2

Data for Health Equity Analysis: Requirements, Sources, and Sample Design

The first step in health equity analysis is to identify appropriate data and to understand their potential and their limitations. This chapter provides an overview of the data needs for health equity analysis, considering how data requirements may vary depending on the analytical issues at hand. The chapter also provides a brief guide to different sources of data and their respective limitations. Although there is some scope for using routine data, such as administrative records or census data, survey data tend to have the greatest potential for assessing and analyzing different aspects of health equity. With this in mind, the chapter also provides examples of different types of survey data that analysts may be able to access. Finally, it offers a brief discussion and illustration of the importance of sample design issues in the analysis of survey data.

Data requirements for health equity analysis

Health outcomes and health-related behavior

Data on health outcomes are a basic building block for health equity analysis. But how can health be measured? Murray and Chen (1992) have proposed a classification of morbidity measures that distinguishes between self-perceived and observed measures (see table 2.1).

For most of these measures, data are not collected routinely and can be obtained only through surveys. However, as is discussed further below, surveys differ substantially, both in the range of measures covered and in the approach to measurement. For example, some surveys include only short questions about illness episodes. Other surveys, such as the Indonesia Family Life Survey, use trained health workers in enumerator teams and collect detailed "observed" morbidity data, including measured height, weight, hemoglobin status, lung capacity, blood pressure, and the speed with which the respondent was able to stand up five times from a sitting position.

Health equity analysis can also be concerned with health-related behavior. The most obvious question in this respect concerns the utilization of and payment for health services. Questions on these issues have been included in many surveys, although the level of detail has varied considerably. But health-related behavior extends beyond the utilization of health services. Other variables relevant to health equity analyses include (i) behavior with an effect on health status (smoking,

Table 2.1 *A Classification of Morbidity Measures*

Self-Perceived	
Symptoms and impairments	Occurrence of illness or specific symptoms during a defined time period
Functional disability	Assessment of ability to carry out specific functions and tasks, or restrictions on normal activities (activities of daily living, e.g., dressing, preparing meals, or performing physical movement)
Handicap	Self-perceived functional disability within a specifically defined context
Observed	
Physical and vital signs	Aspects of disease or pathology that can be detected by physical examination (e.g., blood pressure and lung capacity)
Physiological and pathophysiological indicators	Measures based on laboratory examinations (e.g., blood, urine, feces, and other bodily fluids), body measurements (anthropometry)
Physical tests	Demonstrated ability to perform specific functions, both physical and mental (e.g., running, squatting, blowing up a balloon, or performing an intellectual task)
Clinical diagnosis	Assessment of health status by a trained health professional based on an examination and possibly specific tests

Source: Authors.

drinking, and diet), (ii) sexual practices, and (iii) household-level behavior (cooking practices, waste disposal, sanitation, sources of water). Some data on health service use are collected through routine information systems and population censuses (e.g., immunizations), but more detailed data are likely to be available only through surveys.

In the case of both health outcomes and health-related behaviors, it is important to keep in mind that variation in the variable of interest may arise for many reasons. Some of these relate to health system characteristics—for example, features of health financing or service delivery arrangements. But there is also likely to be variation due to biological, environmental, social, and other factors. Although it is often difficult to identify the contribution of different factors in practice, this is clearly an important issue to address in thinking about the policy implications of health equity analysis.

Living standards or socioeconomic status

Concerns for health equity arise in the relationships between health, or health-related behavior, and a variety of individual characteristics, such as social class, ethnic group, sex, age, and location. This book is concerned primarily with health equity defined in relation to socioeconomic status or living standards. The goal is to assess and to understand how health outcome or health-related behaviors vary with some measure of socioeconomic status or living standards. This is not to say that other types of comparisons are not of interest or relevant to policy—they clearly are. However, comparisons across, say sex, ethnic group, or geographic location, typically are not amenable to the techniques described in this book and hence receive less attention in what follows.

For the purposes of analyzing socioeconomic health inequalities, health-related information must be complemented by data on living standards or socioeconomic status. As is discussed in detail in chapter 6, there are many approaches to living standards measurement, including direct approaches (e.g., income, expenditure, or consumption) and proxy measures (e.g., asset index). In practice, the choice of living standards measure is often driven by data availability. Nonetheless, the choice of measure may influence the conclusions, so it is important for analysts to be aware of both the assumptions that underpin the chosen measure and the potential sensitivity of findings.

