The Analysis of Household Surveys

A Microeconometric Approach to Development Policy

Angus Deaton

Published for the World Bank
The Johns Hopkins University Press
Baltimore and London

The Johns Hopkins University Press
Baltimore, Maryland 21211-2190, U.S.A.

Manufactured in the United States of America
First printing July 1997

Photographs on the back cover: *top and bottom*, household interviews during the Kagera Health and Development Survey, 1991–94; *middle*, woman being weighed as part of the Côte d'Ivoire Living Standards Survey, 1986 (top photo by T. Paul Schultz; middle and bottom photos by Martha Ainsworth).

Library of Congress Cataloging-in-Publication Data

Deaton, Angus.
The analysis of household surveys : a microeconometric approach to development policy / Angus Deaton.
p. cm.
Includes bibliographical references and index.
ISBN 0-8018-5254-4
1. Household surveys—Developing countries—Methodology.
2. Developing countries—Economic conditions—Econometric models.
I. Title.
HB849.49.D43 1997
339.4' 07'23—dc21 97-2905
CIP

Contents

Introduction

The collection of household survey data in developing countries is hardly a new phenomenon. The National Sample Survey Organization in India has been collecting such data on a regular basis since the 1940s, and there are many other countries with long-running and well-established surveys. Until recently, however, the handling and processing of large microeconomic data sets was both cumbersome and expensive, so that survey data were not widely used beyond the production of the original survey reports. In the last ten or fifteen years, the availability of cheap and convenient microcomputers has changed both the collection and analysis of household survey data. Calculations that could be done only on multimillion-dollar mainframes in 1980—and then with some difficulty—are now routinely carried out on cheap laptop computers. These same machines can be carried into the field and used to record and edit data as they are provided by the respondents. As a result, survey data are becoming available in a more timely fashion, months rather than years after the end of the survey; freshly collected data are much more useful for policy exercises than are those that are many years old. At the same time, analysts have become more interested in exploring ways in which survey data can be used to inform and to improve the policy process. Such explorations run from the tabulations and graphical presentation of levels of living to more basic research on household behavior.

Purpose and intended audience

This book is about the analysis of household survey data from developing countries and about how such data can be used to cast light on a range of policy issues. Much of the analysis works with household budget data, collected from income and expenditure surveys, though I shall occasionally address topics that require wider information. I shall use data from several different economies to illustrate the analysis, drawing examples of policy issues from economies as diverse as Côte d'Ivoire, India, Pakistan, South Africa, Taiwan (China), and Thailand. I shall be concerned with methodology as well as substance, and one of the aims of the book is to bring together the relevant statistical and econometric methods that are useful for building the bridge between data and policy. The book is not intended as a manual for the analysis of survey data—it is hardly possible to reduce policy research to a formula—but it does provide a number of illustrations of what can be

done, with fairly detailed explanations of how to do it. Nor can a "how-to" book provide a comprehensive review of all the development topics that have been addressed with household survey data; that purpose has already been largely met by the microeconomic survey papers in the three volumes of the *Handbook of Development Economics*. Instead, I have focused on topics on which I have worked myself, in the hope that the lack of coverage will be compensated for by the detailed knowledge that can only come from having carried out the empirical research. The restriction to my own work also enables me to provide the relevant computer code for almost all of the empirical results and graphics, something that could hardly be combined with the broad coverage of a genuine survey. The Appendix gives code and programs using STATA; in my experience, this is the most convenient package for working with data from household surveys. The programs are not a package; users will have to substitute their own data sets and will need sufficient basic knowledge of STATA to adapt the code. Nevertheless, the programs provide a template for generating results similar to those presented and discussed here. I have also tried to keep the programs simple, sometimes at the expense of efficiency or elegance, so that it should not be too difficult to translate the logic into other packages.

I hope that the material will be of interest to development practitioners, in the World Bank and elsewhere, as well as to a more academic audience of students of economic development. The material in the first two chapters is also designed to help readers interpret applied econometric work based on survey data. But the audience that I most want to reach is that of researchers in developing countries. Statistical offices, research institutes, and universities in developing countries are now much less constrained by computation than they were only a few years ago, and the calculations described here can be done on personal computers using readily available and relatively inexpensive software. I have also tried to keep the technical presentation at a relatively modest level. I take for granted most of what would be familiar from a basic course in econometrics, but I devote a good deal of space to expositions of useful techniques—such as nonparametric density and regression estimation, or the bootstrap—that are neither widely taught in elementary econometrics courses nor described in standard texts. Nevertheless, there are points where there is an inevitable conflict between simplicity, on the one hand, and clarity and precision, on the other. When necessary I have "starred" those sections or subsections in which the content is either necessarily technical or is of interest only to those who wish to try to replicate the analysis. Occasional "technical notes," usually starred, are shorter digressions that can readily be skipped at a first reading.

Policy and data: methodological issues

Household surveys provide a rich source of data on economic behavior and its links to policy. They provide information at the level of the individual household about many variables that are either set or influenced by policy, such as prices, transfers, or the provision of schools and clinics. They also collect data on outcomes that we care about and that are affected by the policy variables, such as levels of nutrition, expenditure patterns, educational attainments, earnings, and health. Many impor-

tant research questions concern the link between the instruments of policy and the outcome variables: the rate of return to government-provided schooling, the effectiveness of various types of clinics, the equity and efficiency effects of transfers and taxes, and the nutritional benefits of food subsidies. Because household surveys document these links, they are the obvious data bases for this sort of policy research, for evaluating the welfare benefits of public programs. Of course, associations in the data establish neither causality nor the magnitude of the effects. The data from household surveys do not come from controlled experiments in which the effects of a "treatment" can be unambiguously and convincingly determined.

In recent years, there has been a great deal of interest in social experiments, including the use of household survey data to evaluate the results of social experiments. Nevertheless, experiments are not always possible, and real experiments usually deviate from the ideal in ways that present their own difficulties of interpretation. In some cases, good luck, inspiration, and hard work throw up circumstances or data that allow a clear evaluation of policy effects in the absence of controlled experiments; these "quasi" or "natural" experiments have been the source of important findings as well as of some controversy. Even without such solutions, it seems as if it ought to be possible to use standard survey data to say something about the policy effects in which we are interested. A good starting point is to recognize that this will not always be the case. Many policy questions are not readily answerable at all, often because they are not well or sharply enough posed, and even when an answer is available in principle, there is no reason to suppose that it can be inferred from the data that happen to be at hand. Only when this is appreciated is there much chance of progress, or even of a realistic evaluation of what can be accomplished by empirical analysis.

Much of the empirical microeconomic literature in development uses econometric and statistical methodology to overcome the nonexperimental nature of data. A typical study would begin with a structural model of the process at hand, for example, of the effects on individual health of opening a new clinic. Integral to the model are statistical assumptions that bridge the gap between theory and data and so permit both the estimation of the parameters of the model and the subsequent interpretation of the data in terms of the theory. I have no difficulty with this approach in principle, but often find it hard to defend in practice. The statistical and economic assumptions are often arbitrary and frequently implausible. The econometric technique can be complex, so that transparency and easy replicability are lost. It becomes difficult to tell whether the results are genuine features of the data or are consequences of the supporting assumptions. In spite of these problems, I shall spend a good deal of space in Chapter 2 discussing the variety of econometric technique that is available for dealing with nonexperimental data. An understanding of these matters is necessary in order to interpret the literature, and it is important to know the circumstances in which technical fixes are useful.

Most of the analysis in this book follows a different approach which recognizes that structural modeling is unlikely to give convincing and clean answers to the policy questions with which we are concerned. Rather than starting with the theory, I more often begin with the data and then try to find elementary procedures for

describing them in a way that illuminates some aspect of theory or policy. Rather than use the theory to summarize the data through a set of structural parameters, it is sometimes more useful to present features of the data, often through simple descriptive statistics, or through graphical presentations of densities or regression functions, and then to think about whether these features tell us anything useful about the process whereby they were generated. There is no simple prescription for this kind of work. It requires a good deal of thought to try to tease out implications from the theory that can be readily checked against the data. It also requires creative data presentation and processing, so as to create useful and interesting stylized facts. But in the end, I believe that we make more progress, not by pretending to estimate structural parameters, but by asking whether our theories and their policy implications are consistent with well-chosen stylized facts. Such facts also provide convenient summaries of the data that serve as a background to discussions of policy. I hope that the examples in this book will make the case that such an approach can be useful, even if its aims are relatively modest.

Structure and outline

Household budget surveys collect information on who buys what goods and services and how much they spend on them. Information on how poor people spend their money has been used to describe poverty and to build the case for social reform since the end of the eighteenth century, and household surveys remain the basis for documenting poverty in developing countries today. When surveys are carried out on a regular basis, they can be used to monitor the welfare of various groups in society and to keep track of who benefits and who loses from development. Large-scale national surveys allow a good deal of disaggregation and allow us to look beyond means to other features of distributions, distinguishing households by occupational, regional, sectoral, and income groups.

In most poor countries, a large fraction of government revenue is raised by indirect taxes on goods and services, and many countries subsidize the prices of commodities such as basic foodstuffs. Household expenditure surveys, by revealing who buys each good and how much they spend, tell us who pays taxes and who benefits from subsidies. They thus yield a reckoning of the gainers and losers from a proposed changes in taxes and subsidies. When data are collected on the use of services provided by the state, such as health and education, we also discover who benefits from government expenditures, so that survey data can be used to assess policy reform and the effectiveness of government taxation and expenditure.

Data from household surveys are also a base for research, for testing theories about household behavior, and for discovering how people respond to changes in the economic environment in which they live. Some recent surveys, particularly the World Bank's Living Standards Measurement Surveys, have attempted to collect data on a wide range of household characteristics and activities, from fertility and physical measurement of weights and heights to all types of economic transactions. Such data allow us to examine all the activities of the household and to trace the behavioral links between economic events and individual welfare.

This book follows the progression of the previous three paragraphs, from data description through to behavioral analysis. Chapters 1 and 2 are preliminary to the main purpose and are concerned with the collection of household survey data, with survey design, and with its consequences for analysis. Chapter 1 is not meant to provide a guide to constructing surveys in developing countries, but rather to describe those features of survey design that need to be understood in order to undertake appropriate analysis. Chapter 2 discusses the general econometric and statistical issues that arise when using survey data for estimation and inference; the techniques discussed here are used throughout the rest of the book, but I also attempt to be more general, covering methods that are useful in applications not explicitly considered. This is not a textbook of econometrics; these two chapters are designed for readers with a basic knowledge of econometrics who want some preparation for working with household survey data particularly, but not exclusively, from developing countries.

Chapter 3 makes the move toward substantive analysis and discusses the use of survey data to measure welfare, poverty, and distribution. I review the theoretical underpinnings of the various measures of social welfare, inequality, and poverty and show how they can be given empirical content from survey data, with illustrations from the Ivorian and South African Living Standards Surveys. I highlight a number of techniques for data analysis that have proved useful in policy discussions, with particular emphasis on graphical methods for displaying large amounts of data. These methods can be used to investigate the distribution of income, inequality, and poverty and to examine changes in the levels of living of various groups over time. The chapter also shows how it is possible to use the data to examine the distributional consequences of price changes directly, without having to construct econometric models. These methods are applied to an analysis of the effects of rice price policy on the distribution of real income in Thailand.

Chapter 4 discusses the use of household budget data to explore patterns of household demand. I take up the traditional topic of Engel curve analysis in developing countries, looking in particular at the demand for food and nutrition. For many people, nutritional issues are at the heart of poverty questions in developing countries, and Engel curve analysis from survey data allows us to measure the relationship between the elimination of hunger and malnutrition and more general economic development, as captured by increases in real disposable income. This chapter also addresses the closely related question of how goods are allocated within the household and the extent to which it is possible to use *household* data to cast light on the topic. One of the main issues of interest is how different members of the household are treated, especially whether boys are favored over girls. Analyses of the effects of household composition on demand patterns can perhaps shed some light on this, as well as on the old but vexed question of measuring the "costs" of children. In most surveys, larger households have more income and more expenditure, but they also have less income or expenditure on a per capita basis. Does this mean that large households are poorer on average or that small households are poorer on average? The answer depends on whether there are economies of scale to large households—whether two people need twice as much as one—and

whether children, who are relatively plentiful in larger households, need less money to meet their needs than do adults. This chapter discusses the extent to which the survey data can be used to approach these questions.

Chapter 5 is about price reform, its effects on equity and efficiency, and how to measure them. Because surveys provide direct information on how much is consumed of each taxed or subsidized good, it is straightforward to calculate the first-round effects of price changes, both on revenue and on the distribution of real income. What are much harder to assess are the behavioral responses to price changes, the degree to which the demand for the good is affected by the change in price, and the extent to which revenues and expenditures from taxes and subsidies on other goods are affected. The chapter discusses methods for estimating price responses using the spatial price variation that is typically quite pronounced in developing countries. These methods are sensitive enough to detect differences in price responses between goods and to establish important cross-price effects between goods, effects that are often large enough to substantially change the conclusions of a policy reform exercise. Reducing a subsidy on one staple food has very different consequences for revenue and for nutrition, depending on whether or not there is a closely substitutable food that is also subsidized or taxed.

Chapter 6 is concerned with the role of household consumption and saving in economic development. Household saving is a major component and determinant of saving in most developing countries, and many economists see saving as the wellspring of economic growth, so that encouraging saving becomes a crucial component of a policy for growth. Others take the view that saving rates respond passively to economic growth, the roots of which must be sought elsewhere. Survey data can be used to explore these alternative views of the relationship between saving and growth, as well as to examine the role that saving plays in protecting living standards against fluctuations in income. The analysis of survey evidence on household saving, although fraught with difficulty, is beginning to change the way that we think about household saving in poor economies.

I have benefited from the comments of many people who have given generously of their time to try to improve my exposition, to make substantive suggestions, and in a few cases, to persuade me of the error of my ways. In addition to the referees, I should like to thank—without implicating any of them—Martha Ainsworth, Harold Alderman, Tony Atkinson, Dwayne Benjamin, Tim Besley, Martin Browning, Kees Burger, Lisa Cameron, David Card, Anne Case, Ken Chay, John Dinardo, Jean Drèze, Eric Edmonds, Mark Gersovitz, Paul Glewwe, Margaret Grosh, Bo Honoré, Susan Horton, Hanan Jacoby, Emmanuel Jimenez, Alan Krueger, Doug Miller, Juan Muñoz, Meade Over, Anna Paulson, Menno Pradhan, Gillian Paull, James Powell, Martin Ravallion, Jeremy Rudd, Jim Smith, T. N. Srinivasan, David Strömberg, Duncan Thomas, and Galina Voronov. I owe special thanks to Julie Nelson, whose comments and corrections helped shape Chapter 5, and to Christina Paxson, who is the coauthor of much of the work reported here. Some of the work reported here was supported by grants from the National Institute of Aging and from the John D. and Catherine T. MacArthur Foundation. The book was written for the Policy Research Department of the World Bank.

1 *The design and content of household surveys*

In his splendid essay on early studies of consumer behavior, Stigler (1954, p. 95) tells how the first collectors of family budgets, the Englishmen Reverend David Davies (1795) and Sir Frederick Morton Eden (1797), were "stimulated to this task by the distress of the working classes at this time." Davies used his results to draw attention to the living conditions of the poor, and to argue in favor of a minimum wage. The spread of working-class socialism in Europe in the late 1840s also spawned several compilations of household budgets, including the one of 200 Belgian households by Edouard Ducpetiaux in 1855 that was used two years later by Ernst Engel, not only as the basis for his law that the fraction of the budget devoted to food is larger for poorer families, but also to estimate the aggregate consumption, not of Belgium, but of Saxony! The use of budget data to expose poverty and living standards, to argue for policy reform, and to estimate national aggregates are all topics that are as relevant today as they were two centuries ago. The themes of the research were set very early in the history of the subject.

The early investigators had to collect data where they could find it, and there was no attempt to construct representative samples of households. Indeed, the understanding that population totals can be estimated from randomly selected samples and the statistical theory to support such estimation were developed only in the first quarter of this century. Around the turn of the century, Kiaer in Norway and Wright in the United States were among the first to use large-scale representative samples, but the supporting statistical theory was not fully worked out until the 1920s, with Bowley, Ronald Fisher, and Neyman making important contributions. The acceptance of sampling is well illustrated by the case of Rowntree, who was unpersuaded by the reliability of sampling when he undertook his survey of poverty in the city of York in 1936. Having collected a full census, he was later convinced by being able to reproduce most of the results from samples drawn from his data (see the supplementary chapter in Rowntree 1985). One of the first large-scale scientific surveys was carried out by Mahalanobis in Calcutta, who estimated the size of the jute crop in Bengal in 1941 to within 2.8 percent of an independent census at less than 8 percent of the cost—see Mahalanobis (1944, 1946) for the classic early accounts, and Seng (1951) and Casley and Lury (1981, ch. 1) for more history and citations to the early literature.

Modern household surveys begin after World War II. Under the leadership of Mahalanobis at the Indian Statistical Institute in Calcutta, the Indian National Sample Survey (NSS) started the annual collection of household consumption data in 1950. Many other economies, both industrialized and developing, now have regular household consumption surveys, sometimes on an annual basis, as in India until 1973–74, or in Taiwan, The Republic of Korea, Britain, and the United States today, but more often less frequently, as for example in India after 1973–74 (quinquennially), the United States prior to 1980, and many other countries. These surveys were often intended to provide data on poverty and income distribution, for example in the form of frequency distributions of households by levels of living—usually defined by per capita income or consumption—but this was by no means their only purpose. In many cases, the surveys were designed to produce *aggregate* data, to help complete the national accounts, to provide weights for consumer price indexes, or to provide the basis for projecting demand patterns in planning exercises. Once begun, it was typically difficult to change the mode of operation or to use the data for purposes different from those in the original design. The former would generate incompatibilities and inconsistencies in the data, while the latter required a computational capacity and willingness to release household-level data that were rarely in evidence. There are, of course, genuine confidentiality issues with household information, but these can be met by removing some information from the publicly available data, and hardly justify their treatment as state secrets.

Recent years have seen a marked change in survey practice, in data collection, and in analysis. Although there are still laggard countries, many government statistical offices have become more open with their data and have given bona fide researchers and international organizations access to the individual household records. Reductions in the real cost of computation have led to more analysis, although it is only in the last few years that mass-storage devices and cheap memory have made it convenient to use microcomputers to analyze large data sets. Perhaps as important have been changes in the design of surveys, and there is now a much wider range of survey instruments in use than was the case a decade ago. Following a number of experimental and innovative surveys in the 1960s and 1970s—particularly the Malaysian Family Life Survey in 1976–77—the World Bank's Living Standards Surveys first collected data in Peru and Côte d'Ivoire in 1985 and incorporated important innovations in data collection and in content. Originally designed to improve the World Bank's ability to monitor poverty and to make international comparisons of living standards, poverty, and inequality, they evolved into vehicles for collecting comprehensive information on a wide range of household characteristics and activities. The rapid availability and ease of analysis of survey data has led to a productive feedback from analysis to design that was rare prior to 1980. In consequence, survey practice and questionnaire design are probably changing more rapidly now than ever before.

