# 1

# *Applied Survey Data Analysis: Overview*

## 1.1 Introduction

Modern society has adopted the **survey method** as a principal tool for looking at itself—"a telescope on society" in the words of House et al. (2004). The most common application takes the form of the periodic media surveys that measure population attitudes and beliefs on current social and political issues:

> Recent international reports have said with near certainty that human activities are the main cause of global warming since 1950. The poll found that 84 percent of Americans see human activity as at least contributing to warming. (*New York Times,* April 27, 2007).

One step removed from the media limelight is the use of the survey method in the realms of marketing and consumer research to measure the preferences, needs, expectations, and experiences of consumers and to translate these to indices and other statistics that may influence financial markets or determine quality, reliability, or volume ratings for products as diverse as automobiles, hotel services, or TV programming:

> CBS won the overall title with an 8.8 rating/14 share in primetime, ABC finished second at 7.7/12…. (http://www.zap2it.com, January 11, 2008)

> The Index of Consumer Sentiment (see Figure 1.1) fell to 88.4 in the March 2007 survey from 91.3 in February and 96.9 in January, but it was nearly identical with the 88.9 recorded last March. (Reuters, University of Michigan, April 2007)

Also outside the view of most of society is the use of large-scale scientific surveys to measure labor force participation, earnings and expenditures, health and health care, commodity stocks and flows, and many other topics. These larger and longer-term programs of survey research are critically important to social scientists, health professionals, policy makers, and administrators and thus indirectly to society itself.

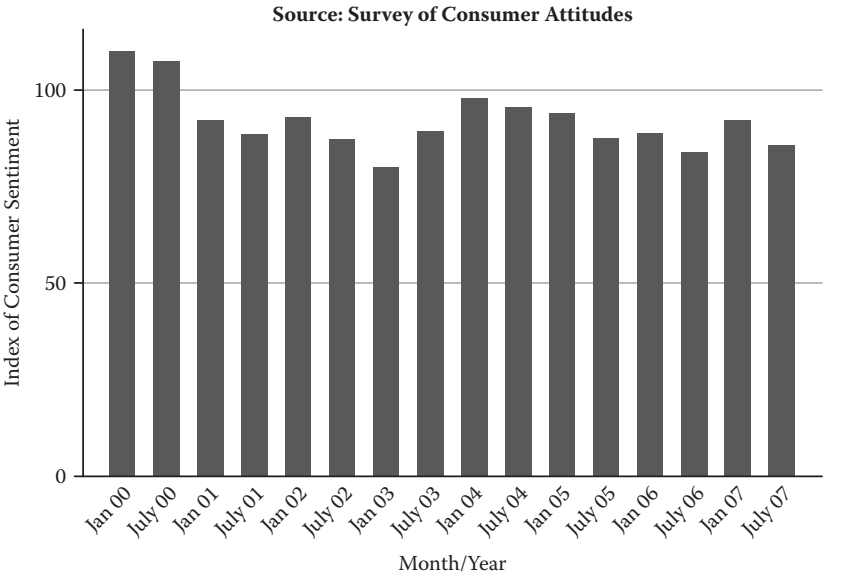


**FIGURE 1.1**
Index of Consumer Sentiment, January 2000–July 2007.

> Real median household income in the United States rose between 2005 and 2006, for the second consecutive year. Household income increased 0.7 percent, from $47,845 to $48,201. (DeNavas-Walt, Proctor, and Smith, 2007)

> In a series of logistic models that included age and one additional variable (i.e., education, gender, race, or APOE genotype), older age was consistently associated with increased risk of dementia ($p < 0.0001$). In these trivariate models, more years of education was associated with lower risk of dementia ($p < 0.0001$). There was no significant difference in dementia risk between males and females ($p = 0.26$). African Americans were at greater risk for dementia ($p = 0.008$). As expected, the presence of one (Odds Ratio = 2.1; 95% C.I. = 1.45 – 3.07) or two (O.R. = 7.1; 95% C.I. = 2.92 – 17.07) APOE e4 alleles was significantly associated with increased risk of dementia. (Plassman et al., 2007)

The focus of this book will be on analysis of **complex sample survey data** typically seen in large-scale scientific surveys, but the general approach to survey data analysis and specific statistical methods described here should apply to all forms of survey data.