It is also important to distinguish between cardinal and ordinal measures of living standards. In the case of cardinal measures—for example, income or consumption in dollars or units of another currency—numbers convey comparable information about magnitude. Ordinal measures only rank individuals or households and do not permit comparisons of magnitudes across units. Some forms of health equity analysis require a cardinal measure of living standards. This is the case, for example, with financing progressivity and the poverty impact of health payments or health events. But in some cases, a ranking of households by some measure of living standards suffices. For example, measures of inequality in health and health care.

Other complementary data

For some forms of health equity analysis, data on the relevant health variables and a measure of living standards suffice. Often, however, other complementary data are required. For example, if multivariate analysis of health-related variables is to be used to better understand why observed inequalities arise, then data on community, household, and individual characteristics are required. This could include, for example, availability and characteristics of health care providers, environmental and climatic characteristics of the community, housing characteristics, education, sex, ethnicity, and so on.

Complementary data are also required to identify the distribution of public health expenditure in relation to living standards, so-called benefit-incidence analysis. The primary requirement is data on unit subsidies to health services. This information tends to be based on public expenditure data, but in some cases, more detailed cost information is available. Taking account of regional variation in unit costs requires data on the geographic location of the individual. Extending the analysis to examine variation in utilization with, for example, sex and ethnicity, requires data on the relevant demographics. Analysis of health financing fairness and progressivity depends on detailed data on user payments for health care.

The data requirements of different types of health equity analysis are summarized in table 2.2. As discussed in the rest of this chapter, the richest data for health equity analysis are likely to be from household surveys, but routine administrative data can also prove useful.

Table 2.2 *Data Requirements for Health Equity Analysis*

	Health variables	*Utilization variables*	*Living standards measure (ordinal)*	*Living standards measure (cardinal)*	*Unit subsidies*	*User payments*	*Back-ground variables*
Health inequality	✓		✓				
Equity in utilization		✓	✓				
Multivariate analysis	✓ or	✓		✓			✓
Benefit-incidence analysis		✓	✓		✓		(✓)
Health financing							
– Progressivity				✓		✓	
– Catastrophic payments				✓		✓	
– Poverty impact				✓		✓	

Source: Authors.

Data sources and their limitations

Household surveys and other nonroutine data

Household surveys are implemented on a regular basis in many countries and are probably the most important source of data for health equity analysis. Some household surveys are designed as multipurpose surveys, with a focus on a broad set of demographic and socioeconomic issues, whereas other surveys focus explicitly on health. Surveys sample from the population and are representative, or can be made representative, of the population as a whole (or whatever target population is defined for the survey). They have the advantage of permitting more detailed data collection than is feasible in a comprehensive census. Although many surveys are conducted on an ad hoc basis, there are an increasing number of multiround integrated survey programs. These include the Living Standards Measurement Study (World Bank), the Demographic and Health Surveys (ORC Macro), the Multiple Indicator Cluster Surveys (UNICEF), and the World Health Surveys (WHO).[1] The Living Standards Measurement Surveys are different from the other surveys in that they collect detailed expenditure data, income data, or both. In that sense, the Living Standards Measurement Surveys are a type of household budget survey.[2] Many countries implement household budget surveys in some form or other on a semiregular basis. A core objective of these surveys is to capture the essential elements of the household income and expenditure pattern. In some countries, the surveys focus exclusively on this objective and are hence of limited use for health equity analysis. However, it is also common for household budget surveys to include additional modules—for example, on health and nutrition—making them

[1] Some surveys, in particular the *Demographic and Health Surveys* and some budget surveys, are repeated on a regular basis and can in that sense be considered “semiroutine” data.

[2] These surveys are sometimes called “family expenditure surveys,” “expenditure and consumption surveys,” or “income and expenditure surveys.”

ideal for detailed analysis of the relationship between economic status and health variables.