This chapter and the next, which are preliminary to the analytical studies in the rest of the book, are concerned with the design and content of household surveys (this chapter) and with its implications for statistical and econometric analysis (the final section of this chapter and Chapter 2). In line with the substantive studies later

in the book, I give disproportionate attention to income and expenditure surveys, or to the income and expenditure sections of broader, integrated surveys such as the Living Standards Surveys. Even so, much of the discussion carries over to other types of household survey, for example to employment or fertility surveys, though I do not give explicit attention to those topics.

The four sections of this chapter are concerned with the design of surveys, with the type of data that they collect, and with the effect of design on the calculation of descriptive statistics such as means. Section 1.1 discusses the practical and statistical issues concerned with choosing households for inclusion into a survey. Section 1.2 is concerned with the types of data that are usually collected, and their likely quality. Section 1.3 focusses on the particular features of the Living Standards Surveys from which data are used in some of the later chapters. Many of the policy analyses that use household survey data were not contemplated when the surveys were designed, so that mechanical calculations that ignore the design of the survey can produce unpredictable results. For example, surveys that are designed to estimate means or population totals may be quite unsuitable for measuring dispersion. In most surveys some types of households are overrepresented relative to their share in the population, while others are underrepresented, so that corrections have to be made to calculate genuinely representative totals. It is also wise to be sensitive to the possibility—in most cases the certainty—of measurement errors, to their effects on the calculations, and to strategies that can be used to protect inference in their presence. These issues are further complicated when, as in some of the Living Standards Surveys, households are observed on more than one occasion, and we are interested in analyzing changes in behavior over time. Section 1.4, which is more technical than the others, presents some of the most useful formulas for estimating means and their sampling variability taking into account the survey design. The discussion is useful both for Chapter 2, where I move from descriptive statistics to a more econometric approach, and for Chapter 3, where I deal with poverty measures, which are a particular kind of descriptive statistic. This section also contains a brief introduction to the bootstrap, a technique that is often useful for calculating standard errors and confidence intervals.

1.1 Survey design

The simplest household survey would be one where there exists a reliable, up-to-date list of all households in the population, where the design assigns an equal probability to each household being selected from the list to participate in the survey and where, in the implementation stage, all households asked to participate actually do so. The sample would then be a simple random sample, with each household standing proxy for an equal number of households in the population. Such samples are easy to use and a few actual surveys approximate this simple structure. However, for a number of good and some not-so-good reasons, most surveys are a good deal more complex. I begin by discussing the list (or frame) from which households are selected and which defines the potential coverage of the survey, and then pass on to stratification and sampling issues.

Survey frames and coverage

A typical household survey collects data on a national sample of households, randomly selected from a "frame" or national list of households. Sample sizes vary widely depending on the purpose of the survey, on the size of the population in the country being surveyed, and on the degree to which regional or other special subsamples are required. Sample sizes of around 10,000 are frequently encountered, which would correspond to a sampling fraction of 1:500 in a population of 5 million households, or perhaps 25 million people. Since the accuracy of sample statistics increases less than proportionally with the sample size—usually in proportion to its square root—sampling fractions are typically smaller in larger populations, a tendency that is reinforced by limits on the size of survey that can be mounted by many data collection agencies. Nevertheless, there are some very large surveys such as the current Indian NSS, where a full national sample contains around a quarter of a million households.

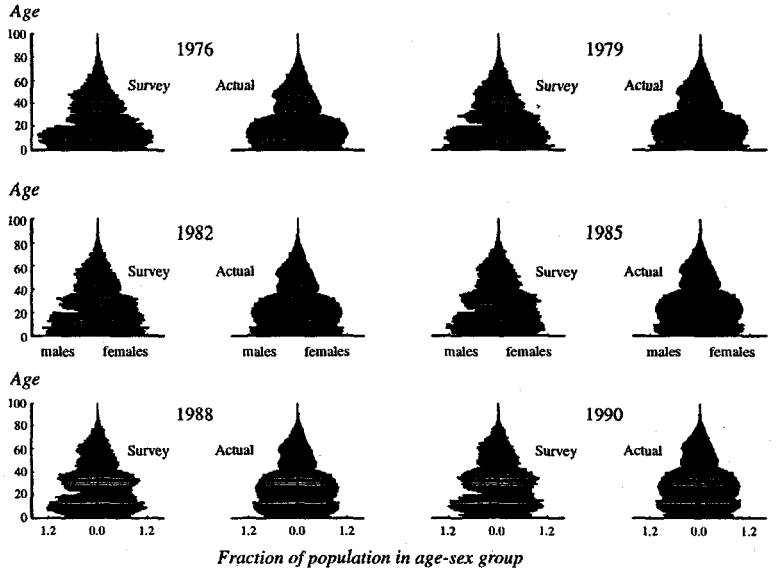
The frame is often a census, which in principle provides a list of all households and household members, or at least of all dwellings. However, there are many countries where there is no up-to-date census, or no reliable recent census, so that other frames have to be constructed, usually from administrative records of some kind (see Casley and Lury 1981, ch. 6, who discuss some of the possibilities). Perhaps the most common method of selecting households from the frame uses a two-stage design. At the first stage, selection is from a list of "clusters" of households, with the households themselves selected at the second stage. In rural areas, the clusters are often villages but the choice will depend on the frame. Censuses have their own subunits that are suitable for first-stage sampling. Once the clusters are chosen, households can be selected directly if an up-to-date list is available, and if the list is detailed enough to allow identification in the field. Otherwise, all households in the selected clusters can be listed prior to the second stage. Since it is often possible to include some household information at the listing stage, the procedure allows the second-stage selection of individual households to be informed by prior knowledge, a possibility to which I shall return in the next subsection. Note finally that two-stage sampling is not inconsistent with each household in the population having an equal chance of selection into the sample. In particular, if clusters are randomly selected with probability proportional to the number of households they contain, and if the same number of households is selected from each cluster, we have a *self-weighting* design in which each household has the same chance of being included in the survey.

The use of outdated or otherwise inaccurate frames is an important source of error in survey estimates. It should also be noted that in some countries—including the United States—censuses are politically sensitive so that various interest groups can be expected to try to interfere with the count. Even when the frame is accurate in itself, its coverage of the population will typically not be complete. Homeless people are automatically excluded from surveys that start from *households*, and in many countries people living in various institutional settings—the armed forces or workers' dormitories—will be excluded.

One example of the differences between a sample and population is provided by the data in Figure 1.1, which shows age-sex pyramids for Taiwan for selected years between 1976 and 1990. There are two pyramids for each year; those on the left-hand side were calculated from household survey data, and were previously reported in Deaton and Paxson (1994a), while those on the right-hand side are calculated from the official population data in Republic of China (1992). The survey data, which are described in detail in Republic of China (1989), come from a set of surveys that have been carried out on a regular basis since 1976, and that are carefully and professionally conducted. The 1976 sample has data on some 50,000 individuals, while the later years cover approximately 75,000 persons; the population of Taiwan grew from 16.1 million in 1975 to 20.4 million in 1990.

The differences between the two sets of pyramids is partly due to sampling error—each year of age is shown in the graphs—but there are also a number of differences in coverage. The most obvious of these causes the notches in the sample pyramids for men aged 18 to 20. Taiwanese men serve in the military during those years and are typically not captured in the survey, and roughly two-thirds of the age group is missing. These notches tend to obscure what is one of the major common features of both population and sample, the baby boom of the early 1950s. In 1988 and 1990, there is some evidence that the survey is missing young women, although the feature is much less sharp than for men and is spread over a wider age range. The design feature in this case is that the survey does not include women attending college nor those living in factory dormitories away from home. As the

Figure 1.1. Age and sex pyramids for survey data and population, Taiwan (China), selected years, 1976–90



Source: Author's calculations using survey data tapes, and Republic of China (1992).

population pyramids make clear, some of these women—together with the men in the same age group—are genuinely "missing" in the sense that the cohort of babies born around 1975 is substantially smaller than those immediately preceding or succeeding it. A number of other distinctive features of these graphs are not design effects, the most notable being the excess of men over women that is greatest at around age 45 in 1976, and moves up the age distribution, one year per year, until it peaks near age 60 in 1990. These men are the survivors of Chiang Kai-shek's army who came to Taiwan after their defeat by the communists in 1949.

The noncoverage of some of the population is typical of household surveys and clearly does not prevent us from using the data to make inferences. Nevertheless, it is always wise to be careful, since the missing people were not missed at random and will typically have different characteristics from the population as a whole. In the Taiwanese case, we should be careful not to infer anything about the behavior of young Taiwanese males. Another notable example comes from Britain, where the annual Family Expenditure Survey (FES) regularly underestimates aggregate alcohol consumption by nearly a half. Much of the error is attributed to coverage; there is high alcohol consumption among many who are excluded from the survey, primarily the military, but also innkeepers and publicans (see Kemsley and others 1980). To the effects of noncoverage by design can be added the effects of nonrespondents, households that refuse to join the survey. Nonresponse is much less of a problem in developing countries than in (for example) the United States, where refusal to participate in surveys has been increasing over time. Although many surveys in developing countries report almost complete cooperation, there will always be specific cases of difficulty, as when wealthy households are asked about incomes or assets, or when households are approached when they are preoccupied with other activities. Once again, some of the low alcohol reports in the British data reflect the relatively low response rates around Christmas, when alcohol consumption is highest (see Crooks 1989, pp. 39–44). It is sometimes possible to study survey nonresponse patterns by tracing refusals in a contemporaneous census, using data from the census to assess the determinants of refusal in the survey (see Kemsley 1975 for such an exercise for the British FES using the 1971 census). Sometimes the survey itself will collect some information about nonrespondents, for example about housing. Groves (1989, ch. 4) discusses these and other techniques for assessing the consequences of nonresponse.

Strata and clusters

A two-stage sample design, first selecting clusters and then households, generates a sample in which sample households are not randomly distributed over space, but are geographically grouped. This arrangement has a number of advantages beyond the selection procedure. It is cost-effective for the survey team to travel from village to village, spending substantial time in each, instead of having to visit households that are widely dispersed from one another. Clustered samples also facilitate repeat visits to collect information from respondents who may not have been present at the first visit, to monitor the progress of record keeping, or to ask

supplementary questions about previous responses that editing procedures have marked as suspect. That there are several households in each village also makes it worthwhile to collect village-level information, for example on schools, clinics, prices, or agroclimatic conditions such as rainfall or crop failures. (Though the clusters defined for statistical purposes will not always correspond to well-defined "communities.") For all these reasons, nearly all surveys in developing countries (and elsewhere, with telephone surveys the notable exception) use clustered samples.

The purposes of the survey sometimes dictate that some groups be more intensively sampled than others, and more often that coverage be guaranteed for some groups. There may be an interest in investigating a "target" group that is of particular concern, and, if members of the group are relatively rare in the population as a whole, a simple random sample is unlikely to include enough group members to permit analysis. Instead, the sample is designed so that households with the relevant characteristic have a high probability of being selected. For example, the World Bank used a Living Standards–type survey in the Kagera region of Tanzania to study the economic effects of AIDS. A random sample of the population would not produce very many households with an infected person, so that care was taken to find such households by confining the survey to areas where infection was known to be high and by including questions about sickness at the listing stage, so that households with a previous history of sickness could be oversampled.

More commonly, the survey is required to generate statistics for population subgroups defined (for example) by geographical area, by ethnic affiliation, or by levels of living. *Stratification* by these groups effectively converts a sample from one population into a sample from many populations, a single survey into several surveys, and guarantees in advance that there will be enough observations to permit estimates for each of the groups.

There are also statistical reasons for departing from simple random samples; quite apart from cost considerations, the precision of any given estimate can be enhanced by choosing an appropriate design. The fundamental idea is that the surveyor typically knows a great deal about the population under study prior to the survey, and the use of that prior information can improve the efficiency of statistical inference about quantities that are unknown. Stratification is the classic example.

Suppose that we are interested in estimating average income, that we know that average rural incomes are lower than average urban incomes, and we know the proportions of the population in each sector. A stratified survey would be two identical surveys, one rural and one urban, each of which estimates average income. (It would not necessarily be the case that the sampling fractions would be the same in each stratum.) The average income for the country as a whole, which is the quantity in which we are interested, is calculated by weighting together the urban and rural means using the proportions of the population in each as weights—which is where the prior information comes in. The precision of this combined estimate is assessed (inversely) from its variance over replications of the survey. Because the two components of the survey are independent, the variance of the overall mean is the sum of the variances of the estimates from each strata. Hence, variance de-

pends only on *within*-sector variance, and not on *between*-sector variance. If instead of a stratified survey, we had collected a simple random sample, the variance of the overall mean would still have depended on the within-sector variances, but there would have been an additional component coming from the fact that in different surveys, there would have been different fractions of the sample in rural and urban. If rural and urban means are different, this variability in the composition of the sample will contribute to the variability of the estimate of the overall mean. In consequence, stratification will have the largest effect in reducing variance when the stratum means are different from one another, and when there is relatively little variation within strata. The formulas that make this intuition precise are discussed in Section 1.4 below.

In household income and expenditure surveys, rural and urban strata are nearly always distinguished, and sometimes there is additional geographic stratification, by regions or provinces, or by large and small towns. Ethnicity is another possible candidate for stratification, as is income or its correlates if, as is often the case, some indication of household living standards is included in the frame or in the listing of households—landholdings and housing indicators are the most frequent examples. Stratification can be done explicitly, as discussed above, or "implicitly." The latter arises using "systematic" sampling in which a list of households is sampled by selecting a random starting point and then sampling every *j*th household thereafter, with *j* set so as to give the desired sample size. Implicit stratification is introduced by choosing the order in which households appear on the list. An example is probably the best way to see how this works. In the 1993 South African Living Standards Survey, a list was made of clusters, in this case "census enumerator subdistricts" from the 1991 census. These clusters were split by statistical region and by urban and rural sectors—the *explicit* stratification—and then in order of percentage African—the *implicit* stratification. Given that the selection of clusters was randomized only by the random starting point, the implicit stratification guarantees the coverage of Africans and non-Africans, since it is impossible for a sample so selected not to contain clusters from high on the list, which are almost all African, and clusters low on the list, which are almost all non-African.

While stratification will typically enhance the precision of sampling estimates, the clustering of the sample will usually reduce it. The reason is that households living in the same cluster are usually more similar to one another in behavior and characteristics than are households living in different clusters. This similarity is likely to be more pronounced in rural areas, where people living in the same village share the same agroclimatic conditions, face similar prices, and may belong to the same ethnic or tribal group. As a result, when we sample several households from the same cluster, we do not get as much information as we would from sampling several households from different clusters. In the (absurd) limit, if everyone in the same cluster were replicas or clones of one another, the effective sample size of the survey would not be the number of households, but the number of clusters. More generally, the precision of an estimate will depend on the correlation within the cluster of the quantity being measured; once again, the formulas are given in Section 1.4 below.

A useful concept in assessing how the sample design affects precision is Kish's (1965) "design effect," often referred to as *deff*. *Deff* is defined as the ratio of the variance of an estimate to the variance that it would have had under simple random sampling; some explicit examples are included in Section 1.4. Stratification tends to reduce *deff* below one, while clustering tends to increase it above one. Estimates of the means of most variables in stratified clustered samples have *deffs* that are greater than one (Groves 1989, ch. 6), so that in survey design the practical convenience and cost considerations of clustering usually predominate over the search for variance-reduction.

Unequal selection probabilities, weights, and inflation factors

As we have seen, it is possible for a survey to be stratified and clustered, and for each household in the population to have an equal probability of inclusion in the sample. However, it is more common for probabilities of inclusion to differ, because it costs more to sample some households than others, because differential probabilities of inclusion can enhance precision, and because some types of households may be more likely to refuse to participate in the survey. Because noncooperation is rarely taken into account in design, even samples that are meant to have equal probabilities of selection often do not do so in practice.

Variation in costs is common, for example between rural and urban households. In consequence, the cost of any given level of precision is minimized by a sample in which urban households are overrepresented and rural households underrepresented. The use of differential selection probabilities to enhance precision is perhaps less obvious, but the general principle is the same as for stratification, that prior information can be used to tell us where to focus measurement. To fix ideas, suppose again that we are estimating mean income. The estimate will be more precise if households that contribute a large amount to the mean—high-income households—are overrepresented relative to low-income households, who contribute little. This is "probability proportional to size," or *p.p.s.*, sampling. Of course, we do not know household income, or we would not have to collect data, but we may have information on correlated variables, such as landholdings or household size. Overrepresentation of large households or large landholding households will typically lead to more precise estimates of mean income (see again Section 1.4 for formulas and justification).

When selection probabilities differ across households, each household in the survey stands proxy for or represents a different number of households in the population. In consequence, when the sample is used to calculate estimates for the population, it is necessary to weight the sample data to ensure that each group of households is properly represented. Sample means will not be unbiased estimates of population means and we must calculate weighted averages so as to "undo" the sample design and obtain estimates to match the population. The rule here is to weight according to the reciprocals of sampling probabilities because households with low (high) probabilities of selection stand proxy for large (small) numbers of households in the population. These weights are often referred to as "raising" or

"inflation" factors because if we multiply each observation by its inflation factor we are estimating the total for all households represented by the sample household, and the sum of these products over all sample households is an estimate of the population total. Inflation factors are typically included in the data sets together with other variables.

Table 1.1 shows means and standard deviations by race of the inflation factors for the 1993 South African survey. This is an interesting case because the original design was a self-weighting one, in which there would be no variation in inflation factors across households. However, when the survey was implemented there were substantial differences by race in refusal rates, and there were a few clusters that could not be visited because of political violence. As a result, and in order to allow the calculation of unbiased estimates of means, inflation factors had to be introduced after the completion of the fieldwork. The mean weight for the 8,848 households in the survey is 964, corresponding to a population of households of 8,530,808 (= 8,848 × 964). Because whites were more likely to refuse to participate in the survey, they attract a higher weight than the other groups.