To set the historical context for contemporary methodology, Section 1.2 briefly reviews the history of developments in theory and methods for applied survey data analysis. Section 1.3 provides some needed background on the data sets that will be used for the analysis examples in Chapters 2–12. This short overview chapter concludes in Section 1.4 with

a general review of the sequence of steps required in any applied analysis of survey data.

## 1.2 A Brief History of Applied Survey Data Analysis

Today's survey data analysts approach a problem armed with substantial background in statistical survey theory, a literature filled with empirical results and high-quality software tools for the task at hand. However, before turning to the best methods currently available for the analysis of survey data, it is useful to look back at how we arrived at where we are today. The brief history described here is certainly a selected interpretation, chosen to emphasize the evolution of probability sampling design and related statistical analysis techniques that are most directly relevant to the material in this book. Readers interested in a comprehensive review of the history and development of survey research in the United States should see Converse (1987). Bulmer (2001) provides a more international perspective on the history of survey research in the social sciences. For the more statistically inclined, Skinner, Holt, and Smith (1989) provide an excellent review of the development of methods for descriptive and analytical treatment of survey data. A comprehensive history of the impacts of sampling theory on survey practice can be found in O'Muircheartaigh and Wong (1981).

### 1.2.1 Key Theoretical Developments

The science of survey sampling, survey data collection methodology, and the analysis of survey data date back a little more than 100 years. By the end of the 19th century, an open and international debate established the **representative sampling method** as a statistically acceptable basis for the collection of observational data on populations (Kaier, 1895). Over the next 30 years, work by Bowley (1906), Fisher (1925), and other statisticians developed the role of randomization in sample selection and large-sample methods for estimation and statistical inference for simple random sample (SRS) designs.

The early work on the representative method and inference for simple random and stratified random samples culminated in a landmark paper by Jerzy Neyman (1934), which outlined a cohesive framework for estimation and inference based on estimated confidence intervals for population quantities that would be derived from the probability distribution for selected samples over repeated sampling. Following the publication of Neyman's paper, there was a major proliferation of new work on survey sample designs, estimation of population statistics, and variance estimation required to develop confidence intervals for sample-based inference, or what in more recent times has been labeled **design-based inference** (Cochran, 1977; Deming, 1950;

Hansen, Hurwitz, and Madow, 1953; Kish, 1965; Sukatme, 1954; Yates, 1949). House et al. (2004) credit J. Steven Stock (U.S. Department of Agriculture) and Lester Frankel (U.S. Bureau of the Census) with the first applications of area probability sampling methods for household survey data collections. Even today, the primary techniques for sample design, population estimation, and inference developed by these pioneers and published during the period 1945–1975 remain the basis for almost all descriptive analysis of survey data.

The developments of the World War II years firmly established the probability sample survey as a tool for describing population characteristics, beliefs, and attitudes. Based on Neyman's (1934) theory of inference, survey sampling pioneers in the United States, Britain, and India developed optimal methods for sample design, estimators of survey population characteristics, and confidence intervals for population statistics. As early as the late 1940s, social scientists led by sociologist Paul Lazarsfeld of Columbia University began to move beyond using survey data to simply describe populations to using these data to explore relationships among the measured variables (see Kendall and Lazarsfeld, 1950; Klein and Morgan, 1951). Skinner et al. (1989) and others before them labeled these two distinct uses of survey data as *descriptive* and *analytical*. Hyman (1955) used the term *explanatory* to describe scientific surveys whose primary purpose was the analytical investigation of relationships among variables.

During the period 1950–1990, analytical treatments of survey data expanded as new developments in statistical theory and methods were introduced, empirically tested, and refined. Important classes of methods that were introduced during this period included log-linear models and related methods for contingency tables, generalized linear models (e.g., logistic regression), survival analysis models, general linear mixed models (e.g., hierarchical linear models), structural equation models, and latent variable models. Many of these new statistical techniques applied the method of maximum likelihood to estimate model parameters and standard errors of the estimates, assuming that the survey observations were *independent* observations from a known probability distribution (e.g., binomial, multinomial, Poisson, product multinomial, normal). As discussed in Chapter 2, data collected under most contemporary survey designs do not conform to the key assumptions of these methods.