Aside from large-scale household surveys, there are often a wealth of other nonroutine data that can be used for health equity analysis. This may include small-scale, ad hoc household surveys and special studies. It may also be possible to analyze data from facility-based surveys of users (exit polls) from an equity perspective. Relative to household surveys, exit polls are cheap to implement (in particular if they are carried out as a component of a health facility survey) and are an efficient means of collecting data on health service use and perceptions. With exit polls it is also easier to associate outcomes of health-seeking behavior (e.g., client perceptions of quality, payments, receipt of drugs) with a particular provider and care-seeking episode. This is often difficult in general household surveys, in which typically specific providers are not identified and in which recall periods of up to 4 weeks can result in considerable measurement error. However, unlike a household survey, an exit poll provides information only about users of health services.

Although survey data can be of considerable value for health equity analysis, it is important to be aware of their limitations. For one thing, large-scale surveys are expensive to conduct and, as a result, they tend to be implemented only periodically. Moreover, the scope, focus, and measurement approaches can vary across surveys and over time, limiting the scope for comparisons. Another challenge concerns the way the survey sample is selected and what this implies for making inferences from the data. It is important for analysts to be aware of the "representativeness" of the survey data and to take this into account when drawing conclusions about the wider population. It is also important to be aware of how to adjust the analysis for departures from simple random sampling, arising from, for example, stratification or multistage sampling. These issues are discussed in more detail below. Finally, survey data can be misleading, or "biased," because of problems in both the sample design and the way the survey is implemented (see box 2.1). Both of these problems can lead analysts to draw inappropriate inferences from survey data.

Box 2.1 *Sampling and Nonsampling Bias in Survey Data*

When analyzing survey data, analysts must be aware of potential sources of sampling and nonsampling bias. Sampling bias refers to a situation in which the sample is not representative of the target population of interest. For example, it is inappropriate to draw inferences about the general population on the basis of a sample drawn from users of health facilities. The reason is that different groups in the population use health facilities to different degrees—for example, due to differences in access or need. Sampling bias can also arise from the practice of "convenience sampling" aimed at avoiding remote or inaccessible areas or from the use of an inaccurate or inappropriate sampling frame. These potential problems point to the need for analysts to be well aware of the sampling procedure.

There are also many potential forms of nonsampling bias that can arise in the process of survey implementation. For example, nonresponse or measurement errors may be systematically related with variables of interest—for example, nonresponse about utilization of health services may be higher among the poor. If this were the case, analysts should be cautious in interpreting results and drawing inferences about the general population. In some cases, it may be possible to correct for this bias by modeling nonresponse. Other potential sources of nonsampling bias include errors in recording or data entry.

Source: Authors.

Routine data: health information systems and censuses

Some forms of routine data may be suitable for health equity analysis. Health information systems (HIS) collect a combination of health data through ongoing data collection systems. These data include administrative health service statistics (e.g., from hospital records or patient registration), epidemiological and surveillance data, and vital events data (registering births, deaths, marriages, etc.). HIS data are used primarily for management purposes, for example, for planning, needs assessments, resource allocation, and quality assessments. However, in some contexts, HIS data include demographic or socioeconomic variables that permit equity analysis. This is the case, for example, in Britain, where mortality data based on death certificates have been used for tabulations of mortality rates by occupational group since the 19th century. Similar analysis has been undertaken in other countries by ethnic group or educational level. Although many HIS do not routinely record socioeconomic or demographic characteristics, this may change in the future as the importance of monitoring health system equity becomes more recognized.

Periodic population and housing censuses are another form of routine data. Censuses are an important source of data for planning and monitoring of population issues and socioeconomic and environmental trends, in both developed and developing countries. National population and housing censuses also provide valuable statistics and indicators for assessing the situation of various special population groups, such as those affected by gender issues, children, youth, the elderly, persons with a disability, and the migrant population. Population censuses have been conducted in most countries in recent years.[3] Census data often contain only limited information on health and living standards, but have sometimes been used to study health inequalities by linking the information to HIS data. For example, socioeconomic differences in disease incidence and hospitalization have been studied by linking cause-of-death or hospital discharge records with census data. In the United States, there have also been efforts to link public health surveillance data with area-based socioeconomic measures based on geocoding. Although poor data quality and availability may currently preclude such linking in low-income countries, census data may be used to study equity issues by constructing need indicators for geographic areas based on demographic and socioeconomic profiles of the population.