This South African case illustrates an important general point about survey weights. Differences in weights from one household to another can come from different probabilities of selection *by design*, or from different probabilities *by accident*, because the survey did not conform to the design, because of non-response, because households who cooperated in the past refused to do so again, or because some part of the survey was not in fact implemented. Whether by design or accident, there are ex post differences in sampling probabilities for different households, and weights are needed in order to obtain accurate measures of population quantities. But the design weights are, by construction, the reciprocals of the sampling probabilities, and are thus controlled in a way that accidental weights are not. Weights that are added to the survey ex post do not have the same pedigree, and are often determined by judgement and modeling. In South Africa, the response rate among White households was lower, so the weights for White households were adjusted upwards. But can we be sure that the response rate was truly determined by race, and not, for example, by some mixture of race, income, and location? Adoption of survey weights often involves the implicit acceptance of modeling decisions by survey staff, decisions that many investigators would prefer to keep to them-

Table 1.1. Inflation factors and race, South Africa, 1993

Race	*Mean weight*	*Standard deviation*	*Households in sample*
Blacks	933	79	6,533
Coloreds	955	55	690
Asians	885	22	258
Whites	1,135	219	1,367
All	964	133	8,848

Source: Author's calculations using the South African Living Standards Survey, 1993.

selves. At the least, survey reports should document the construction of such weights, so that other researchers can make different decisions if they wish.

Sample design in theory and practice

The *statistical* arguments for stratification and differential sampling probabilities are typically less compelling in developing-country surveys than are the *practical* arguments. Optimal design for precision works well when the aim of the survey is the measurement of a single magnitude—average consumption, average income, or whatever. Once this objective is set, all the tools of the sample survey statistician can be brought to bear to design a survey that will deliver the best estimate at the lowest possible cost. Such single-purpose surveys do indeed occur from time to time and more frequently there is a main purpose, such as the estimation of weights for a consumer price index, or the measurement of poverty and inequality. Even in these cases, however, it is recognized that there are other uses for the data, and in general-purpose household surveys there is a range of possible applications, each of which would mandate a different design. Precision for one variable is imprecision for another, and it makes no sense to design a survey for each. In addition, optimizing for one purpose can make it difficult to use the survey for other purposes. A good example is the Consumer Expenditure Survey in the United States, where the main aim is the calculation of weights for the consumer price index. That object is relentlessly pursued, with some expenditures obtained by interviewing some households, some expenditures obtained by diary from other households, and each household is visited five times over fifteen months but with different kinds of data collected at each visit. All of this allows a relatively small sample to deliver good estimates of the average American spending pattern, but the complexity of the design makes it difficult—sometimes even impossible—to make calculations that would have been possible under simpler designs.

Another problem with optimal schemes is that the selection of households according to efficiency criteria can compromise the usefulness of the data. For example, the use of public transport is efficiently estimated by interviewing travelers, and travelers are most easily and economically found by conducting "on-board" surveys on trains, on buses, or at stations. But if we are to study what determines the demand for travel, and who benefits from state subsidies to public transport, we need to know about nontravelers too, information that is better collected in standard household surveys. Indeed, if observations are selected into the sample according to characteristics that are correlated with the magnitude being studied—precisely the recipe in *p.p.s.* sampling—attempts to estimate models that explain that magnitude are likely to be compromised by the selection of the sample. This "choice-based sampling" problem has been studied in the literature (see Manski and Lerman 1977, Hausman and Wise 1977, 1981, and Cosslett 1993) and there exist techniques for overcoming the difficulties. But once again it is much easier to work with a simpler survey, and the results are likely to be more comprehensible and more convincing if they do not require complex corrections, especially when the corrections are supported by assumptions that are difficult or impossible to check.

There are also good practical reasons for straightforward designs. In their book on collecting data in developing countries, Casley and Lury (1981, p. 2) summarize their basic message in the words "keep it simple." As they point out:

> The sampling errors of any rational design involving at least a moderate sample size are likely to be substantially smaller than the nonsampling errors. Complications of design may create problems, resulting in larger nonsampling errors, which more than offset the theoretical benefits conferred.

As we shall see, the econometric analysis will have to deal with a great many problems, among which nonsampling errors are not the least important. Correction for complex designs is an additional task that is better avoided whenever possible.

Panel data

The standard cross-sectional household survey is a one-time affair and is designed to obtain a snapshot of a representative group of households at a given moment in time. Although such surveys take time to collect (frequently a year) so that the "moment in time" varies from household to household, and although households are sometimes visited more than once, for example to gather information on income during different agricultural seasons, the aim of the survey is to gather information from each household about a given year's income, or about consumption in the month previous to the interview, or about the names, sexes, and ages of the members of the household on the day of the interview.

By contrast, longitudinal or panel surveys track households over time, and collect multiple observations on the same household. For example, instead of gathering income for one year, a panel would collect data on income for a number of years, so that, using such data, it is possible to see how survey magnitudes change for individual households. Thus, the great attraction of panel data is that they can be used to study dynamics for individual households, including the dynamics of living standards. They can be used to address such issues as the persistence of poverty, and to see who benefits and who loses from general economic development, or who gains and loses from a specific shock or policy change, such as a devaluation, a structural adjustment package, or a reduction in the prices of commodity exports. However, as we shall see in Chapter 2, panel data are *not* required to track outcomes or behavior for *groups* of individuals—that can be done very well with repeated cross-sectional surveys—but they are the only data that can tell us about dynamics at the individual level. Panel surveys are relatively rare in general, and particularly so in developing countries. The panel that has attracted the greatest attention in the United States is the Michigan Panel Study of Income Dynamics (PSID), which has been following the members of about 4,800 original households since 1968. The most widely used panel data from a developing country come from the Institute for Crop Research in the Semi-Arid Tropics (ICRISAT) in Hyderabad, India, which followed some 40 agricultural households in each of six villages in southwestern India for five or ten years between 1975 and 1985.

These long-standing surveys are not the only way in which panel data can be collected; an alternative is a *rotating panel* design in which some fraction of households is held over to be revisited, with the rest dropped and replaced by new households. Several of the Living Standards Surveys—to be discussed in Section 1.3 below—have adopted such a design. For example, in Côte d'Ivoire, 1,600 households were selected into the 1985 survey, 800 of which were retained in 1986. To these original panelists 800 new households were added in 1986, and these were retained into 1987. By the pattern of rotation, no household is observed for more than two years, so that while we have two observations in successive years on each household (apart from half of the start-up households) we do not get the long-term observation of individual households that come from panel data.

A third way of collecting data is to supplement cross-sectional data. Occasionally this can be done by merging administrative and survey data. More often, an earlier cross-sectional survey is used as the basis for revisiting households some years after the original survey. If records have been adequately preserved, this can be done even when there was no intention in the original survey of collecting panel data; indeed, it is good practice to design any household survey so as to maximize the probability of recontacting the original respondents. Such methods have been successful in a number of instances. Although the Peruvian Living Standards Survey of 1985–86 was designed as a cross section, households living in Lima were revisited in 1990; of the 1,280 dwellings in the original survey, 1,052 were reinterviewed (some dwellings no longer existed, or the occupants refused to cooperate) and, of these, 745 were occupied by the same family (Glewwe and Hall 1995). In 1988–89, RAND carried out a successful reinterview of nearly three-quarters of the individuals in the original 1976–77 Malaysian Family Life Survey (Haaga, DaVanzo, Peterson, and Peng 1994). Bevan, Collier, and Gunning (1989, Appendix) also appear to have been successful in relocating a high fraction of households in East Africa; in Kenya a 1982 survey reinterviewed nearly 90 percent of survey households first seen in 1977–78, while in Tanzania, 73 percent of the households in a 1976–77 survey were reinterviewed in 1983.

In some cases, panel data can be constructed from a single interview by asking people to recall previous events. This works best for major events in people's lives, such as migration or the birth or death of a child; it is likely to be much more difficult to get an accurate recollection of earnings or expenditures in previous years. There is a substantial literature on the accuracy of recall data, and on the various biases that are induced by forgetting and selective memory (see Groves 1989, ch. 9.4). In the context of developing countries, Smith, Karoly, and Thomas (1992) and Smith and Thomas (1993) use their repeat of the Malaysian Family Life Survey to compare recollections about migrations in the first and second surveys.

As well as their unique advantages, panel data have a number of specific problems. One of the most serious is *attrition*, whereby for one reason or another, households are lost from the survey, so that as time goes on, fewer of the original households remain in the survey. The extent of attrition is affected by the design of the panel, whether or not the survey follows individuals who leave the original households or who move away from the original survey area. Another reason for

attrition is refusal; households that have participated once are sometimes unwilling to do so again. Refusal rates are typically lower in surveys in developing countries, and presumably attrition is too. In industrial countries with long-running surveys such as the PSID, there can be a substantial loss of panel members in the first few years until the panel "settles down." Becketti and others (1988) show that although 12 to 15 percent of the individuals in the PSID do not reappear after the first interview, the subsequent attrition rate is much lower so that, for example, of the individuals in the first wave in 1968, 61.6 percent were still present fourteen years later.

Even when households are willing to cooperate, there may be difficulties in finding them at subsequent visits; individuals may move away, and the households may cease to exist if the head dies, or if children split off to form households of their own. Depending on whether the survey attempts to follow these migrants and "splits," as well as whether new births or immigrants are added to the sample, the process of household dissolution and formation can result in changes in the representativeness of the sample over time. (Or what appears to be a panel may not be, if enumerators substitute the new household for the old one without recognizing or recording the change.) There is therefore likely to be a tradeoff between, on the one hand, obtaining a representative sample, which is best done by drawing a new sample each year and, on the other hand, tracking individual dynamics, which requires that households be held over from year to year. Even so, Becketti and others (1988) found no serious problems of representativeness with the PSID when they compared the fourteen-year-old panel with the population of the United States.

Although the main attractions of panel data are for analytical work, for the measurement of dynamics and for controlling for individual histories in assessing behavior, panel designs can also enhance the precision of estimates of aggregate or average quantities. The standard example is estimating changes. Suppose that we compare the case of two independent cross sections with a panel, in which the same households appear in the two time periods. From both designs, the change in average income, say, would be estimated by the difference in average incomes in the two periods. The variance of the estimate from the two cross sections would be the sum of the variances in the two periods because each cross-sectional sample is drawn independently. In the panel survey, by contrast, the same households appear in both periods, so that the variance of the difference is the sum of the variances of the individual means *less* twice the covariance between the two estimates of mean income. If there is a tendency for the same households to have high (or low) incomes in both periods—which we should expect for incomes and will be true for many other quantities—the covariance will be positive and the variance of the estimated change will be less than the sum of the variances of the two means.

The greatest precision will be obtained from a panel, a rotating panel, or independent cross sections depending on the degree of temporal autocorrelation in the quantity being estimated. The higher the autocorrelation, the larger the fraction of households that should be retained from one period to the next. The formulas are given in Hansen, Hurwitz, and Madow (1953, pp. 268–72) and are discussed in the context of developing countries by Ashenfelter, Deaton, and Solon (1986). Provided precision is the main aim, a rotating panel is a good compromise; for example,

retaining only half the households from one period to the next will give a standard error for the change that is at most 30 percent larger than the standard error from the complete panel that is the optimal design, and will do better than this when the autocorrelation is low. Given that most surveys are multipurpose, and that there is a need to measure levels as well as changes, there is a good argument for considering rotating panels.

When using panel data to measure differences, it is important to be alive to the possibility of measurement (nonsampling) error and to its consequences for various kinds of analysis; indeed, the detection and control of measurement error will be one of the main refrains of this book. Suppose, to fix ideas, that household i in period t reports, not the true value x_{it} but x_{it}^* defined by

$$x_{it}^* = x_{it} + \epsilon_{it} \tag{1.1}$$

where ϵ_{it} is a mean zero measurement error with cross-sectional variance ω^2, and where I assume for convenience that the error variance is the same in both periods. If the reporting error is uncorrelated with the truth, then the variance of measured x is the variance of the true x—the signal—plus the variance of the measurement error—the noise. If the observations are differenced over time, we have

$$\Delta x_{it}^* = \Delta x_{it} + \Delta \epsilon_{it}: \quad \text{var}(\Delta x_{it}^*) = \text{var}(\Delta x_{it}) + 2\omega^2(1-\rho) \tag{1.2}$$

where ρ is the correlation between the errors in the two periods.

There are several important consequences of (1.2). Note first that the presence of measurement error is likely to further enhance the advantages of panel data over independent cross sections for measuring changes in the means, at least if the same individuals tend to make the same reporting errors period after period. Second, the signal-to-noise ratio will be different for the changes in (1.2) than for the levels in (1.1). The prototypical example is where the underlying variable x changes only slowly over time, so that the variance of the true changes is smaller than the variance of the levels. By contrast, the variance of the measurement error in changes will be double that in levels if $\rho = 0$, and will be increased by the differencing unless $\rho > 0.5$. There is no general result here, but there will be many cases in household survey data where the variance of the measured changes will be dominated by measurement error, even when the measurement of the levels is relatively accurate.

The data in Table 1.2 are taken from the 1985–86 panel of the Living Standards Survey of Côte d'Ivoire and illustrate a number of these issues. The figures shown are summary statistics for consumption and income for 730 panel households who were in the survey in both years. Although the original design called for 800 panelists, not all households could be found in the second wave, nor yielded useful data. While there is no direct way of assessing the size of the measurement error, both magnitudes are hard to estimate, and for the reasons discussed in the next section, the individual data are likely to be very noisy. The upper part of the table shows that means and standard deviations of consumption and income are of similar size, so that with 730 households, the standard errors of the estimates of the mean levels

Table 1.2. Consumption and income for panel households, Côte d'Ivoire, 1985–86
(thousands of CFAs per month)

	Mean	*Standard deviation*	*Median*	*Interquartile range*
Levels				
Consumption, 1985	1,561	1,513	1,132	1,344
Consumption, 1986	1,455	1,236	1,070	1,090
Income, 1985	1,238	1,464	780	1,137
Income, 1986	1,332	1,525	871	1,052
Differences 1986–85				
Consumption	-106	987	-17	679
Income	94	1,128	92	723

Notes: The figures shown are for total consumption and disposable income, both including imputed rental values of housing and durable goods. Data from 730 panel households.
Source: Author's and World Bank calculations using the Côte d'Ivoire Living Standards Surveys, 1985–86.

are about 4 percent ($1/\sqrt{(730)} \approx 0.04$) of the means. The changes in the bottom two rows have much smaller means than do the levels, and although the standard deviations are also smaller—the correlation coefficients between the years are 0.76 for consumption and 0.72 for income—the standard errors of the estimates of mean change are now relatively much larger. The correlation between the variables is also affected by the differencing. In the micro data, the levels of consumption and income are strongly correlated in both years, 0.81 in 1985 and 0.78 in 1986, but the correlation is only 0.46 for the first differences. There is, of course, nothing in these figures that proves that measurement error is in fact present, let alone that it is determining the outcomes. As we shall see in Chapter 6, there is no difficulty in accounting for the results in Table 1.2 even under the implausible assumption that consumption and income are perfectly measured. However, the phenomena that we observe here are typical of what happens when measurement error is present and important.

1.2 The content and quality of survey data

One of the main reasons for collecting household survey data is the measurement and understanding of living standards. At the least, such measurement requires data on consumption, income, household size, and prices. For broader concepts of living standards, we also want information on health, nutrition, and life expectancy, and on levels of education, literacy, and housing. Moving from measurement to modeling extends the scope a good deal wider. To understand consumption, we need to know about income and assets, and about their determinants, saving behavior, inheritances, education, and the opportunities for working in the labor market, on a farm, or in small businesses of one kind or another. We also need information

about public goods, such as schools and hospitals, and on individuals' access to them. In the past, different types of data often have been collected by different kinds of surveys. Budget surveys collected data on consumption and its components; income and employment surveys have collected data on sources of income, on occupations, and on unemployment; fertility surveys have collected data on children ever born, contraceptive practices, and attitudes towards fertility; and nutritional surveys have collected data on how much food people consume, how it is prepared and eaten, and its calorie, protein, and nutrient content. The Living Standards Surveys, described in Section 1.3, collect data on almost all of these topics from the same households in a single survey, sacrificing sample size for an integrated treatment of a relatively small number of households. This section is concerned with data quality issues that arise more or less independently whether we are dealing with multipurpose or more focussed surveys. Again, I make no attempt to be comprehensive; this section is not a list of what does or ought to appear in the ideal survey. My aim is rather to introduce a number of definitional and measurement issues that will recur throughout the book when using the data, and of which it is necessary to be aware in order to analyze them appropriately.

Individuals and households

The standard apparatus of welfare economics and welfare measurement concerns the well-being of individuals. Nevertheless, a good deal of our data have to be gathered from households, and while in some cases—earnings or hours worked—data are conceptually and practically available for both individuals and households, this is not the case for those measures such as consumption that are most immediately relevant for assessing living standards. Some goods are consumed privately by each member of the household, but many others are shared, and even for food, the most important nonshared good in developing countries, information about each person's consumption usually cannot be inferred from the data on household purchases of food that are typically observed. Chapter 4 will take up these questions in some detail, and review methods that have been proposed for inferring individual welfare from household-level data, as well as a broader avenue of research that uses household data to draw conclusions about allocations within the household. A prior question is the definition of the household in the data, why it is that some people are grouped together and others not. Broadly defined households, containing servants and distant relatives, will be larger and more likely to have a membership that responds to changes in the economic environment. While this will be inconvenient for some purposes, and will certainly make it difficult to impute living standards, it may still be the unit that is relevant for decisions like migration or the allocation of work.

There is no uniformity in definitions of the household across different surveys, although all are concerned with living together and eating together, and sometimes with the pooling of funds. A range of possibilities is reviewed by Casley and Lury (1981, pp. 186–88), who point out that different criteria are often in conflict, and emphasize that household arrangements are often not constant over time. Many of

the problems are associated with the complex structure of living arrangements in developing countries, and the fact that households are often production as well as consumption units so that a definition that is sensible for one may be inappropriate for the other. When men have several wives, each wife often runs what is effectively a separate household within a larger compound presided over by the husband. Even without polyandry, several generations or the families of siblings may live in a single compound, sometimes eating together and sometimes not, and with the group breaking up and reforming in response to economic conditions. In some countries, there are lineages to which groups of households belong, and the head of the lineage may have power to command labor, to order migration, to tax and reward individuals, and to control communal assets. Even so, members of the lineage will typically live in separate households, which will nevertheless not be the appropriate units for the analysis of at least some decisions.

An example of the consequences of alternative definitions comes from Thailand, where compound living arrangements are common. The National Statistical Office changed its survey practice between the 1975–76 and 1981 surveys, and now counts subunits as separate households. Between the two surveys, average household size fell from 5.5 persons to 4.5 persons, and at least some of the difference can be attributed to the change in procedure. A decision to separate previously pooled households should not affect estimates of average consumption or income per head, but will increase measures of inequality, since the previous single estimate for the pooled household is replaced by multiple estimates for each of the subhouseholds, estimates that are not necessarily the same. Splitting households has the same effect on the distribution of income or consumption as an increase in dispersion with no change in mean, and so must increase measures of inequality (see Kanbur and Haddad 1987 and Chapter 3 below).