As Skinner et al. (1989) point out, survey statisticians were aware that straightforward applications of these new methods to complex sample survey data could result in underestimates of variances and therefore could result in biased estimates of confidence intervals and test statistics. However, except in limited situations of relatively simple designs, exact determination of the size and nature of the bias (or a potential correction) were difficult to express analytically. Early investigations of such "design effects" were primarily empirical studies, comparing design-adjusted variances for estimates with the variances that would be obtained if the

data were truly identically and independently distributed (equivalent to a simple random sample of equal size). Over time, survey statisticians developed special approaches to estimating these models that enabled the survey analyst to take into account the complex characteristics of the survey sample design (e.g., Binder, 1983; Kish and Frankel, 1974; Koch and Lemeshow, 1972; Pfeffermann et al., 1998; Rao and Scott, 1981). These approaches (and related developments) are described in Chapters 5–12 of this text.

### 1.2.2 Key Software Developments

Development of the underlying statistical theory and empirical testing of new methods were obviously important, but the survey data analyst needed computational tools to apply these techniques. We can have nothing but respect for the pioneers who in the 1950s fitted multivariate regression models to survey data using only hand computations (e.g., sums, sums of squares, sums of cross-products, matrix inversions) performed on a rotary calculator and possibly a tabulating machine (Klein and Morgan, 1951). The origin of statistical software as we know it today dates back to the 1960s, with the advent of the first mainframe computer systems. Software systems such as BMDP and OSIRIS and later SPSS, SAS, GLIM, S, and GAUSS were developed for mainframe users; however, with limited exceptions, these major software packages did not include programs that were adapted to complex sample survey data.

To fill this void during the 1970s and early 1980s, a number of stand-alone programs, often written in the Fortran language and distributed as compiled objects, were developed by survey statisticians (e.g., OSIRIS: PSALMS and REPERR, CLUSTERS, CARP, SUDAAN, WesVar). By today's standards, these programs had a steep "learning curve," limited data management flexibility, and typically supported only descriptive analysis (means, proportions, totals, ratios, and functions of descriptive statistics) and linear regression modeling of multivariate relationships. A review of the social science literature of this period shows that only a minority of researchers actually employed these special software programs when analyzing complex sample survey data, resorting instead to standard analysis programs with their default assumption that the data originated with a simple random sample of the survey population.

The appearance of microcomputers in the mid-1980s was quickly followed by a transition to personal computer versions of the major statistical software (BMDP, SAS, SPSS) as well as the advent of new statistical analysis software packages (e.g., SYSTAT, Stata, S-Plus). However, with the exception of specialized software systems (WesVar PC, PC CARP, PC SUDAAN, MicroOSIRIS, CLUSTERS for PC, IVEware) that were often designed to read data sets stored in the formats of the larger commercial software packages, the microcomputing revolution still did not put tools for the analysis of complex

sample survey data in the hands of most survey data analysts. Nevertheless, throughout the late 1980s and early 1990s, the scientific and commercial pressures to incorporate programs of this type into the major software systems were building. Beginning with Version 6.12, SAS users had access to PROC SURVEYMEANS and PROC SURVEYREG, two new SAS procedures that permitted simple descriptive analysis and linear regression analysis for complex sample survey data. At about the same time, the Stata system for statistical analysis appeared on the scene, providing complex sample survey data analysts with the "`svy`" versions of the more important analysis programs. SPSS's entry into the world of complex sample survey data analysis came later with the introduction of the Complex Samples add-on module in Version 13. Appendix A of this text covers the capabilities of these different systems in detail.

The survey researcher who sits down today at his or her personal computing work station has access to powerful software systems, high-speed processing, and high-density data storage capabilities that the analysts in the 1970s, 1980s, and even the 1990s could not have visualized. All of these recent advances have brought us to a point at which today's survey analyst can approach both simple and complex problems with the confidence gained through a fundamental understanding of the theory, empirically tested methods for design-based estimation and inference, and software tools that are sophisticated, accurate, and easy to use.

Now that we have had a glimpse at our history, let's begin our study of applied survey data analysis.

## 1.3 Example Data Sets and Exercises

Examples based on the analysis of major survey data sets are routinely used in this book to demonstrate statistical methods and software applications. To ensure diversity in sample design and substantive content, example exercises and illustrations are drawn from three major U.S. survey data sets.