Notwithstanding the potential for using routine data for health equity analysis, it is important to be aware of the common weaknesses of such data. In particular, coverage is often incomplete and data quality may be poor. For example, as a result of spatial differences in the coverage of health facility infrastructure, routine data are likely to be more complete and representative in urban than in rural areas. Similarly, better-off individuals are more likely to seek and obtain medical care and, hence, to be recorded in the HIS. Moreover, in cases in which routine data are used for management purposes, there may exist incentives for staff to record information inaccurately.

Data sources and their limitations are summarized in table 2.3.

[3] Information about dates of censuses in different countries can be found on http://unstats.un.org/unsd/demographic/census/cendate/index.htm.

Table 2.3 *Data Sources and Their Limitations*

Type of data	*Examples*	*Advantages*	*Disadvantages*
Survey data (household)	Living Standards Measurement Study (LSMS), Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), World Health Surveys (WHS)	Data are representative for a specific population (often nationally), as well as for subpopulations Many surveys have rich data on health, living standards, and other complementary variables Surveys are often conducted on a regular basis, sometimes following households over time	Sampling and nonsampling errors can be important Survey may not be representative to of small subpopulations of interest
Survey data (exit poll)	Ad hoc surveys, often linked to facility surveys	Cost of implementation is relatively low	Exit polls provide no information about nonusers
		Detailed information that can be related to provider characteristics is provided about users of health services	Data often contain limited information about household and socioeconomic characteristics
		Data on payments and other characteristics of visit are more likely to be accurate	Survey responses may be biased from “courtesy” to providers or fear of repercussions
Administrative data	HIS, vital registration, national surveillance system, sentinel site surveillance	Data are readily available	Data may be of poor quality Data may not be representative for the population as a whole Data contain limited complementary information, e.g., about living standards
Census data	Implemented on a national scale in many countries	Data cover the entire target population (or nearly so)	Data contain only limited data on health Data collection is irregular Data contain limited complementary information, e.g., about living standards

Source: Authors.

Examples of survey data

Demographic and Health Surveys (DHS and DHS+*)*

The Demographic and Health Surveys (DHS) have been an important source of individual and household-level health data since 1984[4] The design of the DHS drew on the experiences of the World Fertility Surveys[5] (WFS) and the Contraceptive Prevalence Surveys, but included an expanded set of indicators in the areas of population, health, and nutrition. DHS are nationally representative, with sample sizes typically ranging from 5,000 to 30,000 households.

The standard Demographic and Health Surveys consist of a household questionnaire and a women's questionnaire (ages 15–49). The core questionnaire concentrates on basic indicators and is standardized across countries. The household questionnaire covers basic demographic data for all household members, household and dwelling characteristics, and nutritional status of young children and women ages 15 through 49. The women's questionnaire contains information on general background characteristics, reproductive behavior and intentions, contraception, maternity care, breastfeeding and nutrition, children's health, status of women, AIDS and other sexually transmitted diseases, husband's background, and other topics. Some surveys also include special modules tailored to meet particular needs.

Aside from the standard DHS, interim surveys are sometimes implemented to collect information on a reduced set of performance-monitoring indicators. These surveys have a smaller sample size and are often conducted between rounds of DHS. In addition, many of the DHS have included tools to collect community-level data (Service Availability Modules). More recently, detailed facility surveys—Service Provision Assessments—have been implemented alongside household surveys with a view to providing information about the characteristics of health services, including their quality, infrastructure, utilization, and availability.

Further information, including a list of past and ongoing surveys, survey reports, questionnaires, and information on how to access the data, can be found on http://www.measuredhs.com.

The Living Standards Measurement Study

The Living Standards Measurement Study (LSMS) was established by the World Bank in 1980 to explore ways of improving the type and quality of household data collected by government statistical offices in developing countries. LSMS surveys are multitopic surveys, designed to permit four types of analysis: (i) simple descriptive statistics on living standards, (ii) monitoring of poverty and living standards

[4] For further information about the history of DHS, see http://www.measuredhs.com/aboutdhs/history.cfm. In 1997 DHS changed its name to DHS+ to reflect the integration of DHS activities under the MEASURE program. Under that mandate, DHS+ is charged with collecting and analyzing demographic and health data for regional and national family planning and health programs.