Reporting periods

When households are asked to report their income or consumption, a choice has to be made about the reference or reporting period. As with optimal sample design, the ideal reporting period depends on the purpose of the survey. For example, if the object of the exercise is to estimate average consumption over a year, one extreme is to approach a sample of households on January 1 and ask each to recall expenditures for the last year. The other extreme is to divide the sample over the days of the year, and to ask each to report consumption for the previous day. The first method would yield a good picture of each household's consumption, but runs the risk of measurement error because people cannot recall many purchases long after they have been made. The second method is likely to be more economical, because the survey effort is spread over the year, and will give a good estimate of mean consumption over all households. However, unless each household is visited repeatedly, the survey will yield estimates of individual expenditures that, while accurate on average, are only weakly related to the mean or normal expenditures that are appropriate measures of individual standards of living. Nor are short recall periods immune to recall errors, such as "boundary" or "start-up" bias whereby respon-

dents report events that occurred just before the beginning of the reporting period, in an effort to be helpful and to ensure that the enumerator does not miss "relevant" events.

Scott and Amenuvegbe (1990) cite a number of studies showing that reported rates of consumption diminish with the length of the recall period, and their own experiments with households from the Ghanaian Living Standards Survey showed that for 13 frequently purchased items, reported expenditures fell at an average of 2.9 percent for every *day* added to the recall period. They found no evidence of start-up bias; rather their results confirm that recall deteriorates with time, even over a matter of days. If these conclusions are more generally valid, recall periods of even two weeks—as is often the case—will result in downward-biased estimates of consumption.

Even in the absence of reporting errors, so that sample means are unaffected by the choice of recall period, different designs applied to the same underlying population will give different estimates of inequality and of poverty. Many people receive no income on any given day, and many (albeit fewer) will spend nothing on any given day, but neither fact is a real indication that the individual is truly poor, nor that differences between individuals on a given day are indicative of the true extent of inequality. In practice, most surveys adopt sensible designs that tradeoff potential recall bias from long reporting periods against potential variance from short periods. For consumption, frequently purchased items like food have a recall period of between a week and a month, while larger or rarer items, like durable goods, are asked about on an annual recall basis. Even when the recall period is a day, households are revisited every day or two or, when practical, are asked to keep diaries for one or two weeks. Aggregate annual consumption can then be estimated for each household by multiplying monthly food expenditures by twelve and adding the durable and other items. Such a calculation will typically give a useful indication of household consumption, and the average over the households in the survey is likely to be a good estimate of average household consumption in the population.

However, good estimates of means do not imply good estimates of dispersion. For any given household, expenditures will vary from one reporting period to another, so that even in a "survey" that repeatedly interviewed the same household once a week for a year and used the weekly reports to calculate 52 annual estimates, not all would be the same. The measured dispersion of annualized expenditures will contain both intrahousehold and interhousehold components, the former from the within-year dispersion for each household, and the latter from the genuine inequality across households in annual expenditure. Since it is the latter in which we are usually interested, the use of reporting periods shorter than a year will overestimate dispersion. Some of the intrahousehold variation is seasonal, and seasonal patterns can be estimated from the data and used to make corrections. The same is not true for the random nonseasonal variation across weeks or months for each household. Scott (1992) makes calculations for this case and gives a plausible example in which the standard deviation of annual expenditures is overestimated by 36 percent from a survey that gathers consumption data on a monthly basis.

In several of the Living Standards Surveys (LSS), respondents were asked to report expenditures for more than one period. The standard LSS format calls for two visits, roughly two weeks apart, and the interviewer asks how much was spent on each food item "since my last visit." Respondents are also asked in how many months of each year they buy the item, and what they "normally" spend in each of those months. For some nonfoods, households report expenditures both "since the last visit" as well as "in the last year." Results of comparisons are reported in Grosh, Zhao, and Jeancard (1995) for the Living Standards Surveys of Ghana and Jamaica, and in Deaton and Edmonds (1996) for Côte d'Ivoire, Pakistan, and Viet Nam. Estimates of mean expenditure tend to be larger for shorter reporting periods, which is consistent both with recall failure over time, which would bias down the long-period data, but also with boundary effects, which would bias up the short-period data. It is unclear which (or if either) measure is correct. For total expenditures, the ratio of means (short-to-long) are 1.04, 1.08, 1.10, and 1.01 for Côte d'Ivoire, Ghana, Jamaica, and Viet Nam, respectively. There is also evidence, from Côte d'Ivoire and Pakistan, but not from Viet Nam, that when the time between visits is longer, reported expenditures do not increase proportionately, so that the *rate* of expenditure is lower the longer the recall period; once again, this is consistent with progressive amnesia about purchases. Measures of dispersion are also somewhat higher for shorter reporting periods; in Côte d'Ivoire for example, the standard deviations of monthly per capita consumption are 368 and 356 thousand CFAs for short and long periods, respectively. In general, "normal" expenditures on food are usually not very different from those reported "since the last visit," and while the discrepancies in both means and variances are larger for nonfoods, they account for a smaller share of the total budget. Given the other uncertainties associated with defining and measuring consumption—see the next subsection—we should perhaps not be too concerned with the discrepancies that are attributable to differences in reporting periods, at least over the practical range. Of course, such a conclusion does not provide reassurance that the measures of inequality and poverty from the surveys correspond to the measures of inequality and poverty that we should like.

Measuring consumption

If our main concern is to measure living standards, we are often more interested in estimating total consumption that its components. However, some individual items of expenditure are of interest in their own right because their consumption is of direct interest—health care, education, food, especially nutrient-rich foods such as milk—or because the items are subsidized or taxed at differential rates, so that the pattern of demand has implications for public expenditures and revenues. We also need a separate accounting of public goods and their contribution to welfare, and we need to separate expenditures on nondurables from durables, since the latter do not contribute to living standards in the same way as the former. Forecasts of demand patterns are often useful, and are essential for the sort of planning exercises that were once a routine part of development policy. The rate at which consumption

switches from food to manufactures and services, and, in poorer economies, the rate at which a largely vegetarian and cereal-based diet is supplemented with meat, exert a powerful influence on the fraction of the population employed in agriculture and on the type and intensity of agricultural production.

Even when the survey is concerned only with the measurement of living standards, questions about total expenditure are unlikely to provoke accurate responses, and it is necessary to disaggregate to some extent in order to obtain satisfactory estimates. There are often other sources of information about components of consumption—crop surveys and trade data about cereals, for example—so that assessing the reliability of the survey typically requires disaggregation. Traditional household surveys in developing countries have surveyed consumption in great detail, and the Indian NSS, the Indonesian Survei Sosial Ekonomi Nasional (SUSENAS), and many other surveys collect information on around 200 separate food items alone, both in physical quantity and monetary units, together with several dozen more nonfood items. The World Bank's Living Standard Surveys have usually been less detailed, on the grounds that the detail and the data on physical quantities are necessary only for calculating calorie and nutrient intake, but not to obtain accurate estimates of total consumption and living standards.

There is mixed evidence on whether it is possible to obtain accurate estimates of total consumption from a small number of expenditure questions. A test survey in Indonesia (World Bank 1992, Appendix 4.2) subjected 8,000 households to both short and long questionnaires. In the former, the number of food items was reduced from 218 to 15, and the number of nonfood items from 102 to 8. Total measured food expenditures differed little between the questionnaires, either in mean or distribution, although the long questionnaire yielded about 15 percent more nonfood expenditure. But these encouraging results have not been replicated elsewhere. A similar experiment in El Salvador with 72 versus 18 food and 25 versus 6 nonfood items gave ratios (long-to-short) of 1.27 for food and 1.40 overall (Jolliffe and Scott 1995). A 1994 experiment in Jamaica produced similar results, with a long-to-short ratio of 1.26 for both food and nonfood (Statistical Institute and Planning Institute of Jamaica 1996, Appendix III). Although the shorter questionnaires can sometimes lead to dramatic reductions in survey costs and times —in Indonesia from eighty minutes to ten—it seems that such savings come at a cost in terms of accuracy.

The quality of consumption data has been subject to a good deal of debate. Minhas (1988) and Minhas and Kansal (1989) have compared various item totals from the Indian NSS consumption surveys with the independently obtained production-based totals of the amounts of various foods available for human consumption. While the results vary somewhat from food to food, and while it is important not to treat the production figures as necessarily correct, there is typically very close agreement between the two sets of estimates. The survey figures are, if anything, somewhat higher, for example by 4 percent for cereals in 1983. In Britain, the annual FES underestimates total consumption, although as we have already seen, some of the discrepancy is due to the downward bias in alcohol and tobacco expenditures, a part of which reflects the coverage of the survey. In the United States, the

Consumer Expenditure Survey also appears generally to underestimate consumption; again alcohol and tobacco are major offenders, underestimating the national accounts figures by a half and a third, respectively, but there are problems with other categories, and even food expenditures were some 15 percent lower than the estimates from the national income and product accounts (NIPA), with the discrepancy growing over time (see Gieseman 1987).

There are peculiar sampling problems associated with variables whose distribution in the population is extremely positively skewed. Assets—including land—are the most obvious example, but the same is true to a lesser extent of income, consumption, and many of its components. Consider the most extreme case, when one household owns all of the assets in the economy. Then surveys that do not include this household will yield an estimate of mean assets of zero, while those that do will yield an overestimate of the mean, by the ratio of population to sample size. Although the estimate of mean assets over all samples is unbiased, it will usually be zero. More generally, sample means will inherit some of the skewness of the distribution in the population, so that the modal survey estimate will be less than the population mean.

There are two other issues that tend to compromise the quality of consumption data in developing countries. Both are associated with the fact that most agricultural households are producers as well as consumers, and both reflect the difficulty of disentangling production and consumption accounts for people who have no reason to make the distinction. The first problem, which will arise again in Chapter 4, is that wealthy households hire workers, both domestic servants and agricultural workers, and in many cases supply them with food, explicitly or implicitly as part of their wages. Food expenditures for wealthy households will therefore usually include expenditures for items that are not consumed by the immediate family and for large agricultural households the discrepancy can be large.

The second problem relates to consumption of home-produced items, typically food grown or raised on the farm or in kitchen gardens. Such items, often referred to as *autoconsommation*, are properly recorded as both income and consumption, but are often difficult to value, especially in economies—as in much of West Africa —where some markets are not well developed, and where home production and hunting may account for a large share—perhaps more than a half—of total consumption. In cases where prices are available, and where the items being consumed are similar to those that are sold nearby, imputation is not difficult, although there are often difficulties over the choice between buying and seeling prices. Where there are differences in quality, or where the item is rarely sold, some price must be imputed, and the choice is nearly always difficult. In one extreme example, a comprehensive survey in West Africa went so far as to collect data on the amount of water that people fetched from local rivers and ponds and used for cooking and washing. Regarding this as an item of *autoconsommation*, its price was calculated using an algorithm applied to all such items, which was to select the price for a similar traded item in the geographically nearest market. In this case, river water was effectively valued at the current price of *L'Eau Perrier* in the nearest city, thus endowing rural households with immense (but alas illusory) riches.

Quite apart from being aware of the general measurement error that is likely to be introduced when imputations are responsible for a large fraction of total consumption, it is also important to recognize that such errors will be common to measures of both income and consumption, since imputations are added to both. As a result, if we are interested in the relationship between consumption and income, the same measurement error will be present in both dependent and independent variables, so that there is a spurious correlation between the two, something that needs to be taken into account in any analysis.

Measuring income

All of the difficulties of measuring consumption—imputations, recall bias, seasonality, long questionnaires—apply with greater force to the measurement of income, and a host of additional issues arise. Income is often a more sensitive topic than is consumption, especially since the latter is more obvious to friends and neighbors than the former. Accurate estimates of income also require knowledge of assets and their returns, a topic that is always likely to be difficult, and where respondents often have incentives to understate.

Perhaps most important of all is the fact that for the large number of households that are involved in agriculture or in family business, personal and business incomings and outgoings are likely to be confused. Such households do not need the concept of income, so that respondents will not know what is required when asked about profits from farms or own enterprises. The only way to obtain such measures is by imposing an accounting framework on the data, and painstakingly constructing estimates from myriad responses to questions about the specific components that contribute to the total. Even in the industrialized countries, the measurement of self-employed income is notoriously inaccurate; for example, Coder (1991) shows that estimates of nonfarm self-employment income from the March round of the Current Population Survey (CPS) in the United States are 21 percent lower than independent estimates from fiscal sources, while the estimates for farm self-employment income are 66 percent lower. Yet the ratio of the CPS estimate to the tax estimate for wages and salaries is 99.4 percent. A farmer in a developing country (or in the United States for that matter) who buys seeds and food in the same market on the same day has no reason to know that, when computing income, it is only expenditure on the former that should be deducted from his receipts. A street trader selling soft drinks may report that his profits are zero, when the fact is that at the end of each day, after buying food and giving some money to his wife and children, he has just enough left to finance the next day's inventory. Those close to subsistence, whose outgoings are close to incomings, are quite likely to report that income is zero. To get better estimates, the survey must collect detailed data on all transactions, purchases of inputs, sales of outputs, and asset transactions, and do so for the whole range of economic activities for wage earners as well as the self-employed. This is an enormous task, especially in countries where households are large and complex and where people are involved in a wide range of income-generating activities.

The practical and conceptual difficulties of collecting good income data are severe enough to raise doubts about the value of trying; the costs are large and the data may not always be of great value once collected. Apart from some early experiments, the Indian NSS has not attempted to collect income data in their consumer expenditure surveys. Very few households refuse to cooperate with the consumption surveys, and, as we have seen, their estimates of aggregate consumption cross-check with independent estimates. The belief is that the attempt to collect income could compromise this success, and would lead, not only to poor income figures, but also to a deterioration in the quality of the estimates of consumption. That said, the World Bank's experience with the various Living Standards Surveys as well as that of RAND with the Malaysian and Indonesian Family Life Surveys has been that it is possible to collect data on income components—accurately or inaccurately—without any effect on response rates. On the conceptual issues, anyone who has made the calculations necessary to assemble a household income estimate from a detailed integrated survey, such as the Living Standards Survey, with the hundreds of lines of computer code, with the arbitrary imputations, and with allowances for depreciation and appreciation of capital goods and livestock, will inevitably develop a lively skepticism about the behavioral relevance of such totals, even if the calculation is useful as a rough measure of the flow of resources into the household.

Survey-based estimates of income are often substantially less than survey-based estimates of consumption, even when national income estimates show that households as a whole are saving substantial fractions of their incomes, and even in industrialized countries where self-employment is less important and income easier to measure. Although there are often good reasons to doubt the absolute accuracy of the national income figures, the fact that surveys repeatedly show large fractions of poor people dissaving, and apparently doing so consistently, strongly suggests that the surveys underestimate saving. What little we know about the accuracy of the consumption estimates indicates that consumption is more likely to be underestimated than overestimated, so that it seems likely that most survey estimates of income are too low. Some of the underestimation may come from positive skewness in the distribution of income, so that survey estimates of the mean will also be skewed with a mode below the mean (see p. 28 above.) But underestimation of individual incomes is almost certainly important too. While this conclusion is far from well documented as a general characteristic of surveys in developing countries, many statisticians find it plausible, given the conceptual and practical difficulties of measuring income. Discovering more about the discrepancies between national income and survey-based estimates of saving should be given high priority in future research. Most theories of saving relate to individual or family behavior, and yet much of the concern about saving, growth, and macroeconomic performance relates to national aggregates. If household data cannot be matched to national data, it is very difficult to make progress in understanding saving behavior.

Table 1.3 presents data on income, expenditure, and saving from the Socioeconomic Survey of the Whole Kingdom of Thailand in 1986. The left-hand panel groups the households in the survey according to a comprehensive definition of

Table 1.3. Household saving by income and expenditure deciles, Thailand, 1986
(baht per month)

By income decile			*By expenditure decile*		
Decile	*Income*	*Saving*	*Decile*	*Income*	*Saving*
1	690	-683	1	950	49
2	1,096	-737	2	1,409	-6
3	1,396	-741	3	1,768	-23
4	1,724	-775	4	2,119	-73
5	2,102	-746	5	2,523	-145
6	2,589	-593	6	3,005	-221
7	3,231	-503	7	3,649	-103
8	4,205	-358	8	4,579	-379
9	5,900	82	9	6,119	-308
10	13,176	2,761	10	12,283	-1079
All	3,612	-229	All	3,841	-229

Notes: Figures are averages over all households in each decile; there are 10,918 households in the survey. In the left-hand panel, households are grouped by deciles of total income, in the right-hand panel, by deciles of total expenditure.
Source: Author's calculations using the Social Survey of the Whole Kingdom of Thailand, 1986 (see Example 1.1 in the Code Appendix).

household total income, and shows the average saving figures for households in each decile group. According to these estimates, households dissave in total, while the National Income Accounts for Thailand (U.N. 1991) estimate that household saving was 12.2 percent of household income in 1986. Moreover, low-income households dissave more than higher-income households, and the bottom nine deciles each show negative saving on average. Although such behavior is not universal—see, for example, the results for Taiwan (China) in Chapter 6—it is common in surveys in developing countries (see, for example, Visaria and Pal 1980). Such evidence makes it easy to see why early observers of economic development inferred that saving was confined to rich households. However, the right-hand side of the table shows that such an explanation is not correct, at least not in any simple way. Households that are well-off in terms of income are also likely to be well-off in terms of consumption, so that if saving rates are higher for richer households, saving should also rise across consumption deciles. But the date show that the largest amount of dissaving is done by households in the top expenditure decile.

These findings probably owe a good deal to measurement error. If income and consumption are independently measured, at least to some extent, households who overstate their incomes will also, on average, overstate their savings, while households who overstate their consumption will correspondingly understate their saving. The top deciles of income contain a large fraction of households with overstated incomes and will thus show the highest saving rates, with the opposite effect for consumption. Of course, exactly the same story can be told with "transitory in-

come" and "transitory consumption" replacing "measurement error" in income and consumption, respectively, and it is this analogy that lies at the heart of Friedman's (1957) permanent-income theory of consumption. Indeed, since Friedman defined transitory and permanent income as if they were measurement error and the unobserved "true" income, respectively, an explanation in terms of measurement error can always be recast as a permanent-income story. Even so, the results in Table 1.3 are consistent with the importance of measurement error in income, as is the discrepancy between the national accounts and survey results.

Paxson (1992) has argued that the presence of inflation also tends to overstate consumption relative to income given that surveys usually have different reference (reporting) periods for consumption and income. As we have seen, the reference period for consumption varies from item to item, but is often a week or two weeks for food—which may account for two-thirds or more of the budget—while the importance of seasonality in incomes means that reference periods for income are usually a year. Consumption is then denominated in more recent, higher prices than is income, imparting a downward bias to measures of saving. Since inflation in Thailand in 1986 was only 2 percent per annum, the appropriate correction makes little difference to the figures in Table 1.3, though Paxson shows that the corrections for Thailand in the previous 1980–81 survey, when inflation was 16 percent, increase saving by around 7 percent of income.