### 1.3.1 The National Comorbidity Survey Replication (NCS-R)

The NCS-R is a 2002 study of mental illness in the U.S. household population ages 18 and over. The core content of the NCS-R is based on a lay-administered interview using the World Health Organization (WHO) CIDI (Composite International Diagnostic Interview) diagnostic tool, which is designed to measure primary mental health diagnostic symptoms, symptom severity, and use of mental health services (Kessler et al., 2004). The NCS-R was based on interviews with randomly chosen adults in an equal probability, multistage sample of households selected from the University of Michigan

National Sample master frame. The survey response rate was 70.9%. The survey was administered in two parts: a Part I core diagnostic assessment of all respondents ($n$ = 9,282), followed by a Part II in-depth interview with 5,692 of the 9,282 Part I respondents, including all Part I respondents who reported a lifetime mental health disorder and a probability subsample of the disorder-free respondents in the Part I screening.

The NCS-R was chosen as an example data set for the following reasons: (1) the scientific content and, in particular, its binary measures of mental health status; (2) the multistage design with primary stage stratification and clustering typical of many large-scale public-use survey data sets; and (3) the two-phase aspect of the data collection.

### 1.3.2 The Health and Retirement Study (HRS)—2006

The Health and Retirement Study (HRS) is a longitudinal study of the American population 50 years of age and older. Beginning in 1992, the HRS has collected data every two years on a longitudinal panel of sample respondents born between the years of 1931 and 1941. Originally, the HRS was designed to follow this probability sample of age-eligible individuals and their spouses or partners as they transitioned from active working status to retirement, measuring aging-related changes in labor force participation, financial status, physical and mental health, and retirement planning. The HRS observation units are age-eligible individuals and "financial units" (couples in which at least one spouse or partner is HRS eligible). Beginning in 1993 and again in 1998 and 2004, the original HRS 1931–1941 birth cohort panel sample was augmented with probability samples of U.S. adults and spouses/partners from (1) pre-1924 (added in 1993); (2) 1924–1930 and 1942–1947 (added in 1998); and (3) 1948–1953 (added in 2004). In 2006, the HRS interviewed over 22,000 eligible sample adults in the composite panel.

The HRS samples were primarily identified through in-person screening of large, multistage area probability samples of U.S. households. For the pre-1931 birth cohorts, the core area probability sample screening was supplemented through sampling of age-eligible individuals from the U.S. Medicare Enrollment Database. Sample inclusion probabilities for HRS respondents vary slightly across birth cohorts and are approximately two times higher for African Americans and Hispanics. Data from the 2006 wave of the HRS panel are used for most of the examples in this text, and we consider a longitudinal analysis of multiple waves of HRS data in Chapter 12.

### 1.3.3 The National Health and Nutrition Examination Survey (NHANES)—2005, 2006

Sponsored by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC), the NHANES is a survey of the adult, noninstitutionalized population of the United States. The NHANES

is designed to study the prevalence of major disease in the U.S. population and to monitor the change in prevalence over time as well as trends in treatment and major disease risk factors including personal behaviors, environmental exposure, diet, and nutrition. The NHANES survey includes both an in-home medical history interview with sample respondents and a detailed medical examination at a local mobile examination center (MEC). The NHANES surveys were conducted on a periodic basis between 1971 and 1994 (NHANES I, II, III), but beginning in 1999, the study transitioned to a continuous interviewing design. Since 1999, yearly NHANES data collections have been performed in a multistage sample that includes 15 primary stage unit (PSU) locations with new sample PSUs added in each data collection year. Approximately 7,000 probability sample respondents complete the NHANES in-home interview phase each year and roughly 5,000 of these individuals also consent to the detailed MEC examination. To meet specific analysis objectives, the NHANES oversamples low-income persons, adolescents between the ages of 12 and 19, persons age 60 and older, African Americans, and Hispanics of Mexican ancestry. To ensure adequate precision for sample estimates, NCHS recommends pooling data for two or more consecutive years of NHANES data collection. The NHANES example analyses provided in this text are based on the combined data collected in 2005 and 2006. The unweighted response rate for the interview phase of the 2005–2006 NHANES was approximately 81%.