[5] The WFSs were a collection of internationally comparable surveys of human fertility conducted in 41 developing countries in the late 1970s and early 1980s. The project was conducted by the International Statistical Institute (ISI), with funding from USAID and UNFPA.

over time, (iii) description of the incidence and coverage of government programs, and (iv) measurement of the impact of policies and programs on household behavior and welfare (Grosh et al. 2000). The first surveys were implemented in Côte d'Ivoire and Peru. Other early surveys followed a similar format, although considerable variation has been introduced over time.

The household questionnaire forms the heart of the LSMS survey. Typically, it includes a health module that provides information on (i) health-related behavior; (ii) utilization of health services; (iii) health expenditures; (iv) insurance status; and (v) access to health services. The level of detail of the health section has, however, varied across surveys. Complementary data are typically collected through community and price questionnaires. In addition, detailed service provider (health facility or school) data have been collected in some LSMS surveys. The facility surveys have been included to provide complementary data primarily on prices of health care and medicines and health care quality.

Further information, including a list of past and ongoing surveys, survey reports, questionnaires, and information on how to access the data, can be found at http://www.worldbank.org/lsms/.

UNICEF multiple indicator cluster surveys

The multiple indicator cluster surveys (MICS) were developed by UNICEF and others in 1998 to monitor the goals of the World Summit for Children. By 1996, sixty developing countries had carried out stand-alone MICS and another 40 had incorporated some of the MICS modules into other surveys.

The early experience with MICS resulted in revisions of the methodology and questionnaires. These revisions drew on the expertise and experience of many organizations, including WHO, UNESCO, ILO, UNAIDS, the United Nations Statistical Division, CDC Atlanta, MEASURE (USAID), and academic institutions.

The MICS typically include three components: a household questionnaire, a women's questionnaire (15–49 years), and a child (under 5 years) questionnaire. The precise content of questionnaires has varied somewhat across countries. Household questionnaires often cover education, child labor, maternal mortality, child disability, water and sanitation, and salt iodization. The women's questionnaires have tended to include sections on child mortality, tetanus toxoid, maternal health, contraceptive use, and HIV/AIDS. Finally, the child questionnaire covers birth registration, vitamin A, breast-feeding, treatment of illness, malaria, immunizations, and anthropometry.

Further information, including a list of past and ongoing surveys, survey reports, questionnaires, and information on how to access the data can be found at http://www.childinfo.org/index2.htm.

WHO World Health Survey

WHO has developed a World Health Survey (WHS) to compile comprehensive baseline information on the health of populations and on the outcomes associated with the investment in health systems. These surveys have been implemented in 70 countries across the full range of development in collaboration with the people involved in routine HIS. The overall aims of the WHS are to examine the way populations report their health, understand how people value health states, and measure the performance of health systems in relation to responsiveness. In addition, it addresses various issues such as health care expenditures, adult mortality, birth history, various risk factors, and the like.

In the first stage, the WHS targets adult individuals living in private households (18 years or older). A nationally representative sample of households is drawn, and adult individuals are selected randomly from the household roster. Sample sizes vary from 1,000 to 10,000 individuals.

The content of the questionnaires varies across countries but, in general, covers general household information, geocoding, malaria prevention, home care, health insurance, income indicators, and household expenditure (including on health). In addition, a specific module is administered to household members who are trained or are working as health professionals. This module covers a limited set of issues, including occupation, location of work, hours of work, main activities in work, forms and amount of payment, second employment, reasons for not working (if applicable), and professional training. The individual questionnaire includes sections on sociodemographic characteristics, health state descriptions, health state valuations, risk factors, mortality, coverage, health system responsiveness, and health goals and social capital.

Further information, including country reports and questionnaires can be found at http://www.who.int/healthinfo/survey/en/index.html.