1.3 The Living Standards Surveys

The Living Standards Measurement Study (LSMS) was begun in the World Bank in 1979 in the last months of the McNamara presidency. The original aim was to develop the World Bank's ability to monitor levels of living, poverty, and inequality in developing countries, to allow more accurate statements about the number of people in poverty around the world, and to permit more useful comparisons between countries. In conjunction with host statistical offices, the project fielded its first surveys in Peru and in Côte d'Ivoire in 1985–86. Since then, there have been several dozen LSMS or LSMS-related surveys. These surveys are different from the typical earlier survey in developing countries, and the experience with them has been influential in shaping current survey practice. It is therefore useful to devote some space to a description of their special features, and to attempt some assessment of what has been learned from the LSMS experience.

A brief history

In the late 1970s, it became clear that it was impossible for the World Bank—or for anyone else—to make well-supported statements about world poverty, especially statements that required internationally comparable data. There was no firm basis assessing such fundamental topics as the extent of poverty in the world, which countries were the poorest, or whether the inequality within and between nations was expanding or contracting. Even within countries, the simplest statements about distributional outcomes were difficult. One particularly important case was that of

Brazil, where there was dispute as to the extent that poor people had benefited from the "economic miracle" of the 1960s. According to an analysis by Fields (1977), the poor had done much better than the nonpoor, but Ahluwalia and others (1980) showed that neither this result, *nor any other useful conclusion* could be supported by the available evidence. While national income data were not above criticism, then as now, statistical offices worldwide were generating useful, credible, and comparable data on average economic performance, but there were no corresponding data on distribution. At the same time, writers in the "basic needs" literature, for example Streeten and others (1981), were arguing for a reevaluation of the relationship between economic growth and poverty. While there was no lack of economists on either side of the issue, the data did not exist to settle what was largely a factual question. Even in India, the motherland of household surveys, evidence on poverty trends was controversial and hotly debated (see Bardhan 1971, Lal 1976, Ahluwalia 1978, and Griffin and Ghose 1979). Because data on consumer expenditures were collected only every five years, and because rural poverty in India is so sensitive to fluctuations in the harvest, it was impossible before the mid-1980s to separate trend from fluctuations, and be sure from the NSS data that poverty rates were indeed falling (see Ahluwalia 1985).

The original aim of the LSMS project was to remedy this situation, by collecting —or at least by helping others to collect—comparable survey data across countries, and so allowing comparisons of poverty and inequality over time and space. In retrospect, it is unclear how such an objective could have been achieved except by the establishment of international standards for surveys that were comparable, for example, to the U.N.'s system of national accounts (SNA); even then, it would have been much more difficult to establish a common set of protocols for estimating dispersion than for estimating means. However, by the time that the first full LSMS survey was ready for implementation in Côte d'Ivoire in 1985, attention had shifted from measurement towards a more ambitious program of gathering data to be used to understand the processes determining welfare at the household level. Although either set of objectives would have required a multipurpose household income and expenditure survey of some kind, the new aims were best served by an intensive, integrated survey in which each household was asked about every aspect of its economic and domestic activity. While the expense of such a design necessitated a smaller sample size, as well as less detail on individual topics than had been the case in traditional single-purpose surveys, such as agricultural or consumption surveys, these were accepted as the costs of the opportunity to model household behavior as a whole.

The shift in emphasis owed much to "Chicago" views of household and farm behavior, particularly Schultz's arguments that households respond rationally and purposively to prices and incentives and to the development by Becker of the "new household economics." These views first influenced survey practice in RAND's 1976–77 Malaysian Family Life Survey, the experience of which helped shape the first Living Standards Surveys. The latter have in turn influenced later RAND surveys, particularly the 1988–89 second Malaysian Family Life Survey and the 1993 Indonesian Family Life Survey. Behind these designs rests the belief that policy

advances should rest on an enhanced empirical understanding of how such households respond to their economic and physical environments, and on the role of government policy in shaping those environments. Although such a perspective was different from the original one, the practical consequences were confined to the tradeoff between sample size and the amount of information from each individual. Nothing in the new design would prevent the data being used for its original purposes and indeed, in recent years, one of the main uses of LSMS surveys has once again been for the measurement of poverty.

Another theme in the original design was an emphasis on collecting at least some panel data. As shown above, a panel design is often an efficient way to collect information on changes over time, which was one of the aims of the LSMS project. Panel data also seemed to be well suited to the documentation of the losses and gains from economic development, or from structural adjustment. In the late 1970s and early 1980s, there was also a great deal of interest in academic circles in the econometric possibilities associated with panel data, so that the collection of such data, which was rare in developing countries, was an exciting and promising new endeavor.

The Ivorian Living Standards Survey collected data in 1985, 1986, 1987, and 1988, with an intended sample size of 1,600 households. There were three panels of 800 households each, which ran for two years each in 1985–86, 1986–87, and 1987–88; no household was retained for longer than two years, but in principle, all households would be interviewed at two visits a year apart, except for half of those in the first and last years. A larger (5,120-household) single-year survey was carried out in Peru in 1985–86, followed by a two-year panel survey with 3,200 households in Ghana in 1987–88 and 1988–89. Since then there have been LSMS or LSMS-related surveys in Mauritania (1988), Morocco (1990–91), Pakistan (1991), Venezuela (1991–93), Jamaica (1988 to date), Bolivia (1989 to date), the Kagera region of Tanzania (1991–93), Peru (1990, 1991, and 1994), Russia (1991–92), South Africa (1993), Viet Nam (1993), Kyrgyz Republic (1993), Nicaragua (1993), Guyana (1993), Ecuador (1994), Romania (1994–95), Bulgaria (1995), and the Hebie and Liaoning provinces of China (1995). At the time of writing, LSMS surveys are in the design or implementation stage for Brazil, Kazakhstan, Mongolia, Nepal, Paraguay, Tunisia, Turkmenistan, and Uzbekistan. These surveys are far from identical, although all have common design elements as described below. But the surveys have been adapted to different needs in different countries, so that, for example, the Jamaican survey is a compromise between a full LSMS survey and a previously existing and long-running labor force survey, while the survey in Kagera is concerned with monitoring economic responses to the AIDS epidemic. For further details, see Grosh and Glewwe (1995), from which the information in this paragraph is culled, or the LSMS homepage on the World Wide Web.

Design features of LSMS surveys

The design and implementation of the Ivorian survey is described in Ainsworth and Muñoz (1986) and their description remains a good account of a prototypical LSMS

survey. The implementation manual by Grosh and Muñoz (1995) provides a fuller and more recent account, incorporating the lessons of past experience, and is essential reading for anyone designing an LSMS or LSMS-related survey.

There are three separate questionnaires in the "standard" LSMS survey; a household questionnaire, a community questionnaire, and a price questionnaire. The first is long by previous standards, and comprises (up to) seventeen sections or modules, some of which are technical, such as the section that identifies suitable respondents for subsequent modules, or the section that links individuals across years for panel households. Most modules are substantive and cover a long list of topics: household composition, housing and its characteristics, education, health, economic activities and time use, migration, agricultural and pastoral activities, nonfarm self-employment activities, food expenditures, durable-goods expenditures and inventories, fertility, other sources of income including remittances, saving, assets, and credit markets, and anthropometric measurement of household members. The community questionnaire, which is sometimes used only in rural areas, gathers data from knowledgeable local people (such as chiefs, village headmen or elders, medical personnel, or teachers) about local demographics (population, ethnicity, religion, and migration), and about local economic and service infrastructure, such as transportation, marketing, extension services, primary and secondary schools, and health and hospital facilities. The price questionnaire, administered country wide, was collected by enumerators visiting local markets and observing prices, mostly of foods.

The surveys are designed to produce high-quality data, and to deliver the results quickly; to this end, use is made of microcomputers both for the design of questionnaires and for data entry and editing. The use of computer software to turn questions into a printed questionnaire, with appropriate pagination and skip patterns, not only cuts down on error, but also permits rapid redesign and reprinting after field tests. Many responses are precoded on the questionnaire, so that there is no coding by keyboard operators at the data entry stage. In each round, households are visited twice with an interval of two weeks, and roughly half the questionnaire administered at each visit. Between the two visits, the first data are entered into the computers and are automatically subject to editing and consistency checks by the software as they are entered. Such procedures not only minimize data entry errors, but permit the enumerators to correct some response errors during the second visit. They also mean that much less time is required to edit the data after the survey, so that instead of the several years that are common in some countries, the preliminary data and tabulations are available in three to six months after the completion of the fieldwork.

What have we learned?

It is too early for any final verdict on the contribution of LSMS surveys to data collection in developing countries in general, let alone to their ultimate aim, of improving policymaking and the understanding of economic behavior in developing countries. At the time of writing it is only eleven years since the first household

was interviewed. New data are continuously becoming available and new uses in policy and analysis are being developed. However, the LSMS surveys have taught us a great deal that we did not know in 1980, and the experience has certainly changed the way household data are—or at least ought to be—collected in developing countries.

The experiments with microcomputers have been successful. Computer-assisted questionnaire design works and the procedures for precoding, software-controlled data entry, and checking enhance data quality and speed delivery. Although local conditions typically preclude using computers during interviews, so that there are limits on the use of the computer-aided interview techniques that are rapidly developing in the United States and elsewhere, it is feasible to install microcomputers at local survey headquarters, and these are used to enter and check the data within a few days of capture. Given the cost and robustness of microcomputers, there is every reason for their use to be universal in household survey practice around the world. There is no good reason—except for entrenched bureaucracy and vested interests—why survey results and tabulations should be delayed until years after fieldwork is completed.

We also know now that long and complex household questionnaires are practical. Because there are two visits separated by two weeks, and because different members of the household are interviewed for different parts of the questionnaire, no one person is subjected to an impossibly long interview. At the beginning of the LSMS project, opinions were sharply divided on the feasibility of long questionnaires, with experienced surveyors arguing both for and against. In practice, and although each round of the LSMS survey was two to three hours long, depending on the number of household members, there were no refusals based on length, and no appearance of declining cooperation as the interviews progressed. Of course, at any given cost, there is a tradeoff between length of the questionnaires and the number of households that can be covered in the sample. The fact that LSMS surveys favor the former has meant that survey results are less useful than traditional, larger surveys for disaggregated measurement of living standards, by region, occupation, or other target group. As a result, and as interest has turned back to the use of survey data to assess poverty, the later surveys have tended to be simpler and to have larger sample sizes, and there has been some reversion towards the original aims of measurement rather than analysis.

While it is difficult to assess data quality without some sort of controlled experiment, the feedback from users has usually been positive, and there is no evidence to suggest that the timeliness of the data has been bought at the price of lower quality compared with data sets where the cleaning and editing process has taken a great deal longer. To be sure, there are difficulties with the LSMS data, but these appear to be the sort of problems that can and do arise in any survey, irrespective of design. The community and price questionnaires are elements that are more experimental in the LSMS design and there currently appears to be little hard evidence on their usefulness. The community questionnaire can be regarded as an efficient way of collecting information that could, in principle, have been collected from individual households. The data on the provision of services has been routine-

ly combined with the household data in various analyses, for example of the determinants of access to education and health facilities. One difficulty lies in the concept of a community. The simplest idea is of a village, whose inhabitants share common health, educational, and other facilities, and who buy and sell goods in the same markets. But communities may not conform to this model; people may belong to different "communities" for different purposes, and in some parts of the world, there are no well-defined villages at all. Nor is there any guarantee that the primary sampling units or clusters in the survey necessarily correspond to units that have any unified social or administrative structure. In consequence, even when the community questionnaire yields data on schools, clinics, or transportation, we do not always have a clear delineation of the population served by those facilities, or on its relationship to the survey households in the cluster.

Information about prices is not easy to collect. Enumerators are given a list of well-defined items, and are required to price at three different sites in local markets. For obvious reasons, an enumerator is not given money to make actual purchases, but instead approaches the seller, explains that he or she is conducting a survey (which has nothing to do with taxes or law enforcement), and asks the price of an item. There is no haggling, but the enumerator is supplied with scales and asks the seller's permission to weigh the potential purchase. While it is easy to see the problems that might accompany such a procedure, it is harder to devise alternatives. "Market price" is a concept that is a good deal more complex in an African market than in an American supermarket or an economics textbook. Different people pay different amounts, there are quantity discounts, and many foods are presented for sale in discrete bundles (half a dozen yams, or a bunch of carrots) whose weights or volumes may vary even at the same price. There are also wide geographical differences in the range of items in local markets, so that it is difficult to collect the sort of price data that will permit reliable calculation of the cost-of-living indexes required to compare real living standards across different areas of the country, and between rural and urban households. Even so, in those LSMS surveys (for example Pakistan and Viet Nam) where consumers reported both expenditures and physical quantities, the unit values from these reports are well correlated across space with the prices from the price questionnaire (see Deaton and Edmonds 1996). An alternative procedure is recommended by Grootaert (1993) who suggests collecting prices using the standard price surveys that usually exist to gather cost-of-living data, although such a solution sacrifices the exact match between survey households and the prices that they face.

The experience of collecting panel data has been somewhat mixed. The original emphasis has been muted over time, and few of the recent surveys have been explicitly designed as panels. While the LSMS surveys should not be judged on their success (or lack of it) in collecting panel data, the experience has been useful, particularly given the rarity of such data from developing countries. Some critics had suggested that it would be difficult to locate many households for the second annual visit and these worst fears were not realized. Even so, there is a good deal of migration of household members, and there are occasional refusals, so that actual panels are smaller than the design. In Côte d'Ivoire, where the design called

for 800 panel households out of a total sample of 1,600, there were 793 "panel-designate" households from whom valid data were collected in 1985; 730 of them (92 percent) provided valid data in 1986. In the second panel in 1986–87, 693 (87 percent) of the 800 designated households provided data, and in 1987–88, there were 701 (88 percent) out of 800. Since attrition is generally at its worst in the first year, and since the LSMS surveys make no attempt to follow households that have moved, these fractions are impressively high, comparing well with the attrition rates in the PSID.

Perhaps less satisfactory is the occasional difficulty of determining whether two so-called panel households in two years are in fact the same household. When panel households could not be located in the second year, replacement households were selected, and these were not always labeled as clearly as they should have been. The section of the questionnaire that links panel households and their members is a great help in this regard, and is a testament to the fact that the designers clearly anticipated the problems that arise with panel households, but there are cases where examination of this section, and the difficulty of matching individuals, raises questions about whether what is labeled as a panel household is not in fact a replacement. The underlying fear is that, in the absence of detailed supervision, enumerators may too readily substitute new households for the intended panelists, and that the fact that they have done so may not be easily detectable after the event. The "linking" section of the questionnaire is extremely important, and has to be a focus in any satisfactory panel survey. Linking is best done by having a list of the names of all household members, which can conflict with the usual promises of anonymity.

There are also questions about the general value of the data obtained from two-year rolling panels as in the Ghanaian and Ivorian surveys. One issue is measurement error, particularly in income and, to a lesser extent, in consumption. Given that we want to measure changes in levels of living at the individual level, and given that in much of Africa economic progress has been slow at best, measured changes over a single year will be dominated by a combination of measurement error and the normal fluctuations of agricultural income, neither of which is of primary interest. It is probably also true that attrition is at its worst between the first and second year of a panel, so that a two-year design suffers a larger proportional loss per household year than would a longer panel. Longer panels also offer greater opportunities for assessing and controlling for measurement error (see, for example, Griliches and Hausman 1986 and Pischke 1995).

What would be more useful is one of two things; first, estimates of change over much longer periods, five or ten years say, and second, estimates of change over periods when there have been major policy changes, such as those induced by structural adjustment programs, or clearly defined outside shocks that have affected living standards. Obtaining either of these can be a matter of good fortune, of having a survey in place at just the right time, but will be guaranteed only when surveys are organized on a continuing basis by a permanent local survey organization. As I have already noted, it is also sometimes possible to revisit households from an earlier survey even when there was no original intention of constructing

a panel. Such revisits can allow long observation periods without having to maintain a permanent survey, and can be designed on an ad hoc basis to examine some event of interest (see Bevan, Collier, and Gunning 1989, who looked at the consequences of booms in coffee prices in Kenya and Tanzania in the mid-1980s, or Glewwe and Hall 1995, who looked at the effects of macroeconomic shocks in Peru). Although the Ivorian and Ghanaian surveys were originally intended to be permanent, data collection has ceased in both. The timing was particularly unfortunate in Côte d'Ivoire because data collection ceased just before the decline in world prices forced the government to cut the procurement prices of cocoa and coffee, which are the major sources of agricultural income in much of the country. Studying the effects of this policy change, which terminated a thirty-year regime of approximately constant real prices, could have produced important insights about policy and welfare but the opportunity was lost. In general, it is unclear that, except by chance, there are major benefits to be expected from gathering short-term panel data in slowly growing or stagnant economies.

The usefulness of the LSMS data for policymaking and research is another question on which a verdict is premature, although there is certainly a strong demand by the countries themselves for such surveys. For several of the LSMS countries, the surveys have replaced an almost total absence of information with at least some information, and the tabulations are routinely used in policy operations, project evaluation, and poverty assessment inside the World Bank and in the countries themselves. The surveys have also produced large amounts of microeconomic data that are potentially available for the sort of research described in this book. The LSMS group within the World Bank has also produced an impressive (and impressively long) list of working papers, most of which are concerned with policy-related empirical analysis of the data. However, there is less evidence that the data are being used to capacity in the countries from which they originally came.

One of the aims of the LSMS project, from its first inception, was to provide an easily accessible data base for the analysis of policy on a daily basis. While no one supposed that cabinet ministers would sit in front of terminals displaying the LSMS data, there was hope that their assistants and advisors would, and that tabulations or graphics could be produced in hours, not months or years (or not at all) as is often the case. Questions about whether food subsidies are reaching their intended recipients, or who would be hurt by an increase in fertilizer prices, are examples of the sort of questions that arise regularly among policymakers and their advisors, and that can be quantified and clarified by survey data (see Chapter 3). Such distributional analyses are a regular feature of policymaking in the United States or Britain, for example, but for whatever reason—computing facilities, software, or personnel— similar analyses are rare in developing countries. Even so, there are signs of progress. Most notable is the case of Jamaica, where there was an unusually high degree of (powerful) local interest from the start, and where the LSMS data have been used in a wide range of domestic policy exercises, on poverty assessment, on the effectiveness of food stamps, and on health questions. The data are even used as part of the standard training in quantitative methods at the University of the West Indies (Grosh and Glewwe 1995).

