25	4	7	26.4	2
26	4	7	31.6	2
27	4	7	32.9	1
28	4	7	11.1	1
29	4	8	30.7	2
30	4	8	33.9	1
31	4	8	37.7	1
32	4	8	28.1	2

The fundamental results illustrated here for SRS are that, for a sample of size n, unbiased estimates of the population value and the standard error of the sample estimate can be computed. For samples of reasonably large size, a confidence interval for the population parameter of interest can be derived. As our discussion transitions to more complex samples and more complex statistics we will build on this basic framework for constructing confidence intervals for population parameters, adapting each step to the features of the sample design and the analysis procedure.

2.5 Complex Sample Design Effects

Most practical sampling plans employed in scientific surveys are not SRS designs. Stratification is introduced to increase the statistical and administrative efficiency of the sample. Sample elements are selected from naturally occurring clusters of elements in multistage designs to reduce travel costs and improve interviewing efficiency. Disproportionate sampling of population elements may be used to increase the sample sizes for subpopulations of special interest, resulting in the need to employ weighting in the descriptive estimation of population statistics. All of these features of more commonly used sampling plans will have effects on the accuracy and precision of survey estimators, and we discuss those effects in this section.

2.5.1 Design Effect Ratio

Relative to SRS, the need to apply weights to complex sample survey data changes the approach to estimation of population statistics or model parameters. Also relative to SRS designs, stratification, clustering, and weighting all influence the size of standard errors for survey estimates. Figure 2.2 illustrates the general effects of these design features on the standard errors of survey estimates. The curve plotted in this figure represents the SRS standard

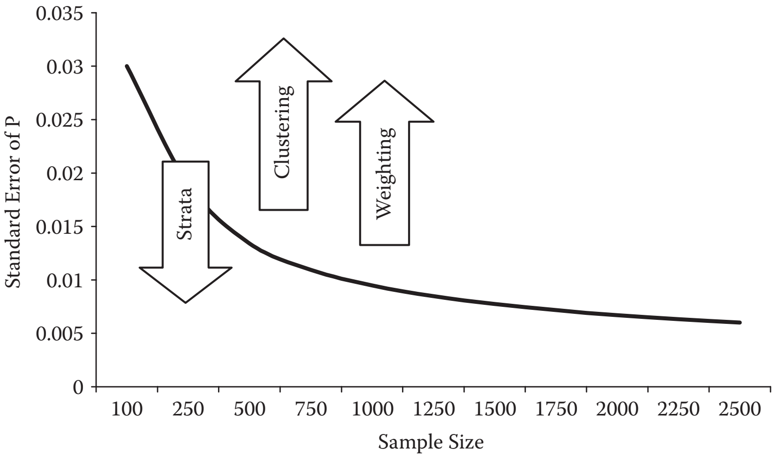

FIGURE 2.2
Complex sample design effects on standard errors.

error of a sample estimate of a proportion P as a function of the sample size n. At any chosen sample size, the effect of sample stratification is generally a reduction in standard errors relative to SRS. Clustering of sample elements and designs that require weighting for unbiased estimation generally tend to yield estimates with larger standard errors than an SRS sample of equal size.

Relative to an SRS of equal size, the complex effects of stratification, clustering, and weighting on the standard errors of estimates are termed the **design effect** and are measured by the following ratio (Kish, 1965):

$$D^2(\hat{\theta}) = \frac{SE(\hat{\theta})^2_{complex}}{SE(\hat{\theta})^2_{srs}} = \frac{Var(\hat{\theta})_{complex}}{Var(\hat{\theta})_{srs}} \tag{2.6}$$

where

$D^2(\hat{\theta})$ = the design effect for the sample estimate, $\hat{\theta}$;
$Var(\hat{\theta})_{complex}$ = the complex sample design variance of $\hat{\theta}$; *and*
$Var(\hat{\theta})_{srs}$ = the simple random sample variance of $\hat{\theta}$.

A somewhat simplistic but practically useful model of design effects that survey statisticians may use to plan a sample survey is

$$D^2(\hat{\theta}) \approx 1 + f(G_{strat}, L_{cluster}, L_{weighting})$$

where

G_{strat} = the relative gain in precision from stratified sampling compared to SRS;
$L_{cluster}$ = the relative loss of precision due to clustered selection of sample elements;
$L_{weighting}$ = the relative loss due to unequal weighting for sample elements.

The value of the design effect for a particular sample design will be the net effect of the combined influences of stratification, clustering, and weighting. In Sections 2.5 to 2.7 we will introduce very simple models that describe the nature and rough magnitude of the effects attributable to stratification, clustering, and weighting. In reality, the relative increase in variance measured by D^2 will be a complex and most likely nonlinear function of the influences of stratification, clustering, and weighting and their interactions. Over the years, there have been a number of attempts to analytically quantify the anticipated design effect for specific complex samples, estimates, and subpopulations (Skinner, Holt, and Smith, 1989). While these more advanced models are instructive, the sheer diversity in real-world survey designs and analysis objectives generally requires the empirical approach of estimating design effects directly from the available survey data:

$$d^2(\hat{\theta}) = \frac{se(\hat{\theta})^2_{complex}}{se(\hat{\theta})^2_{srs}} = \frac{var(\hat{\theta})_{complex}}{var(\hat{\theta})_{srs}} \tag{2.7}$$

where

$d^2(\hat{\theta})$ = the estimated design effect for the sample estimate, $\hat{\theta}$;
$var(\hat{\theta})_{complex}$ = the estimated complex sample design variance of $\hat{\theta}$; and
$var(\hat{\theta})_{srs}$ = the estimated simple random sample variance of $\hat{\theta}$.

As a statistical tool, the concept of the complex sample design effect is more directly useful to the designer of a survey sample than to the analyst of the survey data. The sample designer can use the concept and its component models to optimize the cost and error properties of specific design alternatives or to adjust simple random sample size computations for the design effect anticipated under a specific sampling plan (Kish, Groves, and Kotki, 1976). Using the methods and software presented in this book, the survey data analyst will compute confidence intervals and test statistics that incorporate the estimates of standard errors corrected for the complex sample design—generally bypassing the need to estimate the design effect ratio.

Nevertheless, knowledge of estimated design effects and the component factors does permit the analyst to gauge the extent to which the sampling plan for his or her data has produced efficiency losses relative to a simple random sampling standard and to identify features such as extreme clustering or weighting influences that might affect the stability of the inferences that he or she will draw from the analysis of the data. In addition, there are several analytical statistics such as the Rao–Scott Pearson χ^2 or likelihood ratio χ^2 where estimated design effects are used directly in adapting conventional hypothesis test statistics for the effects of the complex sample (see Chapter 6).

2.5.2 Generalized Design Effects and Effective Sample Sizes

The design effect statistic permits us to estimate the variance of complex sample estimates relative to the variance for an SRS of equal size:

$$\begin{aligned} var(\hat{\theta})_{complex} &= d^2(\hat{\theta}) \cdot var(\hat{\theta})_{srs}; \quad \text{or} \\ se(\hat{\theta})_{complex} &= \sqrt{d^2(\hat{\theta})} \cdot se(\hat{\theta})_{srs} \end{aligned} \tag{2.8}$$

Under the SRS assumption, the variances of many forms of sample estimates are approximately proportionate to the reciprocal of the sample size, that is, $var(\hat{\theta}) \propto 1/n$.