WHO multicountry evaluation of the integrated management of childhood illnesses

Currently, WHO is coordinating a multicountry evaluation (MCE) of the integrated management of childhood illnesses (IMCI).[6] Integrated survey instruments for costs and quality have been developed and implemented (or are being implemented) in Bangladesh, Tanzania, Peru, and Uganda. The purpose of the MCEs is to (i) document the effects of IMCI interventions on health workers' performance, health systems, and family behaviors; (ii) determine whether, and to what extent, the IMCI strategy as a whole has a measurable impact on health outcomes (reducing under-5 morbidity and mortality); (iii) describe the cost of IMCI implementation at national, district, and health facility levels; (iv) increase the sustainability of IMCI and other child health strategies by providing a basis for improving implementation; and (v) support planning and advocacy for childhood interventions by ministries of health in developing countries and national and international partners in development. Worldwide there are 30 countries at different stages of implementation of IMCI, among which Uganda, Peru, Bangladesh, and Tanzania will participate in the MCE.

Further information, including country reports, questionnaires, and how to access data can be found at http://www.who.int/imci-mce/.

[6] The Integrated Management of Childhood Illnesses (IMCI) Strategy was developed by WHO and UNICEF to address five leading causes of childhood mortality, namely, malaria, pneumonia, diarrhea, measles, and malnutrition. The three main components addressed by the strategy are improved case management, improved health systems, and improved family and community practices.

RAND surveys

RAND has supported the design and implementation of Family Life Surveys (FLS) in developing countries since the 1970s. Currently available country surveys include Indonesia (1993, 1997, 1998, 2000), Malaysia (1976–7, 1988–9), Guatemala (1995), and Bangladesh (1996). Further information about these surveys and information on how to access the data can be found at http://www.rand.org.

Indonesia Family Life Survey The Indonesia Family Life Survey (IFLS) is an ongoing, multitopic longitudinal survey. It aims to provide data for the measurement and analysis of a range of individual- and household-level behaviors and outcomes. It includes indicators of economic well-being, education, migration, labor market outcomes, fertility and contraceptive use, health status, use of health care and health insurance, intrahousehold relationships, and participation in community activities. In addition, community-level data are collected. These include detailed surveys of service providers (schools and health care providers) in the selected communities. The first wave of the survey (IFSL1) was conducted in 1993/4, covering approximately 7,000 households. The IFLS2 and IFLS2+ were conducted in 1997 and 1998, and a further wave (IFLS3) in 2000.

Malaysian Family Life Surveys The Malaysian Family Life Surveys were conducted in 1976/7 and 1988. The surveys contain extensive histories on employment, marriage, fertility, and migration. Respondents in the first wave were followed in a second wave, and a refreshment sample was added.

Matlab Health and Socioeconomic Survey The Matlab Health and Socioeconomic Survey was implemented in 1996 in Matlab, a rural region in Bangladesh in which there is an ongoing prospective demographic surveillance system. The general focus of the survey was on issues relating to health and well-being for rural adults and the elderly, including the effects of socioeconomic characteristics on health status and health care utilization; health status, social and kin network characteristics, and resource flows; and community services and infrastructure. The study included a survey of individuals and households, a specialized out-migrant survey (sample of individuals who had left the households of the primary sample since 1982), and a community provider survey.

Guatemalan Survey of Family Health The Guatemalan Survey of Family Health is a single cross-section survey that was conducted in rural communities in 4 of Guatemala's 22 departments. The survey was fielded in 1995.

University of North Carolina surveys

The Carolina Population Center at the University of North Carolina at Chapel Hill has been involved in a range of different data collection exercises. Much of the data are publicly available. Information can be found at http://www.cpc.unc.edu/projects/projects.php.

Cebu Longitudinal Health and Nutrition Surveys The Cebu Longitudinal Health and Nutrition Survey is a study of a cohort of Filipino women who gave birth between May 1, 1983, and April 30, 1984, and were reinterviewed, with their children, at three subsequent points in time until 1998/9.

China Health and Nutrition Survey The China Health and Nutrition Survey is a six-wave longitudinal survey conducted in eight provinces of China between 1989 and 2004. It provides a wealth of detailed information on health and nutrition of adults and children, including physical examinations.

Nang Rong (Thailand) projects The Nang Rong projects represent a major data collection effort that was started in 1984 with a census of households in 51 villages. The villages were resurveyed in 1988 and again in 1994/5. New entrants were interviewed, and a subsample of out-migrants was followed.