In recent years there has also been a great deal of progress in the archiving of the data and in making them available to the research community, for example via the LSMS page on the World Wide Web. While there are real problems of ownership and confidentiality with all household survey data, arrangements can be—and in an increasing number of cases have been—worked out to give access to the individual household records (again see the LSMS homepage for up-to-date information). Finally, a note about survey costs. The Implementation Manual (Grosh and Muñoz 1996, Table 8.1) presents total costs for eight actual surveys. These vary from $78 per household for a 2,000-household survey in Jamaica to over $700 per household for a 4,480-household survey in Brazil; in several other cases, the per household cost lies between $150 and $250. Much of the variation is explained by whether or not vehicles had to be purchased. The high estimates are inflated by these costs—only a fraction of which are directly attributable to the one-year survey—but are understated by amount of technical assistance provided by the World Bank, which is not included in these estimates.

1.4 Descriptive statistics from survey data

One of the first calls on survey data is to calculate descriptive statistics, often for a survey report containing a standard set of tables. In this section, I discuss some of the issues that arise in calculating these statistics, and in particular how we make sure that statistics describe the *population* rather than the particular *sample* that is available for analysis. To do so, we need to know how the sample was designed and to understand its relationship to the population of interest. Different sample designs require the data to be processed in different ways to estimate the same magnitude. I present some useful formulas for calculating the standard error of estimated means, again taking the sample design into account. These formulas are useful in themselves, as well as for the descriptive material that appears later in the book, particularly in Chapter 3 on the measurement of poverty. They also provide a starting point for the analytic and econometric material in Chapter 2.

Common examples of descriptive statistics are measures of central tendency, such as means and medians, and of dispersion, such as variances and interpercentile ranges. Living standards are often measured by the means of income or consumption, inequality by their dispersion, and poverty by the fraction of the population whose income or consumption is below the poverty line. The quantities that are to be summarized are sometimes continuous, as with income or consumption, and sometimes discrete, as in poverty where the basic data are indicators of whether or not a household is in poverty. While the estimation of means and standard errors is familiar, it is worth recording the formulas that take account of differential weights (inflation factors) and that allow for stratification and clustering.

Finite populations and superpopulations

To make inferences using survey data we need a framework for thinking about how the data were generated, which means thinking about the population from which

the data came and about how data collection induces randomness into our sample. There are two different approaches. In the first, which is standard among survey statisticians, the population is a finite one—for example all households in Côte d'Ivoire in calendar year 1985—and the sample households are randomly selected from that population just as, in the classic textbook example, balls are drawn from an urn. The survey data are random because replication of the survey would generate different samples so that the variability of an estimate, such as the mean of household income, is assessed by thinking about how it would vary from one sample to another. The quantity of interest—in the example the average reported household income of Ivorian households in 1985—is a fixed number that could be measured with perfect accuracy from a census, which is when the sample coincides with the population. No assumptions are made about the distribution of income in the population; what is being estimated is simply the average income in the population in the survey year, not the parameter of a distribution.

In the second approach, we are less interested in the actual population in the survey year, regarding it as only one of many possible populations that might have existed. The actual population is itself regarded as a sample from all possible such populations, the infinite *superpopulation.* The focus of attention is instead the statistical law or economic process that generated income in the superpopulation, in populations like the one under study, and the mean is of interest less for itself than as a parameter or a characteristic of that law or process.

The finite-population or survey-statistical approach is typically associated with *description,* while the superpopulation approach is associated with *modeling.* The distinction is made by Groves (1989, esp. ch. 6) and will be useful in this book in which I shall be concerned with both. The distinction between description and modeling is often of no more than philosophical interest—there is no dispute about the appropriate formula for a mean or standard deviation. But when we come to the econometric analysis in Chapter 2, where the tradition has been of modeling, there are sometimes sharp differences in the recommended calculation of apparently identical concepts. Recent trends in econometric practice have tended to emphasize description at the expense of modeling—an emphasis shared by this book—so that the traditional gulf between the practice of survey statisticians and econometricians is narrowing. In consequence, I shall sometimes follow one approach, and sometimes the other, whichever seems more appropriate to the problem at hand.

The example of the mean can be used to illustrate the two approaches and provides an opportunity to record the basic formulas. Suppose that we have a simple random sample of n observations from a population of size N, and that the quantity of interest is x, with observations x_i, i running from 1 to n. The sample mean $\bar{x}$ is the obvious estimator of the population mean, where

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i. \tag{1.3}$$

The sample mean, (1.3), is a random variable that will vary from one sample to another from a given population and, in the superpopulation approach, from one population to another.

Suppose that we want to know the expectation of $\bar{x}$ over the different surveys. This seems like a complicated matter for the survey statistician, because there are a large number of different ways in which n objects can be selected from a population of N objects, and it is necessary to calculate the mean for each and its associated probability of occurrence. A simple shortcut (see Cochrane, 1977, p. 28) is to rewrite (1.3) as

$$\bar{x} = n^{-1} \sum_{i=1}^{N} a_i x_i \tag{1.4}$$

where the sum now runs over the whole population, to N rather than n, and a_i is a random variable that indicates whether i is in the sample, taking the value 1 if so, and 0 otherwise. The x's on the right-hand side of (1.4) are no longer random variables, but simply the fixed x's in the (finite) population. Hence, when we take expectations of (1.4), we need only take expectations of the a's. Since we have a simple random sample, each i has an equal probability of being included, and that probability is simply the ratio of n to N. The expectation of each a_i is therefore 1 times the probability of its being 1, which is n/N, plus 0 times the probability of its being 0, which is $(1 - n/N)$, a total of n/N. We then have

$$E(\bar{x}) = n^{-1} \sum_{i=1}^{N} (n/N) x_i = N^{-1} \sum_{i=1}^{N} x_i = \bar{X} \tag{1.5}$$

where $\bar{X}$ is the population average in which we are interested.

A superpopulation approach makes more assumptions, but is in some ways more straightforward. It might postulate that, in the superpopulation, x is distributed with mean μ, a parameter that is the same for all households. We then have immediately, from (1.3),

$$E(\bar{x}) = n^{-1} \sum_{i=1}^{n} \mu = \mu \tag{1.6}$$

also as desired. The finite-population approach is more general, in that it makes no assumptions about the homogeneity of the observations in the sample, but it is also more limited, in that it is specifically concerned with one population only, and makes no claim to generality beyond that population.

The technical note at the end of this subsection shows that the variance of $\bar{x}$ is

$$V(\bar{x}) = \frac{1-f}{n} S^2 \tag{1.7}$$

where S^2 is given by

$$S^2 = (N-1)^{-1} \sum_{i=1}^{N} (x_i - \bar{X})^2 \tag{1.8}$$

and can be thought of as the population variance, and $1-f$ is the "finite-population correction" (*fpc*),

$$1-f = (N-n)/N. \tag{1.9}$$

Except in the unusual situation where the sample is a large fraction of the population, the *fpc* is close enough to unity for the factor $1-f$ in (1.9) to be ignored. Indeed, in this book I shall typically assume that this is the case; sampling texts are more careful, and can be consulted for the more complex formulas where necessary.

The superpopulation approach to the variance would postulate that each x_i is independently and identically distributed with mean μ and variance σ^2, so that from (1.3),

$$V(\bar{x}) = E(\bar{x}-\mu)^2 = n^{-1}\sigma^2. \tag{1.10}$$

Since both σ^2 and S^2 are estimated from the sample variance

$$\hat{s}^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.11}$$

and provided we ignore the *fpc*, there is no operational difference between the two approaches. Both use the same estimate of the mean, and both estimate its variance using the formula

$$\hat{v}(\bar{x}) = n^{-1}\hat{s}^2 = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{1.12}$$

Books on sampling often give separate treatment to the estimation of means and the estimation of proportions. However, a proportion is simply the mean of a binary (0,1) indicator that tells us whether the observation does or does not possess the attribute of interest. In consequence, the formulas above—as well as those in the next subsections—can also be used for estimating proportions and their sampling variability. To see how this works and to link up with the analysis of poverty in Chapter 3, suppose that x_i is 1 if household i is in poverty, and is 0 otherwise. The estimate of the proportion of households in poverty is then the mean of x

$$\hat{p} = n^{-1}\sum_{i=1}^{n}x_i = n_1/n \tag{1.13}$$

where n_1 is the number of households with $x_i = 1$. If we have a simple random sample, the estimated variance of $\hat{p}$ is given by (1.12), which since x_i can only take on the two values 0 or 1, takes the simple form

$$\hat{v}(\hat{p}) = n^{-1}(n-1)^{-1}\left[\sum_{i=1}^{n_1}(1-\hat{p})^2 + \sum_{i=n_1+1}^{n}\hat{p}^2\right] = (n-1)^{-1}\hat{p}(1-\hat{p}). \tag{1.14}$$

Formula (1.14) is useful because it is simple to remember and can be calculated on the back of an envelope. Even so, the same answer is given using the standard formulas and treating the x's as if they were continuous.

*Technical note: the sampling variance of the mean

I follow the derivation in Cochrane (1977, p. 29) which starts from (1.4) and from its implication that

$$V(\bar{x}) = n^{-2}\left[\sum_{i=1}^{N} x_i^2 \operatorname{var}(a_i) + 2\sum_{i=1}^{N}\sum_{j<i}^{N} x_i x_j \operatorname{cov}(a_i a_j)\right]. \tag{1.15}$$

Because a_i is binomial with parameter (n/N), its variance is $(n/N)(1-n/N)$. The random variable $a_i a_j$ is either 1, if both i and j are in the sample, or 0, if not. Since the sample is drawn without replacement, the probability of the former is (n/N) multiplied by $(n-1)/(N-1)$. Hence

$$\begin{aligned}\operatorname{cov}(a_i a_j) &= E(a_i a_j) - E(a_i)E(a_j) \\ &= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n}{N(N-1)}\left(1-\frac{n}{N}\right).\end{aligned} \tag{1.16}$$

If the variance and covariance formulas are substituted into (1.15), and rearranged, we obtain (1.7) and (1.8).

Using weights or inflation factors

In most surveys different households have different probabilities of being selected into the sample. Depending on the purpose of the survey, some types of households are overrepresented relative to others, either deliberately as part of the design, or accidentally, for example because of differential response. In both cases, if the different types of households are different, sample means will be biased estimators of population means. To undo this bias, the sample data are "reweighted" to make them representative of the population. In this subsection, I discuss some of the reasons for different probabilities of selection, and the procedures that can be used to calculate population statistics and to assess sampling variability.

Suppose that each of the N households in the population is assigned a sampling probability π_i. A sample of size n is chosen, and I assume that the selection is done with replacement, so that in principle a given household can appear more than once. Although samples are almost never selected this way in practice, the difference between sampling with and without replacement is only important when the sample size is large relative to the population. Pretending that the sample is drawn with replacement is akin to ignoring finite-population corrections, and has the advantage of simpler formulas and derivations. Note that π_i is not the probability that i is in the sample, but the probability that i is selected at each draw, the sample being constructed from n such identical draws. Sample households with low values of π_i have a low ex ante probability of being selected into the sample, and such households are underrepresented relative to those with high ex ante probabilities. In order to correct this imbalance between sample and population, the observations need to be reweighted, weighting up those that are underrepresented and weighting down those that are overrepresented.

The weights that we need are inversely proportional to π_i; in particular, define for each household the weight w_i

$$w_i = (n\pi_i)^{-1}. \tag{1.17}$$

For a simple random sample with replacement, each household's probability of selection at each trial is $1/N$ so that, in this case, the weights w_i are the same for all observations and equal to N/n, which is the "inflation factor" that blows up the sample to the population. When the probabilities differ, the quantity $n\pi_i$ is the expected number of times that household i shows up in the survey. When the sample is small relative to the population, so that the probability of a household appearing more than once is small, $n\,\pi_i$ is also approximately equal to the probability of i being in the sample. As a result, w_i in (1.17) is approximately equal to the number of population households represented by the sample household i and can therefore be thought of as the household-specific inflation factor.

Consider first the sum of the weights which, since each is a household inflation factor, might be thought to be an estimate of the population size N; hence the notation

$$\hat{N} = \sum_{i=1}^{n} w_i. \tag{1.18}$$

Define the random variable t_i as the number of times that household i shows up in the sample; this will usually take the values 1 or 0 but, since sampling is with replacement, could in principle be larger. Its expected value is $n\pi_i$, the number of trials multiplied by the probability of success at each. The sum of weights in (1.18) can then be rewritten as

$$\hat{N} = \sum_{i=1}^{N} t_i w_i \tag{1.19}$$

with the sum running from 1 to N. Taking expectations,

$$E(\hat{N}) = \sum_{i=1}^{N} E(t_i) w_i = \sum_{i=1}^{N} n\pi_i w_i = N \tag{1.20}$$

so that the sum of the weights is an unbiased estimator of the population size.

Suppose that x_i is the quantity of interest reported by household i. We estimate the *total* of x in the population by multiplying each x_i by its weight w_i and adding up, so that

$$\hat{X}_{tot} = \sum_{i=1}^{n} w_i x_i. \tag{1.21}$$

By a precisely analogous argument to that for $\hat{N}$, we have

$$E(\hat{X}_{tot}) = \sum_{i=1}^{N} E(t_i) w_i x_i = \sum_{i=1}^{N} x_i = X_{tot} \tag{1.22}$$

so that $\hat{X}_{tot}$ is unbiased for the population total. The sampling variance of (1.21) is

$$V(\hat{X}_{tot}) = \frac{1}{n} \sum_{i=1}^{N} \pi_i \left(\frac{x_i}{\pi_i} - X_{tot} \right)^2 = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{x_i^2}{\pi_i} - X_{tot}^2 \right) \tag{1.23}$$

(see the technical note on p. 49 below and Cochrane 1977, pp. 252–54). An unbiased estimate of (1.23) can be obtained from the sample by defining $z_i = w_i x_i$ and using the formula

$$\hat{v}(\hat{X}_{tot}) = \frac{n}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2. \tag{1.24}$$

The sampling variance (1.23) can be used to give a formal answer to the question of why different probabilities can enhance efficiency, and to tell us what the optimal probabilities should be. In particular, it is a simple matter to show that (1.23) is minimized subject to the constraint that the π's add to 1 by selecting the π's to be proportional to the x's. This is sampling with *probability proportional to size (p.p.s.)*; larger values of x contribute more to the mean so that efficiency is enhanced when larger values are overrepresented. Of course, if we knew the x's, there would be no need to sample, so that in practice we can use only approximate *p.p.s.*, in which the π's are set proportional to some other variable that is thought to be correlated with x and that is known prior to sampling.

We are often interested, not in the total of x, but its mean, $\bar{X}$, which can be estimated from the ratio of the estimated total (1.21) to the estimated population, (1.19). This is the *probability-weighted mean*

$$\bar{x}_w = \sum_{i=1}^{n} w_i x_i \Big/ \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} v_i x_i \tag{1.25}$$

where the v_i are the w_i from (1.17) normalized to sum to unity,

$$v_i = w_i \Big/ \sum_{k=1}^{n} w_k. \tag{1.26}$$

A simple example of the effects of weighting is given in Table 1.4, which presents the weighted and unweighted means and medians—see (1.30) below for the definition of a weighted median—for total household expenditure by race in the South African Living Standards Survey. These calculations use the weights summarized in Table 1.1 and discussed on page 16 above. For the Blacks, Coloreds,

Table 1.4. Household total expenditures, weighted and unweighted means, South Africa, 1993
(rand per month)

	Means		*Medians*	
Race	*Weighted*	*Unweighted*	*Weighted*	*Unweighted*
Blacks	1,053	1,045	806	803
Coloreds	1,783	1,790	1,527	1,547
Asians	3,202	3,185	2,533	2,533
Whites	4,610	4,621	4,085	4,083
All races	1,809	1,715	1,071	1,029

Source: Author's calculations using the South African Living Standards Survey, 1993.

and Asians, the weights do not vary much within each group, so that there is little difference between the weighted and unweighted estimates. For Whites, there is more within-group variance in weights, but it has little effect on either estimate because there is little or no correlation between the weights and the level of income within the group. However, as we saw in Table 1.1, lower participation by Whites made the weights higher for Whites than for the other groups, so that when we calculate estimates for the whole country, their higher incomes result in weighted estimates that are larger than the unweighted estimates. This result illustrates the general point that, whenever there is an association between the sampling probabilities and the quantity being measured, unweighted estimates are biased.

Because $\bar{x}_w$ is the ratio of two random variables, it is not an unbiased estimator. However, because the variances of its numerator and denominator, (1.21) and (1.19), both converge to zero as n tends to infinity, it will converge to the population mean. The sampling variance of $\bar{x}_w$ can be evaluated using standard approximation techniques for ratio estimators; in the technical note on p. 49 below, I sketch the argument that leads to

$$V(\bar{x}_w) \approx N^{-2} \sum_{i=1}^{N} w_i (x_i - \bar{X})^2 \tag{1.27}$$

which can be estimated from the sample data using

$$\hat{v}(\bar{x}_w) = \frac{n}{n-1} \sum_{i=1}^{n} v_i^2 (x_i - \bar{x}_w)^2. \tag{1.28}$$

Note that the estimated variance will usually need to be specially coded. In particular and except for simple random sampling, (1.28) is *not* equal to the sample estimate of the population variance—equation (1.29) below—divided by the sample size (compare (1.11) and (1.12)). Formula (1.27) can also be used to find the probabilities that maximize the precision of the probability-weighted mean (1.25). Recalling that $w_i = (n\pi_i)^{-1}$ and choosing the π's to minimize (1.27) subject to the constraint that their sum be 1, it is easily shown that the optimal selection probabilities should be proportional to the absolute value of the deviation from the mean $|x_i - \bar{X}|$. It is information on the exceptional cases that adds most to the precision of the estimated mean.

It is worth noting that the probability-weighted mean (1.25) is not the only possible estimate. In particular, if the population size N is known, an estimate of the mean can be obtained by dividing $\hat{X}_{tot}$ by N. But there are a number of reasons why the weighted mean is frequently more useful. First, the population size is often not known, but is estimated from the survey itself, for example by randomly selecting a set of villages, enumerating all households in each, and using the totals to estimate the number of households in the population. Second, when we come to use the survey data to calculate means or other statistics, some data are unusable, because they are missing, because of transcription errors, or because they take on clearly implausible values. There is then little option but to average over the "good" observations, renormalizing the weights to sum to unity. Third, in many applications we are not interested in means per household, but in means per person.

We want to know the fraction of *people* in poverty, not the fraction of *households* in poverty. Or we may want to know the fraction of elderly, or women, or children who have some characteristic. In such cases, we weight the data, not by the number of *households* represented by the sample household, but by the number of *people* it represents. To get statistics about persons in the population, the quantity surveyed, say household per capita consumption, should be weighted not by the w_i themselves, but by the w_i multiplied by the number of people in household i. The total of these weights is (in expectation) the number of people in the population, not the number of households; as before, this population total may or may not be known in advance, but we often encounter cases where the relevant population total is estimated by summing weights from the survey estimate. Fourth and finally, we often want to calculate means for subgroups, and to do so in a way that is representative of the relevant subpopulations. Once again, $\bar{x}_w$ is the relevant estimator unless we know the size of each subpopulation.