For example, if we ignore the fpc, the simple random sampling variances of estimates of a population proportion, mean, or simple linear regression coefficient are

$$var(p) = \frac{p(1-p)}{(n-1)}$$

$$var(\bar{y}) = \frac{s^2}{n}$$

$$var(\hat{\beta}) = \frac{\hat{\sigma}^2_{y.x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Before today's software was conveniently available to analysts, many public use survey data sets were released without the detailed stratification and cluster variables that are required for complex sample variance estimation. Instead, users were provided with tables of generalized design effects for key survey estimates that had been computed and summarized by the data producer. Users were instructed to perform analyses of the survey data using standard SAS, Stata, or SPSS programs under simple random sampling assumptions, to obtain SRS sampling variance estimates, and to then apply the design effect factor as shown in Equation 2.8 to approximate the correct complex sample variance estimate and corresponding confidence interval for the sample estimate. Even today, several major public use survey data sets, including the National Longitudinal Survey of Youth (NLSY) and the Monitoring the Future (MTF) Survey, require analysts to use this approach.

Survey designers make extensive use of design effects to translate between the simple analytical computations of sampling variance for SRS designs and the approximate variances expected from a complex design alternative. In working with clients, samplers may discuss design effect ratios, or they may choose a related measure of design efficiency termed the **effective sample size**:

$$n_{eff} = n_{complex} / d^2(\hat{\theta}) \tag{2.9}$$

where

n_{eff} = the effective sample size, or the number of SRS cases required to achieve the same sample precision as the actual complex sample design.

$n_{complex}$ = the actual or "nominal" sample size selected under the complex sample design.

The design effect ratio and effective sample size are, therefore, two means of expressing the precision of a complex sample design relative to an SRS of equal size. For a fixed sample size, the statements "the design effect for the proposed complex sample is 1.5" and "the complex sample of size $n = 1000$ has an effective sample size of $n_{eff} = 667$" are equivalent statements of the precision loss expected from the complex sample design.

2.6 Complex Samples: Clustering and Stratification

We already noted that survey data collections are rarely based on simple random samples. Instead, sample designs for large survey programs often feature stratification, clustering, and disproportionate sampling. Survey organizations use these "complex" design features to optimize the variance/cost ratio of the final design or to meet precision targets for subpopulations of the survey population. The authors' mentor, Leslie Kish (1965, 1987), was fond of creating classification systems for various aspects of the sample design process. One such system was a taxonomy of complex sample designs. Under the original taxonomy there were six binary keys to characterize all complex probability samples. Without loss of generality, we will focus on four of the six keys that are most relevant to the survey data analyst and aim to correctly identify the design features that are most important in applications:

Key 1: Is the sample selected in a single stage or multiple stages?

Key 2: Is clustering of elements used at one or more sample stages?

Key 3: Is stratification employed at one or more sample stages?

Key 4: Are elements selected with equal probabilities?

In the full realm of possible sample approaches this implies that there are at least 2^4 or 16 possible choices of general choices for complex sample designs. In fact, the number of complex sample designs encountered in practice is far fewer, and one complex design—multistage, stratified, cluster sampling with unequal probabilities of selection for elements—is used in most in-person surveys of household populations. Because they are so important in major programs of household population survey research, we will cover these multistage probability sampling plans in detail in Section 2.8. Before we do that, let's take a more basic look at the common complex sample design features of clustering, stratification, and weighting for unequal selection probabilities and nonresponse.

2.6.1 Clustered Sampling Plans

Clustered sampling of elements is a common feature of most complex sample survey data. In fact, to simplify our classification of sample designs it is possible to view population elements as "clusters of size 1." By treating elements as single-unit clusters, the general formulas for estimating statistics and standard errors for clustered samples can be applied to correctly estimate standard errors for the simpler stratified element samples (see Chapter 3).

Survey designers employ sample clustering for several reasons:

- Geographic clustering of elements for household surveys reduces interviewing costs by amortizing travel and related expenditures over a group of observations. By definition, multistage sample designs such as the area probability samples employed in the NCS-R, National Health and Nutrition Examination Survey (NHANES), and the Health and Retirement Study (HRS) incorporate clustering at one or more stages of the sample selection.
- Sample elements may not be individually identified on the available sampling frames but can be linked to aggregate cluster units (e.g., voters at precinct polling stations, students in colleges and universities). The available sampling frame often identifies only the cluster groupings. Identification of the sample elements requires an initial sampling of clusters and on-site work to select the elements for the survey interview.
- One or more stages of the sample are deliberately clustered to enable the estimation of multilevel models and components of variance in variables of interest (e.g., students in classes, classes within schools).

Therefore, while cluster sampling can reduce survey costs or simplify the logistics of the actual survey data collection, the survey data analyst must recognize that clustered selection of elements affects his or her approach to variance estimation and developing inferences from the sample data. In almost all cases, sampling plans that incorporate cluster sampling result in standard errors for survey estimates that are greater than those from an SRS of equal size; further, special approaches are required to estimate the correct standard errors. The SRS variance estimation formulae and approaches incorporated in the standard programs of most statistical software packages no longer apply, because they are based on assumptions of independence of the sample observations and sample observations from within the same cluster generally tend to be correlated (e.g., students within a classroom, or households within a neighborhood).

The appropriate choice of a variance estimator required to correctly reflect the effect of clustering on the standard errors of survey statistics depends on the answers to a number of questions:

1. Are all clusters equal in size?
2. Is the sample stratified?
3. Does the sample include multiple stages of selection?
4. Are units selected with unequal probability?

Fortunately, modern statistical software includes simple conventions that the analyst can use to specify the complex design features. Based on a simple set of user-supplied "design variables," the software selects the appropriate variance estimation formula and computes correct design-based estimates of standard errors.

The general increase in design effects due to either single-stage or multistage clustered sampling is caused by correlations (nonindependence) of observations within sample clusters. Many characteristics measured on sample elements within naturally occurring clusters, such as children in a school classroom or adults living in the same neighborhood, are correlated. Socioeconomic status, access to health care, political attitudes, and even environmental factors such as the weather are all examples of characteristics that individuals in sample clusters may share to a greater or lesser degree. When such group similarity is present, the amount of "statistical information" contained in a clustered sample of n persons is less than in an independently selected simple random sample of the same size. Hence, clustered sampling increases the standard errors of estimates relative to a SRS of equivalent size. A statistic that is frequently used to quantify the amount of homogeneity that exists within sample clusters is the **intraclass correlation**, ρ (Kish, 1965). See Kish et al. (1976) for an in-depth discussion of intraclass correlations observed in the World Fertility Surveys.

A simple example is useful to explain why intraclass correlation influences the amount of "statistical information" in clustered samples. A researcher has designed a study that will select a sample of students within a school district and collect survey measures from the students. The objective of the survey is to estimate characteristics of the full student body and the instructional environment. A probability sample of n = 1,000 is chosen by randomly selecting 40 classrooms of 25 students each. Two questions are asked of the students:

1. What is your mother's highest level of completed education? Given the degree of socioeconomic clustering that can exist even among schools within a district, it is reasonable to expect that the intraclass correlation for this response variable is positive, possibly as high as $\rho = 0.2$.
2. What is your teacher's highest level of completed education? Assuming students in a sampled class have only one teacher, the intraclass correlation for this measure is 1.0. The researcher need not ask the question of all n = 1,000 students in the sample. An

identical amount of information could be obtained by asking a single truthful student from each sample classroom. The effective sample size for this question would be 40, or the number of sample classrooms.

When the primary objective of a survey is to estimate proportions or means of population characteristics, the following model can be used to approximate the design effect that is attributable to the clustered sample selection (Kish, 1965):

$$D^2(\bar{y}) = 1 + L_{cluster} \approx 1 + \rho \cdot (B - 1) \tag{2.10}$$

where
ρ = the intraclass correlation for the characteristic of interest;
B = the size of each cluster or primary sampling unit (PSU).