Sample design and the analysis of survey data

Survey data provide information on a subset of a population—a sample. If the sample is appropriately selected, it provides the basis for drawing inferences about the target population, for example, all children under five in a particular country. A sample is selected from a sampling frame, which is a list of sampling units (e.g., households).[7] In a probability sampling design, every element in the sampling frame has a known, nonzero chance of being selected into the survey sample. This is not true with nonprobability methods, such as quota or convenience sampling and random walks.

The most straightforward way of selecting a sample is by simple random sampling–sampling units are selected from the sampling frame with equal probability.[8] In many cases, a single-stage random sampling design is impractical. This may be so because of the difficulty in drawing up a complete list for the entire target population, because of concern that the sample would contain "too few" members of some subpopulations, or because of high costs and logistical constraints in visiting a randomly selected sample. Because of these and other concerns, many surveys have what is referred to as a complex survey design. Three factors that arise from the sample design have important implications for data analysis (Deaton 1997).

- **Stratification** Stratification is the process by which the population is divided into subgroups or subpopulations, and sampling is then done separately for each subpopulation. Stratification can be done on the basis of geography, level of urbanization, socioeconomic zones or administrative areas, and so forth. Stratification is used when there is an expectation of heterogeneity between different subpopulations. It can then reduce sampling error and ensures that representative estimates can be produced for each strata.

[7] The sampling units are often the same as the members of the target population, but that is not always the case. For example, because it would be very difficult to construct a list of all children under 5 in any country, it may be more convenient to consider households as the sampling units and then to include all children under 5 from the selected households in the sample.

[8] In theory, simple random sampling is done with replacement of units after each draw. In practice, sampling is usually without replacement, and there should be a slight adjustment to the standard errors to correct for this (see, for example, Deaton [1997]).

- **Cluster sampling** A cluster is a naturally occurring unit or grouping within the population (e.g., enumeration areas). Cluster sampling entails randomly selecting a number of clusters and then including all or a random selection of units within the cluster. In multistage cluster sampling, further clusters are selected from within the first cluster. For example, enumeration areas may be the primary sampling unit, followed by households as secondary sampling units, and individuals as the final unit. Cluster sampling is useful because it reduces the informational requirement in the sampling process (a complete list of sampling units is required only for selected clusters) and because it can significantly reduce the costs of survey implementation. However, if there is a great deal of homogeneity within clusters, but heterogeneity between clusters, cluster sampling can substantially increase standard errors.
- **Unequal selection probabilities** In many surveys, different observations may have different probabilities of selection. This may be the consequence of stratification or other sample design decisions. In this case, it is necessary to weight each observation in the analysis to generate unbiased estimates of parameters of interest. The weights are equal (or proportional) to the inverse of the probability of being sampled. As a consequence, the weight for a specific observation can be interpreted as the number of elements in the population that the observation represents. In other words, if an element has a very small probability of selection relative to other elements, it should be weighted more heavily in the analysis.

The importance of taking sample design into account: an illustration

Many software packages have preprogrammed features for the analysis of complex survey data. That is the case, for example, with Stata, SPSS, and EpiInfo. For example, in Stata, survey commands can be used for descriptive analysis (e.g., `svydes`, `svymean`, `svyprop`, `svytotal`, `svytab`), estimation (e.g. `svyreg`, `svyprobit`, `svylogit`, `svymlogit`, `svyoprobit`, `svypois`), and postestimation testing (e.g., `svytest`).[9] Issues in the multivariate analysis of complex survey data are discussed in greater detail in chapter 10. Here, we simply illustrate the importance of taking sample design into account when making inferences about a population mean.