Public use versions of each of these three major survey data sets are available online. The companion Web site for this book provides the most current links to the official public use data archives for each of these example survey data sets.

### 1.3.4 Steps in Applied Survey Data Analysis

Applied survey data analysis—both in daily practice and here in this book—is a process that requires more of the analyst than simple familiarity and proficiency with statistical software tools. It requires a deeper understanding of the sample design, the survey data, and the interpretation of the results of the statistical methods. Following a more general outline for applied statistical analysis presented by Cox (2007), Figure 1.2 outlines a sequence of six steps that are fundamental to applied survey data analysis, and we describe these steps in more detail in the following sections.

#### *1.3.4.1 Step 1: Definition of the Problem and Statement of the Objectives*

The first of the six steps involves a clear specification of the problem to be addressed and formulation of objectives for the analysis exercise. For example, the "problem" may be ambiguity among physicians over whether there should be a lower threshold for prostate biopsy following prostate specific antigen (PSA) screening in African American men (Cooney et al., 2001). The

| Step | Activity |
|---|---|
| 1 | Definition of the problem and statement of the objectives. |
| 2 | Understanding the sample design. |
| 3 | Understanding design variables, underlying constructs, and missing data. |
| 4 | Analyzing the data. |
| 5 | Interpreting and evaluating the results of the analysis. |
| 6 | Reporting of estimates and inferences from the survey data. |

**FIGURE 1.2**
Steps in applied survey data analysis.

corresponding objective would be to estimate the 95th percentile and the 95% confidence bounds for this quantity (+/– .2 ng/ml PSA) in a population of African American men. The estimated 95% confidence bounds can in turn be used by medical experts to determine if the biopsy threshold for African American men should be different than for men of other race and ethnic groups.

As previously described, the problems to which survey data analyses may be applied span many disciplines and real-world settings. Likewise, the statistical objectives may vary. Historically, the objectives of most survey data analyses were to describe characteristics of a target population: its average household income, the median blood pressure of men, or the proportion of eligible voters who favor candidate X. But survey data analyses can also be used for decision making. For example, should a pharmaceutical company recall its current products from store shelves due to a perceived threat of contamination? In a population case-control study, does the presence of silicone breast implants significantly increase the odds that a woman will contract a connective tissue disease such as scleroderma (Burns et al., 1996)? In recent decades, the objective of many sample survey data analyses has been to explore and extend the understanding of multivariate relationships among variables in the target population. Sometimes multivariate modeling of survey data is seen simply as a descriptive tool, defining the form of a functional relationship as it exists in a finite population. But it is increasingly common for researchers to use observational data from complex sample surveys to probe causality in the relationships among variables.

#### *1.3.4.2 Step 2: Understanding the Sample Design*

The survey data analyst must understand the sample design that was used to collect the data he or she is about to analyze. Without an understanding of key properties of the survey sample design, the analysis may be inefficient,

biased, or otherwise lead to incorrect inference. An experienced researcher who designs and conducts a randomized block experimental design to test the relative effectiveness of new instructional methods should not proceed to analyze the data as a simple factorial design, ignoring the blocking that was built into his or her experiment. Likewise, an economics graduate student who elects to work with the longitudinal HRS data should understand that the nationally representative sample of older adults includes stratification, clustering, and disproportionate sampling (i.e., compensatory population weighting) and that these design features may require special approaches to population estimation and inference.

At this point, we may have discouraged the reader into thinking that an in-depth knowledge of survey sample design is required to work with survey data or that he or she may need to relearn what was studied in general courses on applied statistical methods. This is not the case. Chapters 2 through 4 will introduce the reader to the fundamental features of **complex sample designs** and will demonstrate how design characteristics such as stratification, clustering, and weighting are easily incorporated into the statistical methods and software for survey estimation and inference. Chapters 5–12 will show the reader that relatively simple extensions of his or her current knowledge of applied statistical analysis methods provide the necessary foundation for efficient and accurate analysis of data collected in sample surveys.