The value of ρ is specific to the population characteristic (e.g., income, low-density cholesterol level, candidate choice) and the size of the clusters (e.g., counties, enumeration areas [EAs], schools, classrooms) over which it is measured. Generally, the value of ρ decreases as the geographical size and scope (i.e., the heterogeneity) of the cluster increases. Typical values of ρ observed for general population characteristics range from .00 to .20 with most between .005 and .100 (Kish et al., 1976).

In practice, it is sometimes valuable to estimate the value of ρ for a survey characteristic y. While it is theoretically possible to estimate ρ as a function of the sample data (y_i, $i = 1, \ldots, n$) and the sample cluster labels ($\alpha = 1, \ldots, a$), in most survey designs such direct estimates tend to be unstable. Kish (1965) suggested a synthetic estimator of ρ, which he labeled the **rate of homogeneity (*roh*)**. Using the simple model for the estimated design effect for a sample mean, the expression is rearranged to solve for *roh* as

$$roh(y) = \frac{d^2(\bar{y}) - 1}{\bar{b} - 1} \tag{2.11}$$

where $\bar{b}$ is the average sample size per cluster.

Algebraically, this synthetic estimate of the intraclass correlation is restricted to a range of values from $-1/(\bar{b} - 1)$ to 1.0. Since ρ is almost always positive for most survey variables of interest, this restricted range of the synthetic approximation *roh* is not a problem in practice. As the analyst (and not the designer) of the sample survey, the calculation of *roh* is not essential to the reader's work; however, when the reader begins to work with a new survey data set, it can be valuable to compute the *roh* approximation for a few key survey variables simply to get a feel for the level of intraclass correlation in the variables of interest.

Let's return to the hypothetical sample data set provided in Table 2.1. Previously, we assumed a simple random sample of size n = 32 and estimated the mean, the SRS standard error, and a 95% confidence interval for the mean of the economic conditions index. Now, instead of assuming SRS, let's assume that eight clusters, each of size four elements, were selected with equal probability. The cluster codes for the individual cases are provided in Column (3) of Table 2.1. Without digressing to the specific estimators of the mean and variance for this simple clustered sample (see Cochran, 1977), let's look at the correct numerical values for the estimated mean, standard error of the mean, and the 95% CI for $\bar{Y}$:

$$\bar{y}_{cl} = 40.77,\ se_{cl}(\bar{y}) = 3.65,\ CI(\bar{y}_{cl}) = (32.12, 49.49)$$

Comparing these results to those for the SRS example based on the same data (Section 2.4.3), the first thing we observe is that the estimate of $\bar{y}$ is identical. The change that occurs when clustering is introduced to the example is that the standard error of the mean is increased (due to the presence of the intraclass correlation that was used in developing the clusters for this example). The width of the 95% CI is also increased, due to the increase in standard error and the reduction in the degrees of freedom for the Student t statistic used to develop the CI limits (degrees of freedom determination for complex sample designs will be covered in Chapter 3). The relative increase in $se(\bar{y})$ for the simple cluster sample compared with an SRS of n = 32 and the synthetic estimate of the intraclass correlation are computed as follows:

$$d(\bar{y}) = \sqrt{d^2(\bar{y})} = \frac{se_{cl}(\bar{y})}{se_{srs}(\bar{y})} = \frac{3.65}{2.41} = 1.51$$

$$roh(y) \approx \frac{1.51^2 - 1}{4 - 1} = 0.43$$

2.6.2 Stratification

Strata are nonoverlapping, homogeneous groupings of population elements or clusters of elements that are formed by the sample designer prior to the selection of the probability sample. In multi-stage sample designs (see Section 2.8), a different stratification of units can be employed at each separate stage of the sample selection. Stratification can be used to sample elements or clusters of elements. As already noted, if elements are viewed as "clusters of size B = 1" then the general formulas for estimating statistics and standard errors for stratified clustered samples can be applied to correctly compute these statistics for the simpler stratified element samples (see Chapter 3). To take this "unification" one step further, if the sample design does not incorporate explicit stratification, it may be viewed as sampling from a single stratum (i.e., $H = 1$).

In survey practice, stratified sampling serves several purposes:

- Relative to an SRS of equal size, stratified samples that employ **proportional allocation** or **optimal allocation** of the sample size to the individual strata have smaller standard errors for sample estimates (Cochran, 1977).
- Stratification provides the survey statistician with a convenient framework to disproportionately allocate the sample to subpopulations, that is, to oversample specific subpopulations to ensure sufficient sample sizes for analysis.
- Stratification of the probability sample can facilitate the use of different survey methods or procedures in the separate strata (see Chapter 3).

Every stratified sample design involves the following four steps in sample selection and data analysis:

1. Strata ($h = 1, \ldots, H$) of N_h clusters/elements are formed.
2. Probability samples of a_h clusters or $a_h = n_h$ elements are independently selected from each stratum.
3. Separate estimates of the statistic of interest are computed for sample cases in each stratum and then weighted and combined to form the total population estimate.
4. Sampling variances of sample estimates are computed separately for each stratum and then weighted and combined to form the estimate of sampling variance for the total population estimate.

Because stratified sampling selects independent samples from each of the $h = 1, \ldots, H$ explicit strata of relative size $W_h = N_h/N$, any variance attributable to differences among strata is eliminated from the sampling variance of the estimate. Hence, the goal of any stratification designed to increase sample precision is to form strata that are "homogeneous within" and "heterogeneous between": Units assigned to a stratum are like one another and different from those in other strata, in terms of the measurements collected. To illustrate how stratification works to reduce sampling variance, let's consider the case of a simple stratified random sample of size

$$n = \sum_{h=1}^{H} n_h$$

where the stratum sample sizes are proportional to the individual strata: $n_h = n \cdot N_h / N$. This **proportionate allocation** of the sample ensures that each population element has an equal probability of being included in the sample, denoted by $f_h = n_n / N_h = n / N = f$. Let's compare $var(\bar{y}_{st,pr})$ for this stratified sample with $var(\bar{y}_{srs})$, ignoring the fpc:

$$\begin{aligned}
\Delta &= var(\bar{y}_{srs}) - var(\bar{y}_{st,pr}) \\
&= \frac{S^2_{total}}{n} - \frac{S^2_{within}}{n} = \frac{\left[S^2_{within} + S^2_{between}\right]}{n} - \frac{S^2_{within}}{n} \\
&= \frac{S^2_{between}}{n} \\
&= \frac{\sum_{h=1}^{H} W_h(\bar{Y}_h - \bar{Y})^2}{n}
\end{aligned} \tag{2.12}$$

The simple random sample of size n includes the variance of y both within the strata, S^2_{within}, and between the strata, $S^2_{between}$, but the stratified sample has eliminated the between-stratum variance component and consequently yields a sampling variance for the mean that is less than or equal to that for an SRS of equal size. The expected amount of the reduction in the variance of the mean is a weighted function of the squared differences between the stratum means, $\bar{Y}_h$, and the overall population mean, $\bar{Y}$. Hence the importance of forming strata that differ as much as possible in the $h = 1, \ldots, H$ values of the $\bar{Y}_h$.