The following example is based on the 1997 Mozambique Living Standards and Measurement Survey. The survey sample was selected through a three-stage process, with stratification by province (11 provinces—the variable `province`) and area (urban/rural—`urban`), primary sampling at the locality level (`locality`), followed by sampling of households within each locality. Sampling weights are recorded in the variable `wgt`. In surveys in which samples are stratified along more than one dimension, a stratification variable (with a unique value for each strata) typically has to be constructed by the analyst. For example in the Mozambique data,

[9] For most Stata commands, adjustment for unequal sampling probabilities can be made by applying the weights option, for example, `[pw=weight]`. Standard errors can also be adjusted for cluster design by the option `cluster()`. Nonsurvey commands do not handle stratified sampling, however.

there are 21 separate strata (two strata (urban/rural) for each of the 11 provinces, except for Maputo City Province, which is only urban). This stratification variable can be easily constructed in Stata using the `group` function of the `egen` command.

```
egen strata = group(province urban)
```

We now have the three variables—`wgt`, `strata`, and `locality`—required to take sample design fully into account in the analysis. Here, we consider how child immunization rates, estimated from a dummy variable `vacc` indicating whether

Table 2.4 *Child Immunization Rates by Household Consumption Quintile, Mozambique, 1997*

Effect on Point Estimates and Standard Errors of Taking Sample Design into Account

A

pweight: -
strata: -
psu: -

Quintile	*Mean*	*s.e.*	*Deff*
Poorest	0.545	0.014	1.000
2	0.659	0.014	1.000
3	0.708	0.013	1.000
4	0.805	0.011	1.000
Richest	0.892	0.008	1.000
Total	**0.728**	**0.006**	**1.000**
n	6,447		
No. strata	1		
No. PSUs	6,447		

B

pweight: *wgt*
strata: -
psu: -

Quintile	*Mean*	*s.e.*	*Deff*
Poorest	0.531	0.017	1.694
2	0.629	0.019	2.196
3	0.621	0.019	2.117
4	0.708	0.024	3.416
Richest	0.843	0.014	1.488
Total	**0.654**	**0.009**	**2.138**
n	6,447		
No. strata	1		
No. PSUs	6,447		

C

pweight: *wgt*
strata: *strata*
psu: -

Quintile	*Mean*	*s.e.*	*Deff*
Poorest	0.531	0.017	1.630
2	0.629	0.019	2.164
3	0.621	0.019	2.075
4	0.708	0.024	3.366
Richest	0.843	0.014	1.456
Total	**0.654**	**0.008**	**1.942**
n	6,447		
No. strata	21		
No. PSUs	6,447		

D

pweight: *wgt*
strata: *strata*
psu: *locality*

Quintile	*Mean*	*s.e.*	*Deff*
Poorest	0.531	0.028	4.469
2	0.629	0.033	6.577
3	0.621	0.026	4.014
4	0.708	0.029	5.092
Richest	0.843	0.018	2.485
Total	**0.654**	**0.017**	**8.313**
n	6,447		
No. strata	21		
No. PSUs	273		

Source: Authors.

a child is immunized, vary across consumption quintiles (`quint`). Four different cases are considered:

A. sample design not taken into account

```
svyset
```

B. sample weights taken into account

```
svyset [pw=wgt]
```

C. sample weights and stratification taken into account

```
svyset [pw=wgt], strata(strata)
```

D. sample weights, stratification, and clustering taken into account

```
svyset locality [pw=wgt], strata(strata)
```

In each case, the `svyset` command is followed by

```
svy: mean vacc, over(quint)
```

As can be seen from table 2.4, the application of weights has a substantial impact on both point estimates and standard errors. In this application, taking stratification into account reduces the standard errors only slightly, whereas taking clustering into account increases the standard errors substantially. This illustrates that application of weights is not sufficient to correct for the sample design. It corrects the point estimates, but not the standard errors, confidence intervals, and test statistics.

These effects are described by the design effect (`deff`), which is a measure of how the survey design affects variance estimates. deff is calculated as the design-based variance estimate divided by an estimate of the variance that would have been obtained if a similar survey had been carried out using simple random sampling. It is obtained from the command `estat effects` following `svy`.

References

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Baltimore, MD: Published for the World Bank [by] Johns Hopkins University Press.

Grosh, M. E., P. Glewwe, and World Bank. 2000. *Designing Household Survey Questionnaires for Developing Countries : Lessons from 15 Years of the Living Standards Measurement Study.* Washington, DC: World Bank.

Murray, C., and L. Chen. 1992. "Understanding Morbidity Change." *Population and Development Review* 18(3): 481–503.