#### *1.3.4.3 Step 3: Understanding Design Variables, Underlying Constructs, and Missing Data*

The typical scientific survey data set is **multipurpose**, with the final data sets often including hundreds of variables that span many domains of study—income, education, health, family. The sheer volume of available data and the ease by which it can be accessed can cause survey data analysts to become complacent in their attempts to fully understand the properties of the data that are important to their choice of statistical methods and the conclusions that they will ultimately draw from their analysis. Step 2 described the importance of understanding the sample design. In the survey data, the key features of the sample design will be encoded in a series of **design variables**. Before analysis begins, some simple questions need to be put to the candidate data set: What are the empirical distributions of these design variables, and do they conform to the design characteristics outlined in the technical reports and online study documentation? Does the original survey question that generated a variable of interest truly capture the underlying construct of interest? Are the response scales and empirical distributions of responses and independent variables suitable for the intended analysis? What is the distribution of missing data across the cases and variables, and is there a potential impact on the analysis and the conclusions that will be drawn?

Chapter 4 discusses techniques for answering these and other questions before proceeding to statistical analysis of the survey data.

#### *1.3.4.4 Step 4: Analyzing the Data*

Finally we arrive at the step to which many researchers rush to enter the process. We are all guilty of wanting to jump ahead. Identifying the problem and objectives seems intuitive. We tell ourselves that formalizing that step wastes time. Understanding the design and performing data management and exploratory analysis to better understand the data structure is boring. After all, the statistical analysis step is where we obtain the results that enable us to describe populations (through confidence intervals), to extend our understanding of relationships (through statistical modeling), and possibly even to test scientific hypotheses.

In fact, the statistical analysis step lies at the heart of the process. Analytic techniques must be carefully chosen to conform to the analysis objectives and the properties of the survey data. Specific methodology and software choices must accommodate the design features that influence estimation and inference. Treatment of statistical methods for survey data analysis begins in Chapters 5 and 6 with coverage of univariate (i.e., single-variable) descriptive and simple bivariate (i.e., two-variable) analyses of continuous and categorical variables. Chapter 7 presents the linear regression model for continuous dependent variables, and generalized linear regression modeling methods for survey data are treated in Chapters 8 and 9. Chapter 10 pertains to methods for event-history analysis of survey data, including models such as the Cox proportional hazard model and discrete time logistic models. Chapter 11 introduces methods for handling missing data problems in survey data sets. Finally, the coverage of statistical methods for survey data analysis concludes with a discussion of new developments in the area of survey applications of advanced statistical techniques, such as multilevel analysis, in Chapter 12.

#### *1.3.4.5 Step 5: Interpreting and Evaluating the Results of the Analysis*

Knowledge of statistical methods and software tools is fundamental to success as an applied survey data analyst. However, setting up the data, running the programs, and printing the results are not sufficient to constitute a thorough treatment of the analysis problem. Likewise, scanning a column of $p$-values in a table of regression model output does not inform us concerning the form of the "final model" or even the pure effect of a single predictor. As described in Step 3, interpretation of the results from an analysis of survey data requires a consideration of the error properties of the data. Variability of sample estimates will be reflected in the **sampling errors** (i.e., confidence intervals, test statistics) estimated in the course of the statistical analysis. **Nonsampling errors**, including potential bias due to survey nonresponse

and item missing data, cannot be estimated from the survey data (Lessler and Kalsbeek, 1992). However, it may be possible to use ancillary data to explore the potential direction and magnitude of such errors. For example, an analyst working for a survey organization may statistically compare survey respondents with nonrespondents in terms of known correlates of key survey variables that are readily available on the sampling frame to assess the possibility of nonresponse bias.

As survey data analysts have pushed further into the realm of multivariate modeling of survey data, care is required in interpreting fitted models. Is the model reasonably identified, and do the data meet the underlying assumptions of the model estimation technique? Are there alternative models that explain the observed data equally well? Is there scientific support for the relationship implied in the modeling results? Are interpretations that imply causality in the modeled relationships supported (Rothman, 1988)?

#### *1.3.4.6 Step 6: Reporting of Estimates and Inferences from the Survey Data*

The end products of applied survey data analyses are reports, papers, or presentations designed to communicate the findings to fellow scientists, policy analysts and administrators and decision makers. This text includes discussion of standards and proven methods for effectively presenting the results of applied survey data analyses, including table formatting, statistical contents, and the use of statistical graphics.

With these six steps in mind, we now can now begin our walk through the process of planning, formulating, and conducting analysis of survey data.