Let's return to the example data set in Table 2.1 and assume that the $n = 32$ observations were selected from four population strata of equal size, such that $W_h = N_h/N = 0.25$. The stratum identifiers for each sample case are provided in Column (2) of Table 2.1. Under this proportionately allocated, stratified random sample design, we compute the following:

$$\begin{aligned}
\bar{y}_{st,pr} &= \sum_{h=1}^{H} W_h \cdot \bar{y}_h \\
&= 0.25 \times 51.1 + 0.25 \times 43.32 + 0.25 \times 39.60 + 0.25 \times 29.04 \\
&= 40.77
\end{aligned}$$

$$\begin{aligned}
se(\bar{y}_{st,pr}) &= \sqrt{var(\bar{y}_{st,pr})} = \sqrt{\sum_{h=1}^{H} W_h^2 s_h^2 / n_h} \\
&= 2.04
\end{aligned}$$

$$d(\bar{y}_{st,pr}) = \frac{se(\bar{y}_{st,pr})}{se(\bar{y}_{srs})} = \frac{2.04}{2.41} = 0.85$$

$$d^2(\bar{y}_{st,pr}) = 0.85^2 = 0.72$$

$$n_{eff} = n / d^2(\bar{y}_{st,pr}) = 32 / 0.72 = 44.44$$

Note that the unbiased sample estimate, $\bar{y}_{st,pr} = 40.77$, is identical to that obtained when we assumed that the 32 observations were sampled using SRS, and also identical to that when we assumed that the 32 observations were obtained from a cluster sample. The estimated standard error, $se(\bar{y}_{st,pr}) = 2.04$, is much smaller, resulting in an estimated design effect ratio of 0.72. For this example, it appears that the stratified sample of size $n = 32$ (eight cases from four strata) yields the sample precision of an SRS of more than 44 cases.

The simple algebraic illustration and previous exercise calculation demonstrate how stratification of the sample can lead to reductions in the variance of sample estimates relative to SRS. In contrast to this simple example, the precision gains actually achieved by stratification, G_{strat}, in complex samples are difficult to determine analytically. The gains will be a function of the stratum means, the variances of y within individual strata, and whether the sampling within strata includes additional stages, clustering, unequal probabilities of selection, or other complex sample design features.

2.6.3 Joint Effects of Sample Stratification and Clustering

To see how the two design features of stratification and clustering contribute to the design effect for a complex sample, let's revisit the sample data in Table 2.1, treating it as a stratified sample from $h = 1, \ldots, 4$ strata. Within each stratum, a proportionately allocated sample of two clusters ($a_h = 2$) of size 4 are selected with equal probability. This is a very simple example of a stratified "**two-per-stratum**" or "paired selection" cluster sample design that is very common is survey practice (see Chapter 4). Table 2.2 compares

TABLE 2.2

Sample Estimates of the Population Mean and Related Measures of Sampling Variance under Four Alternative Sample Designs

Sample Design Scenario	Estimator	$\bar{y}$	$se(\bar{y})$	$d(\bar{y})$	$d^2(\bar{y})$	n_{eff}
SRS	$\bar{y}_{srs}$	40.77	2.41	1.00	1.00	32
Clustered	$\bar{y}_{cl}$	40.77	3.66	1.51	2.31	13.9
Stratified	$\bar{y}_{st}$	40.77	2.04	0.85	0.72	44.4
Stratified, clustered	$\bar{y}_{st,cl}$	40.77	2.76	1.15	1.31	24.4

Source: Table 2.1.

the results for this stratified, cluster sample design scenario with those for the simpler SRS, clustered, and stratified designs considered earlier.

Note that for the stratified, cluster sample design scenario the estimated design effect falls between the high value for the clustered-only scenario and the low value for the scenario where only stratification effects are considered. As illustrated in Figure 2.2, real survey designs result in a "tug of war" between the variance inflation due to clustering and the variance reduction due to stratification. From Table 2.2, the design effect for the stratified, clustered sample is greater than 1, indicating a net loss of precision relative to an SRS of size n—clustering "won" the contest. Estimates of design effects greater than 1 are common for most complex sample designs that combine stratification and clustered sample selection. In population surveys, large stratification gains are difficult to achieve unless the survey designer is able to select the sample from a frame that contains rich ancillary data, $\boldsymbol{x}$, that are highly correlated with the survey variables, $\boldsymbol{y}$. The survey data analyst needs to be aware of these design effects on the precision of sample estimates. Correctly recognizing the stratum and cluster codes for sample respondents provided by a data producer for use by data analysts is essential for correct and unbiased analyses of the survey data. We discuss the importance of identifying these codes in a survey data set in Chapter 4.

2.7 Weighting in Analysis of Survey Data

2.7.1 Introduction to Weighted Analysis of Survey Data

When probability sampling is used in survey practice, it is common to find that sample inclusion probabilities for individual observations vary. Weighting of the survey data is thus required to "map" the sample back to an unbiased representation of the survey population. A simple but useful device for "visualizing" the role of case-specific weights in survey data analysis is to consider the weight as the number (or share) of the population elements that is represented by the sample observation. Observation i, sampled with probability $f_i = 1/10$, represents 10 individuals in the population (herself and nine others). Observation j, selected with probability $f_j = 1/20$, represents 20 population elements. In fact, if each sample case is assigned a weight equal to the reciprocal of its probability of selection, $w_i = 1/f_i$, the sum of the sample weights will in expectation equal the population size:

$$E\left(\sum_{i=1}^{n} w_i\right) = N$$

As a survey data analyst, the reader can use this fact during his or her pre-analysis "checklist" to understand the distribution of the sampling weights that the data producer has provided (see Chapter 4).

Generally, the final analysis weights in survey data sets are the product of the sample selection weight, w_{sel}, a nonresponse adjustment factor, w_{nr}, and the poststratification factor, w_{ps}:

$$w_{final,i} = w_{sel,i} \times w_{nr,i} \times w_{ps,i} \tag{2.13}$$

The sample selection weight factor (or **base weight**) is simply the reciprocal of the probability that a population element was chosen to the sample, $w_{sel,i} = 1/f_i$. Under the theory of design-based inference for probability samples, weighted estimation using these "inverse probability" weight factors will yield unbiased (or nearly unbiased) estimates of population statistics. For example,

$$\begin{aligned} \bar{y}_w &= \frac{\sum_{i=1}^{n} w_{sel,i} \cdot y_i}{\sum_{i=1}^{n} w_{sel,i}} \text{ is unbiased for } \bar{Y} \\ s_w^2 &= \frac{\sum_{i=1}^{n} w_{sel,i} \cdot (y_i - \bar{y})^2}{\sum_{i=1}^{n} w_{sel,i} - 1} \text{ is an estimate of } S^2 \end{aligned} \tag{2.14}$$

Note how in each of these estimators, the selection weight "expands" each sample observation's contribution to reflect its share of the population representation. Throughout this book, weighted estimation of population parameters will follow a similar approach, even for procedures as complex as the pseudo-maximum likelihood (PML) estimation of the coefficients in a multivariate logistic regression model.

If we were able to always complete an observation on each selected sample case, the development of the survey analysis weights would be very straightforward. The computation of the weight would require only knowledge of the sample selection probability factors (see Section 2.7.2). In a carefully designed and well-managed survey, these probability factors are known and carefully recorded for each sample case. The computation of w_{sel} is therefore an accounting function, requiring only multiplication of the probabilities of selection at each stage of sampling and then taking the reciprocal of the product of the probabilities.

As we will see in Chapter 4, **survey data producers**, who in most cases have the responsibility for developing the final analysis weights have a more difficult task than the simple arithmetic required to compute w_{sel}. First, weighting by w_{sel} will yield only unbiased estimators of population parameters if all n elements in the original sample are observed. Unfortunately, due to **survey nonresponse**, observations are collected only for r cases of the original probability sample of n elements (where $r <= n$). Therefore, survey data producers must develop statistical models of the conditional probability that a sample element will be an observed case. In general terms, the nonresponse adjustment factor, w_{nr}, in the analysis weight is the reciprocal of the estimated conditional probability that the sample case responds. The objective in applying nonresponse factors in survey weights is to attenuate bias due to differential nonresponse across sample elements. A price that may be paid for the bias reduction through nonresponse weighting takes the form of increases in standard errors for the weighted estimates. The potential magnitude of the increases in standard errors is discussed next in the context of weighting for unequal selection probabilities.