The weights can be used to estimate other population statistics analogously to the mean in (1.25). For example, the population variance S^2 is estimated from the sample by weighting the individual squared deviations from the mean so that

$$\hat{s}_w^2 = \frac{n}{n-1}\sum_{i=1}^{n} v_i(x_i - \bar{x}_w)^2 = \frac{n}{n-1}\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2 / \sum_{i=1}^{n} w_i \tag{1.29}$$

where the adjustment factor $n/(n-1)$ is of no practical significance, but matches (1.29) to the standard unbiased estimator (1.11) in the case where $w_i = N/n$. Expression (1.29) would be used, for example, when using the variance of income or of the logarithm of income to measure inequality. Weights must also be used when ranking households, for example when calculating medians, quartiles, or other percentiles in the population. In the sample, median household income (for example) is that level of income below which (and above which) lie half the sample observations. When estimating the median for the population, we must find instead the level of income such that, when we take all sample households with lower income, the sum of their weights is half the total sum of weights. Formally, the median $\tilde{x}_w$ is defined by

$$\sum_{i=1}^{n} 1(x_i \le \tilde{x}_w)\, v_i = 0.5 \tag{1.30}$$

where the v_i are the normalized weights in (1.26) and the function $1(e)$ is an indicator that takes the value 1 if the statement e is true and is 0 otherwise. Other percentiles are calculated by replacing the 0.5 in (1.30) by the appropriate fraction. In practice, the easiest way to work is to sort the data in order of increasing x_i, and then to calculate a running sum of the normalized weights. The percentiles are then read off from this running sum. Example 1.1 in the Code Appendix gives the STATA program used to calculate the results in Table 1.3; this shows how to label households by their deciles of total household income and expenditure, and to summarize variables by those deciles.

In principle, formulas can be derived for sampling variances for medians, variances, and functions of these statistics. But the calculations are in some cases quite

complex, and the approximations and assumptions required are not always palatable. As we shall see below, the bootstrap often offers a more convenient way to assess sampling variability in these cases.

*Technical note: sampling variation of probability-weighted estimates

The probability-weighted mean is the ratio of two estimates, $\hat{X}_{tot}$ and $\hat{N}$. The variance of such a ratio can be approximated by

$$\text{var}(\bar{x}_w) \approx N^{-2}[\text{var}(\hat{X}_{tot}) - 2\bar{X}\text{cov}(\hat{X}_{tot}\hat{N}) + \bar{X}^2\text{var}(\hat{N})]. \tag{1.31}$$

I illustrate only the derivation of the first term in square brackets; the other variance and the covariance are readily obtained in the same manner. From (1.21),

$$\hat{X}_{tot} = \sum_{i=1}^{N} t_i w_i x_i. \tag{1.32}$$

The only random variables in (1.32) are the t's, so that

$$\text{var}(\hat{X}_{tot}) = \sum_{i=1}^{N} w_i^2 x_i^2 \text{var}(t_i) + \sum_{j \neq i}^{N} \sum_{i=1}^{N} w_i w_j x_i x_j \text{cov}(t_i t_j). \tag{1.33}$$

The t's follow a multinomial distribution, so that the variance of t_i is $n\pi_i(1-\pi_i)$ and the covariance between t_i and t_j is $-n\pi_i\pi_j$. Substituting in (1.33) and rearranging gives (1.23) above. That (1.24) is unbiased for (1.23) is shown by inserting t's, and changing the summation from n to N, and taking expectations (see also Cochrane). Substitution of (1.33) into (1.31) together with the comparable formulas for $\text{var}(\hat{N})$ and $\text{cov}(\hat{X}_{tot},\hat{N})$ gives (1.27). The sample estimate (1.28) is constructed by replacing the square of N in the denominator by the square of its estimate $\hat{N}$, replacing $\bar{X}$ by its estimate $\bar{x}_w$, and replacing the population sum of squares by its sample equivalent, remembering that for every household in the sample, there are w_i in the population. The scaling factor $n/(n-1)$ is conventional and clearly has little effect if n is large (see also (1.29)).

Stratification

The effect of stratification is to break up a single survey into multiple independent surveys, one for each stratum. When we think of the different samples that might be drawn in replications of the survey, the strata will be held fixed while the particular households selected from each will vary from sample to sample. Without stratification, the fraction of the sample in each stratum is left to chance. In consequence, estimates of population parameters vary across samples because each sample has different fractions of observations in each stratum so that, when the means differ across strata, their weighted average will also differ. As a result, stratification can reduce sampling variability whenever the means differ across strata.

Suppose that there are S strata, labeled by s, that we know the total population N as well as the population in each stratum N_s, and that the mean for stratum s is $\bar{X}_s$. The population mean, or "grand" mean, is then

$$\bar{X} = \sum_{s=1}^{S} (N_s/N)\,\bar{X}_s \tag{1.34}$$

which can be estimated from

$$\bar{x} = \sum_{s=1}^{S} (N_s/N)\,\bar{x}_s \tag{1.35}$$

where $\bar{x}_s$ is the estimated mean for stratum s. Note that the calculation of these means will typically involve weights, as in (1.25) above. The population shares may or may not be the same as the sample shares; stratification is about breaking up the sample into subsamples, not about weighting. Nevertheless, it is often the case that the sampling fractions are different in different strata. In such cases, as I shall show below, it is possible to incorporate the stratum weights into the weights for each observation.

Because the strata are independent, the variance of the estimate of the population mean (1.35) takes the simple form, ignoring the *fpc*,

$$V(\bar{x}) = \sum_{s=1}^{S} (N_s/N)^2\, V(\bar{x}_s), \tag{1.36}$$

where $V(\bar{x}_s)$ is the variance of the estimate of the stratum mean. If instead of the stratified survey, we had used a simple random sample, the numbers in each stratum, n_s, would be random variables within the total sample size n. Hence, if we write the sample mean for the simple random sample design analogously to (1.35),

$$\bar{x}_{srs} = \sum_{s=1}^{S} (n_s/n)\,\bar{x}_s \tag{1.37}$$

which looks very similar to (1.35), especially since the expectations of the sample ratios (n_s/n) are the population ratios (N_s/N). However, in the unstratified design, the fractions of the sample in each stratum will vary from sample to sample, so that the variability of the estimate will not only have a component from the variability of the stratum means, as in the stratified sample, but also a component from the variability of the fractions in each stratum. With some algebra, it can be shown that the variances of the two estimates are linked by the approximation

$$V(\bar{x}_{srs}) \approx V(\bar{x}) + n^{-1} \sum_{s=1}^{S} (N_s/N)(\bar{x}_s - \bar{x})^2. \tag{1.38}$$

As would be expected, the variance is larger in the simple random sample than in the stratified sample, and will be the more so the larger is the heterogeneity across strata. When the strata means coincide with the grand mean, there is no increase in efficiency from stratification.

In practice, estimation in stratified samples is usually done using simple adaptations of the formulas for the weighted estimates in the previous subsection. Suppose that we think of each stratum as a separate survey, and write the inflation factors for households in stratum s as w_{is}, where, corresponding to (1.17),

$$w_{is} = (n_s \pi_{is})^{-1} \tag{1.39}$$

where n_s is the sample size from stratum s, and π_{is} is the probability that i is drawn at each trial. The within-stratum sum of these weights is an unbiased estimator of the stratum population, N_s, and the grand sum over all observations is an unbiased estimator of the total population. Hence, we can define a probability-weighted estimate of the grand mean corresponding to (1.35)

$$\bar{x}_w = \sum_{s=1}^{S} (\hat{N}_s/\hat{N})\,\bar{x}_{sw} \tag{1.40}$$

where $\bar{x}_{sw}$ is the probability-weighted mean (1.25) computed for stratum s. Note that in (1.40), unlike (1.35), the fractions of the population in each stratum are estimated, and are therefore random variables. As a result, (1.40) loses what might be expected from stratification, that without variation within strata, there is no variance in an estimate from a stratified sample (see (1.36). That this is not the case is because there are differential weights within strata, so that different samples will give different weights to each stratum mean. Of course, if the ratios N_s/N are available, they can be used to replace the estimates in (1.40), with a gain in precision.

If we substitute the appropriate sums of weights for $\hat{N}_s$ and $\hat{N}$ in (1.40), we get

$$\bar{x}_w = \left(\sum_{s=1}^{S}\sum_{i=1}^{n_s} x_{is}w_{is}\right) \Big/ \left(\sum_{s=1}^{S}\sum_{i=1}^{n_s} w_{is}\right) \tag{1.41}$$

where x_{is} is the observation from household i in stratum s. Note that (1.41) is simply the probability-weighted mean without any explicit allowance for the stratification; each observation is weighted by its inflation factor and the total divided by the total of the inflation factors for the survey. Like (1.25), it is also a ratio estimator; the mean is estimated by the ratio of the estimated total—in the numerator—to the estimated population size—in the denominator. In consequence, we can use the variance formula (1.31) to approximate the variance of (1.41) in terms of the variances and covariance of the totals which, in turn, are sums of stratum-specific terms because sampling is independent within each stratum. The algebra is similar to that used to derive (1.23) and (1.27), and yields

$$V(\bar{x}_w) = \frac{1}{N^2}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{N_s}\pi_{is}\left[\left(\frac{x_{is}}{\pi_{is}} - N_s\bar{X}_s\right) - \bar{X}\left(\frac{1}{\pi_{is}} - N_s\right)\right]^2. \tag{1.42}$$

As was the case for (1.23) and (1.24), a feasible sample-based estimator of (1.42) starts from defining $z_{is} = x_{is}w_{is}$, and using the formula

$$\hat{v}(\bar{x}_w) = \sum_{s=1}^{S}\frac{n_s}{n_s - 1}\sum_{i=1}^{n_s}[(z_{is} - \bar{z}_s) - \bar{x}_w(w_{is} - \bar{w}_s)]^2 \tag{1.43}$$

where $\bar{z}_s$ and $\bar{w}_s$ are the stratum means of the z's and the weights, respectively.

Two-stage sampling and clusters

Within strata, most household surveys collect their data in two stages, first sampling clusters, or primary sampling units (PSUs), and then selecting households from

within each cluster; this is the standard two-stage stratified design. Clustered samples raise different statistical issues from stratified samples. When we imagine replicating a survey, which is how we think about sampling variability, the strata are held constant from sample to sample, but new clusters are drawn every time. For example, each potential survey might select an equal number of households from all of the provinces of the country (the strata), but would always select a new set of villages within provinces (the clusters). Probabilities of selection can differ at either or both stages of the survey, between clusters, or between households within clusters. The formulas for weighted and unweighted means are not affected by the two-stage design any more than they were affected by stratification. But the sampling variability of these estimates *is* affected by the design. Because households within clusters are often similar to one another in their relevant characteristics, it is frequently the case that clustering will *increase* variability compared with simple random sampling. In this subsection, I introduce some notation, record the formulas for the means and their sampling variation, and explain how it is that clustering reduces precision, and the consequences of ignoring the clustering in calculating variability. This is perhaps the most important message of this subsection, that it can be a serious mistake to treat a two-stage sample as if it were a simple random sample; the use of standard formulas can seriously overstate the precision of the estimates.

I start by supposing that there is no stratification, or equivalently, that there is only a single stratum. Because separate strata can be thought of as separate surveys, a simple way to deal with a stratified sample is to work with one stratum at a time, and then to reassemble the survey as a whole from its components. I shall do so, and record the relevant formulas, at the end of the subsection.

We need notation for the numbers of clusters and households in the sample and in the population. Suppose that there are N clusters in the population from which n are selected into the survey; this preserves the previous notation for the case where each observation is a cluster. I use the suffix c to denote a cluster or PSU, and m_c and M_c to denote the number of sample and population households in cluster c. I shall use T for the total number of households in the population

$$T = \sum_{c=1}^{N} M_c. \tag{1.44}$$

Suppose that sampling is with replacement, but with unequal probabilities at both stages. I use π_c to denote the probability of selection for cluster c in the first stage, and π_{ic} for the probability that i is selected at the second stage, conditional on c having been selected at the first. The unconditional probability that household i in cluster c is selected at a single two-stage draw is therefore $\pi_{ic}\pi_c$. Proceeding as before, we can define inflation factors for each stage of the survey. To differentiate these stage-specific inflation factors from the overall inflation factors, I use different notations for each. Define h_c and h_{ic} by

$$h_c = (n\pi_c)^{-1}, \quad h_{ic} = (m_c\pi_{ic})^{-1} \tag{1.45}$$

so that h_c is the number of population *clusters* represented by cluster c, and h_{ic} is the number of cluster-c households represented by household i in cluster c. The overall inflation factor, the number of households in the population represented by household i, is the product of h_c and h_{ic}, which corresponds to our previous survey weight w_{ic},

$$w_{ic} = h_c h_{ic} = (\pi_c \pi_{ic} m_c n)^{-1}. \tag{1.46}$$

Note that the sum of these weights over cluster c,

$$w_c = \sum_{i=1}^{m_c} w_{ic} = h_c \sum_{i=1}^{m_c} h_{ic} \tag{1.47}$$

is an inflation factor that tells us how many population households are represented by the collectivity of sample households in cluster c.

The probability-weighted mean is defined in the standard way; adapting (1.25) to recognize the clusters,

$$\bar{x}_w = \frac{\sum_{c=1}^{n}\sum_{i=1}^{m_c} w_{ic}x_{ic}}{\sum_{c=1}^{n}\sum_{i=1}^{m_c} w_{ic}} = \frac{\sum_{c=1}^{n} w_c \bar{x}_{cw}}{\sum_{c=1}^{n} w_c} = \sum_{c=1}^{n} v_c \bar{x}_{cw} \tag{1.48}$$

where $\bar{x}_{cw}$ is the probability-weighted mean for cluster c and the v's are the cluster weights (1.47) normalized to sum to 1. The evaluation of the variance of (1.48) is complicated by the randomness in the cluster means, as well as in the selection of clusters themselves. The algebra is simplified if we follow Cochrane (1977, pp. 275–76) and calculate expectations and variances in two stages, so that, for the mean $\bar{x}_w$

$$E(\bar{x}_w) = E_1[E_2(\bar{x}_w)] \tag{1.49}$$

where the expectation E_2 is taken with respect to the second-stage sampling, treating the choice of clusters as fixed, and where E_1 is taken with respect to the choice of clusters. The corresponding variance formulas are

$$V(\bar{x}_w) = V_1[E_2(\bar{x}_w)] + E_1[V_2(\bar{x}_w)]. \tag{1.50}$$

The application of (1.49) and (1.50) is relatively straightforward using previous results, and after a good deal of algebra, we reach

$$V(\bar{x}_w) = n^{-1}\sum_{c=1}^{N} \pi_c \phi_c^2 (\bar{X}_c - \bar{X})^2 + T^{-2}\sum_{c=1}^{N}\sum_{j=1}^{M_c} w_{jc}(X_{jc} - \bar{X}_c)^2 \tag{1.51}$$

where ϕ_c is the fraction of the population in cluster c, and $\bar{X}_c$ and $\bar{X}$ are the true means for the cluster and the population, respectively. (For comparison, note that (1.51) reduces to (1.27) when each cluster contains a single household and the weights satisfy $n^{-1}\pi_c = w_c$ and $\phi_c = N^{-1}$). A consistent estimate of (1.51) can be obtained from

$$\hat{v}(\bar{x}_w) = \frac{n}{n-1} \sum_{c=1}^{n} v_c^2 (\bar{x}_{cw} - \bar{x}_w)^2 \tag{1.52}$$

which is identical to (1.28), with households replaced by clusters, and individual data points replaced by cluster means.

It should be emphasized that, in spite of its formal similarity, the variance (1.52) is quite different from the corresponding formula when there is no clustering, and that the use of the incorrect formula can be seriously misleading. While it is sometimes the case that estimated variances are not much altered by allowing for stratification or differential weights, clustering is ignored at one's peril. I illustrate for the simplest case, where there are M households in each of the N clusters, and at the first stage, clusters are selected by simple random sampling. Each cluster is then equally weighted, so that when we estimate the variance from (1.52) we get

$$\hat{v}(\bar{x}) = \frac{1}{n(n-1)} \sum_{c=1}^{n} (\bar{x}_c - \bar{x})^2. \tag{1.53}$$

If we were mistakenly to ignore the clustering and treat each observation as an independent draw in a simple random sample of size mn, we would use (1.12) to give

$$\hat{v}_{srs}(\bar{x}) = \frac{1}{mn(mn-1)} \sum_{c=1}^{n} \sum_{i=1}^{m} (x_{ic} - \bar{x})^2 = \frac{1}{mn} \hat{s}^2. \tag{1.54}$$

If we substitute for the cluster means in (1.53) and rearrange, we get

$$\hat{v}(\bar{x}) \approx \hat{v}_{srs}(\bar{x})[1 + (m-1)\hat{\rho}] \tag{1.55}$$

where $\hat{\rho}$ is defined by

$$\hat{\rho} \approx \frac{\sum_{c=1}^{n} \sum_{j=1}^{m} \sum_{k \neq j}^{m} (x_{jc} - \bar{x})(x_{kc} - \bar{x})}{mn(m-1)\hat{s}^2}. \tag{1.56}$$

The quantity in (1.56) is a sample estimate of the *intracluster correlation coefficient*. Like any correlation coefficient, ρ measures the similarity of values, in this case within the clusters. When all the x's are the same in the same cluster, $\rho = 1$, when they are unrelated, $\rho = 0$. In practice, for quantities like income and consumption in rural areas of developing countries, ρ is often substantially larger than zero and values of 0.3 to 0.4 are frequently encountered.

Equation (1.55) shows how the magnitude of ρ affects the variability of sample estimates, at least in this simple case. When $\rho = 0$, the variance of the estimate from the clustered sample coincides with the variance of the estimate from the simple random sample. At the other extreme, when $\rho = 1$, the factor in square brackets in (1.55) is m, so that $\hat{v}(\bar{x}) = \hat{s}^2/n$ and the effective sample size is not the number of sample *observations*, mn, but the number of sample *clusters*, n. When the observations are the same within each cluster, sampling more than one from each does nothing to increase the precision of the estimate. In the next subsection, I shall give some practical examples of the way in which assumptions about sample

design affect calculations of standard errors, and of the potential for being misled by the wrong assumption.

Two-stage samples often use a "self-weighting" design. At the first stage, clusters are selected with probability proportional to the number of households they contain while, at the second stage, an equal number of households is drawn from each cluster using simple random sampling. This has the effect of making the overall, surveywide, inflation factors w_{ic} the same for all households. To see how this works, set

$$\pi_c = \frac{M_c}{\sum_{c=1}^{N} M_c} = \frac{M_c}{T}, \quad \pi_{ic} = \frac{1}{M_c}, \quad m_c = m \tag{1.57}$$

so that, substituting into (1.46), we have

$$w_{ic} = T/(mn) \tag{1.58}$$

which is the same as in a simple random sample. (Of course, this only applies to the weights and the computation of sampling variability must still allow for the two-stage design. A two-stage self-weighting sample is *not* the same thing as a simple random sample.)