Even after computing w_{sel} and adjusting for nonresponse through the factor w_{nr}, most survey data producers also introduce the poststratification factor w_{ps} into the final weight. As its label implies, the poststratification factor is an attempt to apply stratification corrections to the observed sample "post" or after the survey data have been collected. While not as effective as estimation for samples that were stratified at the time that they were selected, the use of poststratification weight factors can lead to reduced standard errors (variance) for sample estimates. Furthermore, poststratification weighting can attenuate any sampling biases that may have entered the original sample selection due to sample frame noncoverage or omissions that occurred in implementing the sample plan. An example of the latter would be systematic underreporting of young males in the creation of household rosters for the selection of the survey respondent.

In the sections that follow, we describe the development of these three important weighting factors in more detail.

2.7.2 Weighting for Probabilities of Selection

As previously described, the selection weight factor, w_{sel}, is introduced in the calculation of the final analysis weight to account for the probability that a case was selected for the sample. Common reasons for varying probabilities of case selection in sample surveys include the following:

1. Disproportionate sampling within strata to achieve an optimally allocated sample for a specific population estimate (Cochran, 1977).

2. Disproportionate sampling within a stratum or group of strata to deliberately increase the sample size and precision of analysis for certain domains of the survey population.
3. The use of sample screening across the sample strata and clusters to identify and differentially sample subpopulations, such as the NHANES oversampling of household members with disabilities.
4. Unequal probabilities that arise from subsampling of observational units within sample clusters, such as the common procedure of selecting a single random respondent from the eligible members of sample households.
5. Unequal probabilities that reflect information on final sampling probability that can be obtained only in the process of the survey data collection, for example, in a **random digit dialing (RDD)** telephone sample survey, the number of distinct landline telephone numbers that serve a respondent's household.

Due to the multipurpose nature of modern population surveys (many variables, many statistical aims), the use of disproportionate sampling to optimize a sample design for purposes of estimating a single population parameter is rare in survey practice. Far more common are designs in which geographic domains or subpopulations of the survey population are oversampled to boost the sample size and precision for stand-alone or comparative analysis. For example, the HRS employs a roughly two-fold oversampling of age-eligible African Americans and Hispanic individuals to improve precision of cross-sectional and longitudinal analyses of these two population subgroups (Heeringa and Connor, 1995). The NCS-R employs a Part 1 screening interview to obtain profiles of respondents' mental health related symptoms but disproportionately samples persons with positive symptom counts for the more intensive Part 2 questionnaire (Kessler et al., 2004). As these examples illustrate, the selection probability for individual observational units often includes multiple components—for example, a base factor for housing unit selection, a factor for subsampling targeted groups within the full population, and a probability of selecting a single random respondent within the household. For HRS sample respondents, the selection weight factor, w_{sel}, is computed as the product of the reciprocals of three probabilities: (1) $w_{sel,hh}$, the reciprocal of the multistage probability of selecting the respondent's housing unit (HU) from the area frame (see Section 2.8); (2) $w_{sel,sub}$, the reciprocal of the probability of retaining the sample HU under the design objective of a 2:1 oversampling of eligible African American and Hispanic adults; and (3) $w_{sel,resp}$, the reciprocal of the conditional probability of selecting a respondent within the eligible household. This calculation is illustrated as follows:

TABLE 2.3

Illustration of Example w_{sel} Computations Based on the HRS

Description of Sample Case						
Sample ID	**Race/Ethnicity of Respondent**	**Eligible Rs/Household**	$w_{sel,hh}$	$w_{sel,sub}$	$w_{sel,resp}$	w_{sel}
A	Black	2	2000	1	2	4000
B	Black	1	2000	1	1	2000
C	Hispanic	2	2000	1	2	4000
D	White	1	2000	2	1	4000

$$f_{sel} = f_{sel,hh} \times f_{sel,sub} \times f_{sel,resp}$$

$$\Rightarrow$$

$$w_{sel} = \frac{1}{f_{sel}} = w_{sel,hh} \times w_{sel,sub} \times w_{sel,resp}$$

Drawing on the general features of the HRS sample selection, Table 2.3 illustrates the sample selection weight calculations for four hypothetical respondents from different race/ethnicity groups who have different numbers of eligible respondents in their sample households.

2.7.3 Nonresponse Adjustment Weights

As described already, the most widely accepted approach to compensate for **unit nonresponse** in surveys is for the data producer to develop and apply a nonresponse adjustment factor to the sample selection weight that is used in analysis. Underlying the weighting adjustment for nonresponse is a model of the **response propensity**—conditional on sample selection, the probability that the unit will cooperate in the survey request. In a sense, the concept of response propensity treats response to the survey as another step in the "sample selection process." But unlike true sample selection in which the sampling statistician predetermines the sampling probability for each unit, an underlying propensity model—for the most part outside the control of a statistician—determines the probability that a sampled case will be observed. The multiplication of the original sample selection weight for each sample unit by the reciprocal of its modeled response propensity creates a new weight, which, if the model is correct, enables unbiased or nearly unbiased estimation of population statistics from the survey data.

Two related methods for estimating response propensity and computing the nonresponse adjustment are commonly used in survey practice: a simple weighting class adjustment method and a propensity score weighting approach.

2.7.3.1 Weighting Class Approach

The "weighting class method" (Little and Rubin, 2002) assigns all eligible elements of the original sample—survey respondents and nonrespondents—to classes or cells based on categorical variables (e.g., age categories, gender, region, sample stratum) that are predictive of response rates. Original sample cases that were found to be ineligible for the survey are excluded from the nonresponse adjustment calculation. The weighting class method makes the simple assumption that the response propensities for cases within a given weighting class cell are equal (i.e., respondents are MAR, or missing at random, conditional on the categorical variables and interactions implicit in the defined cells). The common response propensity for cases in a cell is estimated by the empirical response rate for the cases assigned to that cell. The weighting class nonresponse adjustment is then computed as the reciprocal of the response rate for the cell $c = 1, ..., C$ to which the case was assigned:

$$w_{nr,wc,i} = \frac{1}{rrate_c} \tag{2.15}$$

where $rrate_c$ = the response rate for weighting class $c = 1, \ldots, C$.

To be effective at reducing potential bias due to differential nonresponse across sample cases, the variables chosen to create the weighting classes must be predictive of survey participation rates. Recent work (Little and Vartivarian, 2005) has shown that reductions in both the bias and variance of an estimate are possible if weighting classes are formed based on variables related to *both* response propensity and the survey variable of interest.

2.7.3.2 Propensity Cell Adjustment Approach

The propensity cell weighting approach also assigns an adjustment factor to each respondent's sample selection weight that is equal to the reciprocal of the estimated probability that they participated in the survey. However, in the propensity adjustment method, the assignment of cases to adjustment cells is based on individual response propensity values estimated (via logit transform) from a logistic regression model:

$$\hat{p}_{resp,i} = prob(respondent = yes \mid X_i) = \left(\frac{e^{X_i\hat{\beta}}}{1+e^{X_i\hat{\beta}}} \right) \tag{2.16}$$

THEORY BOX 2.2 ESTIMATING RESPONSE PROPENSITIES IN NR ADJUSTMENTS

In our discussion, both the weighting class and propensity cell adjustments for nonresponse require estimation of response rates within defined cells. Traditionally, many survey statisticians have employed the sample selection weight, w_{sel}, to compute weighted estimates of response rates in these cells:

$$rrate_{weighted,cell} = \sum_{i \in cell}^{n_{cell}} w_{sel,i} \cdot y_i \Big/ \sum_{i \in cell}^{n_{cell}} w_{sel,i} \quad (2.18)$$

where

y_i = 1 if sample person i is a respondent, 0 otherwise;
$w_{sel,i}$ = the sample selection weight for case i;
n_{cell} = the total count of eligible sample cases in cell $c = 1, \ldots, C$.

Recently, Little and Vartivarian (2005) have suggested that the unweighted response rate may be the preferred estimate of the response propensity for the cell, depending on what design variables are used to construct the cells:

$$rrate_{cell} = \sum_{i \in cell}^{n_{cell}} y_i \,/\, n_{cell} \quad (2.19)$$

Note that under the propensity modeling approach, the computed value of the empirical response rate for cases in a modeled propensity quantile need not always fall within the range of scores for the quantile. For example, the range of propensity scores used to define an adjustment cell might be 0.75–0.79, but the corresponding response rate for cases in the cell could be 0.73. This is a reflection of lack of fit in the propensity model. Standard tests may be used to evaluate the goodness of fit of the logistic regression model of response propensity. We recommend including as many theoretical predictors of responding as possible in the model, and especially predictors that are also correlated with the survey variables of primary interest (Little and Vartivarian, 2005).

where
X_i = a vector of values of response predictors for $i = 1,...,n$;
$\hat{\beta}$ = the corresponding vector of estimated logistic regression coefficients.

Adjustment cells are then defined based on quantile ranges—often deciles, such as 0–0.099 or 0.10–0.199—of the distribution of response propensities (Little and Rubin, 2002). The propensity score nonresponse weighting adjustments are then computed as the reciprocal of the response rate in the quantile range cell $d = 1, ..., D$ to which the case was assigned:

$$w_{nr,prop,i} = \frac{1}{rrate_d} \tag{2.17}$$

where $rrate_d$ = the weighted response rate for propensity cell $d = 1, \ldots,$ D.

The use of individual estimated response propensities for these adjustments (i.e., $\hat{p}_{resp,i}$ as opposed to the response rates, $rrate_d$, within the quantile range cells, as shown in (2.17)) is often not recommended, as extremely low estimated response propensities (e.g., $\hat{p}_{resp,i} = 0.001$) could result in more variance in the weights. This increased variance in the weights would decrease the precision of weighted estimates based on the survey data (see Section 2.7.5).

To employ either method for nonresponse adjustment, the characteristics used to define the cells or model the response propensities must be known for both respondents and nonrespondents. In cross-sectional sample surveys such as the NCS-R or NHANES, this limits the nonresponse adjustment to characteristics of sample persons or households that are known from the sampling frame or are completely observed in the screening process. In the case of the 1992 HRS baseline sample, a simple weighting class adjustment approach was employed to develop the nonresponse adjustment (Heeringa and Connor, 1995). Region (Northeast, South, Midwest, West), urban/rural status of the primary stage unit (PSU), and race of the respondent (Black, Hispanic, White, and Other) were used to define weighting class adjustment cells. See Table 2.4 for an illustration of typical values of w_{nr} for various cells and how the values of this adjustment factor influence the final composite weight.

2.7.4 Poststratification Weight Factors

The nonresponse adjustment procedures just described have the property that only data available for sampled respondent and nonrespondent cases are used to compute weighting adjustments. Another weighting technique used in practice to improve the quality of sample survey estimates is to incorporate known information on the full survey population—borrowing strength from data sources external to the sample. **Poststratification** is one

TABLE 2.4

Illustration of Example w_{final} Computations Including w_{sel}, w_{nr}, and w_{ps} Based on the HRS

Description of Sample Case						
Sample ID	Nonresponse Adjustment Cell	Poststratum	w_{sel} [a]	w_{nr}	w_{ps}	w_{final}
A	Black, Northeast, urban	Age 50–54, male, Northeast, Black	4000	1.3	1.04	5408
B	Black, South, rural	Age 55–61, female, South, Black	2000	1.15	.96	2208
C	Hispanic, West, urban	Age 50–54, male, West, Hispanic	4000	1.25	1.06	5300
D	White, Midwest, rural	Age 55–61, female, Midwest, White	4000	1.18	.97	4578

[a] From Table 2.3.

such method for using population data in survey estimation. **Raking ratio estimation**, **generalized regression (GREG) estimation**, and **calibration** are other forms of postsurvey weight adjustment that may be employed to improve the precision and accuracy of survey estimates. Here we will focus on poststratification, the most common technique applied in general social and health survey weighting.

Simple poststratification forms $l = 1, \ldots, L$ poststrata of respondent cases (just as the sample designer might form $h = 1, \ldots, H$ design strata prior to sample selection). The criteria used to select variables for forming poststrata include (1) variables such as age, gender, and region that define poststrata for which accurate population control totals are available from external sources; (2) poststratification variables that are highly correlated with key survey variables; and (3) variables that may be predictive of noncoverage in the sample frame. Cross-classifications of these variables are then used to define the $l = 1,\ldots, L$ poststrata. To ensure efficiency in the poststratification, poststrata are generally required to include a minimum of $n_l = 15$ to 25 observations. If the cross-classification of poststratification variables results in cells with fewer than this minimum number, similar poststrata are collapsed to form a larger grouping.

Poststratification weighting involves adjusting the final weights for respondent sample cases so that weighted sample distributions conform to the known population distributions across the $l = 1, \ldots, L$ poststrata. The poststratification factor applied to each respondent weight is computed as follows:

$$w_{ps,l,i} = \frac{N_l}{\sum_{i=1}^{n_l}(w_{sel,i} \times w_{nr,i})} = \frac{N_l}{\hat{N}_l} \tag{2.20}$$

where

$w_{ps,1,i}$ = the post-stratification weight for cases in post-stratum $l = 1, \ldots, L$; and

N_l = the population count in post-stratum l obtained from a recent Census, administrative records, or a large survey with small sampling variance.

In the case of the 1992 HRS baseline sample, poststrata of sample respondents were defined based on age category (50–54, 55–61) for the 1931–1941 birth cohorts, gender, race/ethnicity, and Census region. Population control totals, N_l, for each poststratum were obtained as weighted population totals from the March 1992 demographic supplement to the U.S. Current Population Survey (Heeringa and Connor, 1995). See Table 2.4 for an illustration of typical values of w_{ps} for various poststrata and how the values of this adjustment factor influence the final composite sampling weight (w_{final}).

2.7.5 Design Effects Due to Weighted Analysis

The weights that readers will be using in their survey data analysis are therefore compound products of several adjustments—unequal probabilities of selection, differential patterns of response, and poststratification. Each of these factors can have a different effect on the precision and bias reduction for a design-based population estimate. Along with sample stratification and clustering, weighted estimation contributes to the final design effect for a survey estimate. In Figure 2.2 shown earlier in this chapter, the large arrow representing the net effect of weighting could be split into three smaller arrows: the small arrows for selection weighting and nonresponse adjustment pointing up to increased variance and the small poststratification arrow pointing down to lower variance. It is analytically impossible in most surveys to partition out the size of the contribution of each factor (stratification, clustering, selection weights, nonresponse adjustment, poststratification weighting) to the variance of sample estimates.