Self-weighting is simple and elegant. It also had practical utility when computation was so difficult that the additional complexity of weights was best avoided if possible. However, self-weighting designs are rarely self-weighting in practice because adjustments are often made to the weights after the survey, for example to compensate for unanticipated nonresponse by some set of households, and weights have to be used in any case. An example is the South African Living Standards Survey, which had a self-weighting design, but which had to be weighted ex post (see Table 1.1). Since computation is hardly an issue today, it is unclear why the design remains so popular.

In practice, it is necessary to combine the formulas for the clustered case with those that allow for multiple strata. This is conceptually straightforward, although the notation makes the formulas look forbidding. I denote the strata by the subscript s, and rewrite the mean for stratum s from (1.48) as

$$\bar{x}_{sw} = \sum_{c=1}^{n_s} w_{cs}\bar{x}_{csw} \Big/ \sum_{c=1}^{n_s} w_{cs} \tag{1.59}$$

where the only change is to add a suffix s to indicate the stratum. From (1.35), we can compute the grand mean over all the strata using

$$\bar{x}_w = \sum_{s=1}^{S} \hat{N}_s \bar{x}_{sw} / \hat{N} \tag{1.60}$$

where the hats denote the usual estimates from the sums of the weights. Substituting (1.59) into (1.60) gives the probability-weighted estimate of the grand mean in the familiar form of a ratio between the estimated total of X, and the estimated population size,

$$\bar{x}_{sw} = \frac{\sum_{s=1}^{S}\sum_{c=1}^{n_s} w_{cs}\bar{x}_{csw}}{\sum_{s=1}^{S}\sum_{c=1}^{n_s} w_{cs}} = \frac{\sum_{s=1}^{S}\sum_{c=1}^{n_s}\sum_{i=1}^{m_c} w_{ics}x_{ics}}{\sum_{s=1}^{S}\sum_{c=1}^{n_s}\sum_{i=1}^{m_c} w_{ics}} = \frac{\hat{X}_{tot}}{\hat{N}} \tag{1.61}$$

which is simply the weighted mean using all the observations and all the weights in the survey. An estimate of the variance of (1.59) and (1.61) is obtained following the same general procedures as for the ratio estimator (1.41) in the stratified case, but making the adaptions for clustering for the variances within each stratum. The formulas are simplified if we define the cluster level variable

$$z_{cs} = \sum_{i=1}^{m_c} w_{ics}x_{ics}. \tag{1.62}$$

If z_s is the mean of z_{cs} over clusters in stratum s, the variance of (1.61) can be estimated from (compare (1.43))

$$\hat{v}(\bar{x}_{sw}) = \frac{1}{\hat{N}^2}\sum_{s=1}^{S}\frac{n_s}{n_s-1}\sum_{c=1}^{n_s}[(z_{cs}-z_s)-\bar{x}_{sw}(w_{cs}-\bar{w}_s)]^2 \tag{1.63}$$

where w_{cs} is the total weight in cluster c of stratum s, and $\bar{w}_s$ is the stratum-s mean of w_{cs}. Sample code for equation (1.63) is given in Example 1.2 of the Code Appendix; it is also available as a special case of the more general formulas available in Version 5.0 of STATA.

A superpopulation approach to clustering

It is also possible to take a superpopulation approach to clustering, and as was the case with simple random sampling, the results are in many ways simpler. They also provide a useful bridge to the discussion of clustering and regression in Section 1 of Chapter 2. Suppose that there are no weights and that

$$x_{ic} = \mu + \alpha_c + \epsilon_{ic} \tag{1.64}$$

where μ is the mean, α_c is a cluster effect, and ϵ_{ic} is a random variable with mean 0 and variance σ_ϵ^2 that is independently and identically distributed for all i and c. The cluster effects α_c are also random with mean 0 and variance σ_α^2, are independently and identically distributed across clusters, and are independent of the ϵ's. These independence assumptions are the counterpart of the independence of the two stages of the sampling in the finite-population approach and the presence of the α's allows cluster means to differ from the overall mean.

As before, the obvious estimator of μ is the sample mean $\bar{x}$, and straightforward calculation gives

$$E(\bar{x}) = \mu; \quad V(\bar{x}) = n^{-1}\sigma_\alpha^2 + (nm)^{-1}\sigma_\epsilon^2. \tag{1.65}$$

The variance in (1.65) is the counterpart of (1.51) when both stages are by simple

random sampling. It is also instructive to write it in the form

$$(1.66) \qquad V(\bar{x}) = \frac{\sigma_\epsilon^2 + \sigma_\alpha^2}{nm}\left(1 + (m-1)\frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + \sigma_\alpha^2}\right) = \frac{\sigma^2}{nm}[1 + (m-1)\rho],$$

where $\sigma^2 = \sigma_\epsilon^2 + \sigma_\epsilon^2$ is the variance of x_{ic} and ρ, the ratio of σ_α^2 to σ^2 is the intra-cluster correlation coefficient (compare (1.55) above).

An unbiased estimator of $V(\bar{x})$ is given by

$$(1.67) \qquad \hat{v}(\bar{x}) = n^{-1}\hat{s}_1^2 = n^{-1}(n-1)^{-1}\sum_{c=1}^{n}(\bar{x}_c - \bar{x})^2$$

which corresponds to (1.52). In both cases the variance can be computed by considering only the variation of the estimated cluster means, ignoring within cluster variability.

Illustrative calculations for Pakistan

For poverty and welfare calculations, we often use household per capita expenditure (PCE)—total expenditure on goods and services divided by household size—as a measure of living standards. I use measurements of PCE from the Pakistan Living Standards Survey—formally the Pakistan Integrated Household Survey, or PIHS—to illustrate the sort of design that is encountered in practice, as well as the consequences of the design for the calculation of statistics and their standard errors.

The survey documentation will usually explain how the stratification was done and the primary sampling units selected. Identifiers for the stratum and cluster of each household are sometimes included as data in one of the household files, but are more usually incorporated into the household identifiers. This is the case for the PIHS, where the first three digits of the household code gives the stratum, the next three the cluster, and the last three the household within the cluster. Example 1.2 in the Code Appendix shows how the household identifiers are broken down to give stratum and cluster identifiers. In the PIHS there are 22 strata; the four provinces—Punjab, Sindh, North-West Frontier (NWFP), and Baluchistan—which are further stratified by urban and rural and by income level. There are 280 PSUs, or clusters, between 2 and 37 in each of the strata, and there are between 13 and 32 households in each cluster. The probability weights are sufficiently correlated with PCE for the weighting to make a difference; the unweighted average of household PCE is 730 rupees a month, whereas the weighted mean is only 617 rupees a month.

Table 1.5 shows these estimates for the country as a whole and for its four provinces together with various calculated standard errors. The first two columns are doubly incorrect; the estimated means ignore the probability weights, and the standard errors ignore the sample design. The weights are negatively correlated with PCE—in this case, better-off households are oversampled—so that the unweighted means are biased up, which makes the standard errors of little interest. The third column shows the (correct) weighted means, and the other columns show various possible standard errors, each calculated under different assumptions about

Table 1.5. Estimates of mean household per capita expenditure and calculated standard errors, Pakistan, 1991
(rupees per capita per month)

Province	$\bar{x}$	$\hat{s}/\sqrt{n}$	$\bar{x}_w$	$\hat{s}_w/\sqrt{n}$	*Weight*	*Weight, strata*	*Weight, strata, PSU*
Punjab	660	16.0	584	13.5	17.4	17.5	22.6
Sindh	754	21.9	693	18.3	17.4	17.4	35.1
NWFP	963	66.8	647	30.8	24.6	24.6	37.6
Baluchistan	682	30.6	609	31.0	41.0	41.1	96.2
Pakistan	730	14.1	617	9.9	12.0	12.0	17.0

Note: $\bar{x}$ is the unweighted mean, and $\hat{s}/\sqrt{n}$ a standard error calculated according to (1.12). $\bar{x}_w$ is the probability-weighted mean and $\hat{s}_w/\sqrt{n}$ is computed from the weighted sample variance (1.29). The column headed "weight" takes the probability weights into account using (1.28), but ignores stratification and clustering. The column headed "weight, strata" uses (1.28) for each stratum and then adds the stratum variances using (1.36). The final column is the appropriate standard error calculated from (1.60).
Source: Author's calculations using the Pakistan Integrated Household Survey, 1991 (see Example 1.2 in the Code Appendix).

the sample design. The column headed $\hat{s}_w/\sqrt{n}$ is what might be calculated if the weights were used to estimate the standard deviation, using (1.28) rather than (1.11), but the sample was incorrectly assumed to have been drawn as a simple random sample for which (1.10) would be the true variance. The next column recognizes the probability weights explicitly and comes from (1.28), but takes no account of stratification nor clustering. Allowing for the stratification in the next column has very little effect on the calculations because most of the variation is within the strata rather than between them (see equation (1.38) above). The largest changes come in the last column, where the cluster structure is recognized and the standard error calculated from (1.62). Because PCE is correlated within the clusters —the intracluster correlation coefficient for the whole sample is 0.346—the effective sample size is a good deal smaller than 4,800, and the standard errors that recognize the fact are a good deal larger. For the country as a whole, the correctly calculated standard error in the last column is almost twice that in column 4, and for the province of Baluchistan, the ratio is more than three.

The bootstrap

The more complex is the survey design, the more difficult it becomes to assess the variability of estimates based on the results. In the previous subsections, I have discussed only a few of the most important designs, and have provided formulas for variances only for estimates of the mean. There are other designs, some of truly bewildering complexity, and there are other statistics in which we are interested. Books on sampling techniques provide many more results than can be discussed here, but even the full range of formulas often falls short of what we need. For example, it is more difficult to obtain good estimates of the sampling variability of

a median than of a mean, and yet in many situations, the median is the more useful measure of central tendency, if only because it is less influenced by the sort of outliers that often occur in real data. It should also be noted that for ratios of random variables, such as the probability-weighted mean, the variance formulas are approximate, not exact, and the accuracy of the approximation is not always apparent in practice. Yet means, medians, and ratios are among the simplest quantities that we want to calculate from survey data. The econometric analyses in later chapters often involve more elaborate calculations and the derivation of sampling distributions in these cases can present formidable difficulties, especially when we want to allow for the survey design.

The bootstrap is an alternative method of assessing sampling variability. It is no panacea, and it will not always give better results than the variance formulas, even approximate formulas. But it offers a mechanical procedure that can be applied in a wide variety of "difficult" situations, it works in much the same way whether we are estimating something straightforward, like a mean or a median, or something more complex, and it substitutes computer power for statistical analysis and algebra, a substitution that is welcome to all who do not enjoy the contemplation of balls and urns. The bootstrap, which was invented by Efron (1979), samples repeatedly, not from the population, which is of course not available for the purpose, but from the sample. For each resampling, we make whatever calculation we are interested in, and we keep track of the results over the replications. The variability of these resampled estimates is then used to assess the variability of the estimator over different samples from the population. An excellent, readable, and clear introduction to the bootstrap is provided by Efron and Tibshirani (1993).

As always, the simplest case is where we have a simple random sample of (say) n households. The bootstrap works by repeatedly drawing samples of size n from the sample *with replacement.* At each replication, the statistic of interest—mean, median, variance, or whatever—is calculated and stored. After K replications, the K values of the statistic are used to compute a measure of dispersion, for example the standard deviation as a measure of standard error, or—and necessarily in cases where the moments may not exist—percentiles used to estimate percentiles of the sampling distribution. The value of K will vary from application to application. Small values (around 100, say) will typically give a good idea of variance, when the variance exists, but when we need to calculate the fractions of occurrences in the tails of the distributions—as will often be the case for percentiles—much larger numbers of replications may be required. Given a desired level of precision and some idea of the sampling distribution, the required number of replications can be calculated in the usual way.

In simple cases, bootstrapping can be shown to lead back to the usual statistics. For example, suppose that we have a simple random sample $(x_1, x_2, \ldots, x_n)$ from which we draw bootstrap samples, always with replication and of the same size as the original. A typical replication might be denoted $(x_1^b, x_2^b, \ldots, x_n^b)$, with the superscripted b standing for "bootstrap." If we wished to bootstrap the mean, or the weighted mean, we would at each replication calculate the quantities $\bar{x}^b = n^{-1}\Sigma x_i^b$ or $\bar{x}_w^b = \Sigma w_i^b x_i^b / \Sigma w_i^b$, where the w's are drawn simultaneously with the x's.

Finding the means and variances of these expressions over the bootstrap replications can be done in exactly the same way as we found the means and variances for estimators using samples from the population rather than samples from the sample. The calculations are particularly straightforward because the bootstrap sample is drawn by simple random sampling and is the same size as the "population" so that t_i, the number of times each sample x_i appears in the bootstrap sample, is a random variable with mean 1, variance $(1-n^{-1})$, and covariance with t_j of $-n^{-1}$. Using these facts, it is straightforward to show that the mean across replications of the bootstrapped mean converges to the mean in the original sample, and that the variances are given by, for the unweighted mean:

$$V(\bar{x}^b) = n^{-2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.68}$$

and for the weighted mean:

$$V(\bar{x}_w^b) = \sum_{i=1}^{n} v_i^2(x_i-\bar{x})^2. \tag{1.69}$$

Up to the ratio $n/(n-1)$, the inclusion of which is a matter of convention in any case, these are identical to the variance formulas presented above (see (1.12) and (1.28), respectively). Hence, for both the weighted and unweighted mean, we get the same estimate of sampling variance either by direct calculation, or by simulation using the bootstrap, provided we have enough replications. Of course, it would be absurd to use the bootstrap in this case; the simulation is expensive and adds nothing to the direct and straightforward calculations. But there are many other cases where analytical formulas are not available, but where the bootstrap can be used in exactly the same way. And if the bootstrap did not give the right answer in these familiar settings, there would be no reason to trust it in more complex cases.

It should be noted that the formulas (1.68) and (1.69) do not contain any finite-population corrections. More generally, the fact that the bootstrap uses sampling with replacement will prevent it from giving good results when the original sample is large relative to the population (see Rao and Wu 1988 and Sitter 1992 for a discussion of methods of dealing with these cases). Care must also be taken in applying the bootstrap to dependent observations, and it cannot be applied without modification to data that were collected using a two-stage clustered design. Attempts to do so will usually understate sampling variability just as the use of formulas that ignore clustering will usually understate variability. However, the bootstrap can still be applied to a stratified clustered sample if we treat the strata separately, each its own survey, and if we resample, not the basic underlying units—the households—but rather the primary sample units—the clusters. This is straightforward to implement; a list of the n sample clusters is made, a bootstrap sample of size n is drawn with replacement, and the individual cluster-level data merged in (see Example 1.3 in the Code Appendix). Following this procedure for the PIHS and using 100 bootstrap replications gives a bootstrapped standard error for PCE of 16.5, compared with 17.0 from the formula (see the last column of Table 1.5). The median PCE is much lower than mean PCE, 461 as opposed to 617 rupees per month, and

its bootstrapped standard error is only 7.7, less than half the estimated standard error of the mean. Because the median is relatively unaffected by outliers, and because the distribution of PCE is so positively skewed, the median varies much less from one sample to another than does the mean.

These calculations illustrate only the most basic use of the bootstrap although other, more complex, examples will be seen in later chapters. However, the replication of the quantity of interest—the mean, median, or whatever—is not always the best way to use the bootstrap. In particular, when we wish to calculate a confidence interval, the recommended procedure is not to bootstrap the estimate itself, but rather to bootstrap the distribution of the t-value. This is feasible in the frequently occurring situation where we have an approximate or large-sample version of the standard error, but are skeptical about its accuracy in the application at hand. The method works as follows. Start out as usual, drawing repeated samples from the base sample, taking into account the design, and calculating the estimate for each bootstrap replication. But instead of recording the estimate itself, subtract from it the estimate from the original full sample, and divide the difference by the approximate standard error $\hat{\sigma}$, say. The result is a bootstrapped t- or z-value whose distribution would be $N(0,1)$ if the estimates were normally distributed and the approximate standard were correct. But we do not need to assume either normality or accuracy of the approximation. Instead we carry out enough replications of the bootstrap to obtain an idea of the actual distribution of the bootstrapped t-values. In particular, if we want a 90 percent confidence interval for the sample mean, we calculate the fifth and ninety-fifth percentiles of the distribution of the t's, t_{05} and t_{95}, and use them to construct the confidence interval $[\hat{x} - t_{05}\hat{\sigma}, \hat{x} + t_{95}\hat{\sigma}]$. Note that the accurate calculation of the tails of the distribution is likely to require large numbers of bootstrap replications. The benefit is that the procedure will provide more accurate estimates of confidence intervals than either the simple descriptive bootstrap or the approximate standard errors. An explanation of why this should be so is beyond the scope of this book; the interested reader is referred to Hall (1995) for a review.

1.5 Guide to further reading

There are several good texts on survey design, notably Cochrane (1977), Hansen, Hurwitz, and Madow (1953), Kish (1965), Som (1973), Levy and Lemenshow (1991), and Wolter (1985). Much the same ground is covered by Murthy (1977), who also gives a description of the design of the Indian National Sample Survey. The discussion of sample design and sampling variation in Section 1.4 makes most use of Cochrane's treatment; I have also been influenced by the discussion of sample design and poverty measurement by Howes and Lanjouw (1995). Casley and Lury (1981) discuss sample surveys in developing countries, covering sample and questionnaire design and a host of practical matters. Much the same territory for developed countries is covered by Groves (1989), who discusses many of the issues of this chapter, including a much more systematic treatment of the various sources of measurement error. He also, like Casley and Lury, discusses question-

naire design, a major omission from this chapter. Sample design issues in the LSMS surveys are dealt with in Grosh and Muñoz (1996); see also Ainsworth and Muñoz (1986) and Grootaert (1993). The LSMS group is currently preparing a monograph that will deal with the experience to date and make recommendations about the design of similar surveys in the future. Data quality in a broader perspective is covered in the special June 1994 issue of the *Journal of Development Economics.* Pudney (1989) and Skinner, Holt, and Smith (1989) both contain chapters on survey design and its implications for analysis, Pudney from an econometric perspective, and Skinner et al. from a statistical perspective. Both bridge the material in this chapter and the next. An excellent introduction to the bootstrap is provided by Efron and Tishbirani (1993), and Wolter (1985) discusses a number of alternative computation-intensive methods for calculating variance. Version 5 of STATA (which was released in the fall of 1996, and thus too late for the applications in this book) contains a set of commands for dealing with complex survey designs; as often, the documentation is a good introduction to the theory. Among many other things, these commands implement the formulas in Section 1.4. A review of econometric applications of the bootstrap is Jeong and Maddala (1993).