Changes in standard errors due to weighting are related to the variance of the weight values assigned to the individual cases and the correlations of the weights with the values and standard deviations of the variables of interest. We generally find in survey practice that the net effect of weighted estimation is inflation in the standard errors of estimates. This reflects the empirical fact that through the sequence of steps (e.g., selection weighting, nonresponse weighting) case weights can be quite variable. Furthermore, most survey weights are at best only moderately correlated with the distributional properties of the survey variables.

In recent years, the survey literature has used the term *weighting loss* to describe the inflation in variances of sample estimates that can be attributed to weighting. A simple approximation used by sampling statisticians to anticipate $L_{weighting}$, the proportional increase in variance of the sample mean due to weighted estimation, is

$$L_{weighting}(\bar{y}) \approx cv^2(w) = \frac{\sigma^2(w)}{\bar{w}^2} = \left\{ \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i \right)^2} \cdot n \right\} - 1 \tag{2.21}$$

where

$cv^2(w)$ = the relative variance of the sample weights;

$s(w)$ = the standard deviation of the sample weights; and

$\bar{w}$ = the mean of the sample weights.

This simple model of weighting loss was introduced by Kish (1965), and despite its widespread use, it was intended more as a design tool than as a strong model of true weighting effects on variances of sample estimates. The Kish model of weighting loss was originally presented in the context of proportionate stratified sampling and represented the proportional increase in the variance of means (and proportions as means of binary variables) due to arbitrary disproportionate sampling of strata. It assumes that proportionate allocation is the optimal stratified design (i.e., variances of y are approximately equal in all strata) and that the weights are uncorrelated with the values of the random variable y. Little and Vartivarian (2005) show clear examples where this simple model of weighting loss breaks down. As in any model, the quality of the predictions is tied to how closely the data scenario matches the model assumptions. Even within a survey data set, there may be variables for which the weighting actually improves the precision of estimates (due to intended or chance optimality for that variable) and many others for which the variability and randomness of the weights simply produces an increase in the variance of estimates.

To illustrate the application of Kish's (1965) weighting loss model to a sample design problem, consider the following example. A survey is planned for a target population that includes both urban and rural populations of children. A total of 80% of the target population lives in the urban domain, and 20% lives in the rural villages. The agency sponsoring the survey is interested in measuring the mean body mass index (BMI) of these children. The agency would like to have roughly equal precision for mean estimates for urban and rural children and decides to allocate the sample equally (50:50) to the two geographic domains (or strata). They recognize that this will require weighting to obtain unbiased estimates for the *combined* area. Urban cases will need to be weighted up by a factor proportional to 0.80/0.50 = 1.6, and rural cases will need to be down-weighted by a factor proportional to 0.20/0.50 = 0.4. To estimate $L_{weighting}$, they compute the relative variance of these weights for a sample that is

50% urban and 50% rural. They determine that $L_{weighting} \approx cv^2(w) = 0.36$. Ignoring any clustering that may be included in the sample plan, this implies that the final sample size for the survey must be $n_{complex} = n_{srs} \times 1.36$, or 36% larger than the SRS sample size required to meet a set precision level for a estimating mean BMI for the combined population of urban and rural children.

The reader should note that $L_{weighting}$ is often nontrivial. Survey data sets that include disproportionate sampling of geographic areas or other subpopulations should be carefully evaluated: Despite what may appear to be large nominal sample sizes for the total survey, the effective sample sizes for pooled analysis may be much smaller than the simple case count suggests.

Fortunately, a detailed understanding of each of these contributions to the final design effect for a complex sample survey is more important to the survey designer than it is to the survey analyst. In the role of survey analyst, the reader will often not have access to the detailed information that the survey producer's statisticians used to develop the final weight. As analysts, we must certainly be able to identify the correct weight variable to use in our analysis, be able to perform checks that the designated weight has been correctly carried forward to the data set that we will use for analysis, and be familiar with the syntax required to perform weighted estimation in our chosen statistical analysis software. Subsequent chapters in this book will provide the explanation and examples needed to become proficient in applying weights in survey data analysis.

2.8 Multistage Area Probability Sample Designs

The applied statistical methods covered in Chapters 5–12 of this volume are intended to cover a wide range of complex sample designs. However, because the majority of large public-use survey data sets that are routinely used in the social and health sciences are based on multistage area probability sampling of households, it is important to describe this particular probability sampling technique in greater detail. The generic illustration presented here is most applicable to national household sampling in the United States and Canada. With minor changes in the number of stages and the choice of sampling units, similar household sample designs are used throughout Central and South America, Africa, Asia, and Europe (Heeringa and O'Muircheartaigh, 2010).

Figure 2.3 illustrates the selection of a multistage, area probability sample of respondents. The design illustrated in this figure conforms closely to the actual procedure used in the selection of the NCS-R, HRS, and NHANES samples of U.S. household populations. Under the general procedure illustrated by the figure, the selection of each study respondent requires a four-

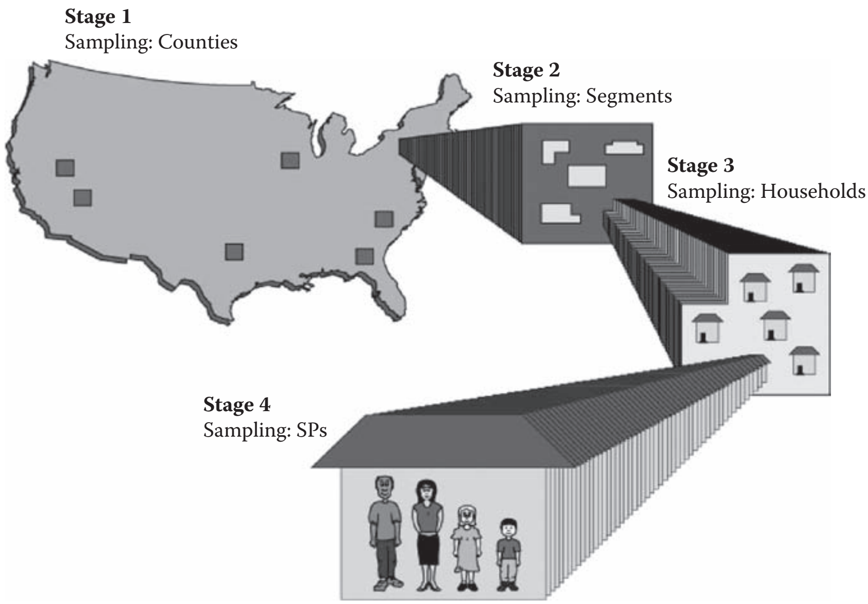

FIGURE 2.3
Schematic illustration of multistage area probability sampling. (From: Mohadjer, L., *The National Assessment of Adult Literacy (NAAL) Sample Design,* Westat, Rockville, MD, 2002. With permission.)

step sampling process: a primary stage sampling of U.S. counties or groups of adjacent counties, followed by a second stage sampling of area segments and a third stage sampling of housing units within the selected area segments, and concluding with the random selection of eligible respondents from the sampled housing units.

2.8.1 Primary Stage Sampling

The **primary stage units** of the multistage sample are single counties or groupings of geographically contiguous counties. The PSUs are therefore the highest-level groupings or "clusters" of sample observations (this will be important to remember when we discuss variance estimation in Chapters 3 and 4). All land area in the target population is divided into PSUs (e.g., the land area of the 50 United States is uniquely divided into roughly 3,100 county and parish [Louisiana] divisions). For comparison, the land area of Chile is uniquely divided into *communa* units, Russia into *raions* (regions), and South Africa and many other countries into Census *enumeration areas*. Ideally, the populations within PSUs will be reasonably heterogeneous to minimize intraclass correlation for survey variables. At the same time, the geographically defined PSUs must be small enough in size to facilitate cost-efficient travel to second stage interview locations. Each designated PSU in the population is assigned to $h = 1, \ldots, H$ sampling strata based on region of the country, urban/rural status, PSU size, geographic location within regions, and population characteristics. Designs with approximately $H = 100$ primary stage strata are common in multistage samples of U.S. households, but the actual number of strata employed in the primary stage sample can range from less than ten to hundreds depending on available stratification variables and the number of PSUs, a_h, that will be selected from each stratum.

Depending on the study sample design, from 12 to 20 of the primary stage strata contain only a single **self-representing (SR)** metropolitan PSU. Each SR PSU is included in the sample with certainty in the primary stage of selection—for these strata, "true sampling" begins at the second stage of selection. The remaining **nonself-representing (NSR)** primary stage strata in each design contain more than one sample PSU. From each of these NSR strata, one or more PSUs is sampled with **probability proportionate to its size (PPS)** as measured in occupied housing unit counts reported at the most recent census. The number of PSUs selected from each primary stage stratum is decided by the sample allocation. The University of Michigan Survey Research Center (SRC) National Sample Design (Heeringa and Connor, 1995) allocates one primary stage selection per stratum. More commonly, a minimum of two sample PSUs are selected from each primary stage stratum. A "two-per-stratum" design in which exactly two PSUs ($a_{\mathrm{h}} = 2$) are selected with PPS from each stratum maximizes the number of strata that can be formed for selecting a primary stage sample of

$$a = \sum_{h=1}^{H} a_h \text{ PSUs.}$$

The two-per-stratum allocation is also common because a minimum of two PSUs per primary stage stratum is required to estimate sampling variances of estimates from complex samples. One-per-stratum primary stage samples, like that used by the University of Michigan's SRC for HRS and NCS-R, maximize the stratification potential but require a collapsing of design strata to create pseudo-strata for complex sample variance estimation (see Chapter 4).

2.8.2 Secondary Stage Sampling

The designated second-stage sampling units (SSUs) in multistage samples are commonly termed **area segments**. Area segments are formed by linking geographically contiguous Census blocks to form units with a minimum number of occupied housing units (typically 50–100 based on the needs of the study). Within sample PSUs, SSUs may be stratified at the county level by geographic location and race/ethnicity composition of residents' households. Within each sample PSU, the actual probability sampling of SSUs is performed with probabilities proportionate to census counts of the occupied housing units for the census blocks that comprise the area segment. The number of SSUs that are selected within each sample PSU is determined by survey statisticians to optimize the cost and error properties of the multistage design. The number of SSUs selected in an SR PSU varies in proportion to the total size of the PSU selected with certainty (e.g., HRS and NCS-R use approximately 48 area segments in the New York self-representing PSU). Depending on the total size of the survey and the cost structure for data collection operations, typically from 6 to 24 SSUs are selected from each NSR PSU.

2.8.3 Third and Fourth Stage Sampling of Housing Units and Eligible Respondents

Prior to the selection of the third stage sample of households, field staff from the survey organization visit each sample SSU location and conduct an up-to-date enumeration or "listing" of all housing units located within the physical boundaries of the selected area segments. A third-stage sample of housing units is then selected from the enumerative listing according to a predetermined sampling rate. The third-stage sampling rates for selecting households in the multistage area probability samples are computed using the following "selection equation" (Kish, 1965):

$$\begin{aligned} f &= f_1 \times f_2 \times f_3 \\ &= \frac{MOS_\alpha \times a_h}{MOS_h} \times \frac{b_\alpha \times MOS_{ssu}}{MOS_\alpha} \times \frac{C_h}{MOS_{ssu}} \end{aligned} \tag{2.22}$$

In the final selection equation derived in (2.22), we have the following notation: f = the overall multistage sampling rate for housing units; MOS_α = total population measure of size in the selected PSU α; MOS_h = total population measure of size in the design stratum h; a_h = number of PSUs to be selected from design stratum h; b_α = number of area segments selected in the PSU α; MOS_{ssu} = total household measure of size for the SSU; C_h = a stratum-specific constant = $(f \times MOS_{stratum})/a_h b_\alpha$.

For example, the third-stage sampling rate for selecting an equal probability national sample from the listed housing units for the selected PSUs and area segment SSUs is

$$f_3 = \frac{f}{f_1 \times f_2} = \frac{f \times MOS_h}{a_h \times b_\alpha \times MOS_{ssu}} \tag{2.23}$$

The third-stage sampling rate is computed for each selected SSU in the sample design. This rate is then used to select a random sample of actual housing units from the area segment listing.

Each sample housing unit is then contacted in person by an interviewer. Within each cooperating sample household, the interviewer conducts a short screening interview with a knowledgeable adult to determine if household members meet the study eligibility criteria. If the informant reports that one or more eligible adults live at the sample housing unit address, the interviewer prepares a complete listing of household members and proceeds to randomly select a respondent for the study interview. The random selection of the respondent is often performed using a special adaptation of the objective household roster/selection table method developed by Kish (1949).

Despite the obvious effort and complexity that goes into fielding a multistage area probability sample, relatively simple specifications of primary stage stratum, primary stage cluster (PSU), and final analysis weight variables will be required for analysts desiring appropriate analyses of survey data from this common household sampling design. A detailed discussion of a unified approach to variance estimation for multistage samples is given in Chapter 4.

2.9 Special Types of Sampling Plans Encountered in Surveys

As previously described, most large-scale social, economic, demographic, and health-related surveys are designed to provide the capability to make descriptive inferences to specific survey populations or to analyze multivariate relationships in a population. Although the techniques for applied survey data analysis presented in the following chapters are generally applicable to all forms of probability sample designs suitable for population estimation and inference, some fields of population survey research (e.g., surveys of businesses, hospitals and other nonhousehold units that vary in size and "importance") have developed special methods that will not be covered in detail in this volume. Researchers who are working with survey data for these populations are encouraged to use the survey literature to determine current best practices for these special population surveys.

Survey research on natural populations in environmental (e.g., forestry or fisheries), geological, and some human and animal epidemiological studies is increasingly turning to **adaptive sample designs** to optimize observation, estimation, and inference. If the reader's data are of these types, they will need to use special procedures for estimation and inference. An excellent reference on adaptive sample design can be found in Thompson and Seber (1996).

Increasingly, adaptive sampling procedures are being employed in major population surveys. Groves and Heeringa (2006) apply the term **responsive design** to surveys that adapt sampling, survey measurement, and nonresponse follow-ups to empirical information that is gathered in the survey process. One sampling technique that is critical to the responsive design of surveys is the use of multiphase sampling, in which sample cases may be subsampled for further contact and interview at a time point, *t*, conditional on the prior disposition (e.g., number of calls, success with contact, resistance to interview) of the case. Presently, the stochastic nature of the sample disposition of each case at time *t* is ignored, and the data are weighted for estimation as though the disposition of cases was a deterministic (fixed) outcome. Current and future research is expected to lead to improved procedures for estimation and inference in multiphase sample designs.

Occasionally, population-based survey methods are employed to perform research that is purely analytical. These include studies that fall in the category of epidemiological case-control designs; randomized population-based experiments including "group randomized trials"; or model-based designs for research on hierarchical or multilevel populations (e.g., research on student, classroom, and school effects). Analysis of "survey" data from these types of analytical research designs requires special approaches. Chapter 12 will explore approaches to several of the more common analytical designs that use survey-like procedures to collect data, but, again, with data of this type, the reader is encouraged to also turn to the statistical literature for an up-to-date and more in-depth description of best practices. See, for example, Burns et al. (1996), Heeringa et al. (2001), and Raudenbush (2000).