Data for health equity analysis

Requirements, sources, and issues in analysis

Introduction

The first step in health equity analysis is to identify appropriate data and to understand its potential and limitations. This Technical Note 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 note 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, for health equity analysis, survey data tend to have the greatest potential for assessing and analyzing different aspects of health equity. With this in mind, this note 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 for health equity analysis

Health outcomes and health related behavior

Data on health outcomes comprise 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).

Table – A classification of morbidity measures

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

For most of these measures, data are not collected on a routine basis, and can only be obtained 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, have included trained health workers in enumerator teams, and collected 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 that merit attention in the analysis of health equities include: (i) behavior with an effect on health status (smoking, drinking, diet); (ii) sexual practices; and, (iii) household level behavior (cooking practices, waste disposal, sanitation, sources of water). Some data on health service use is collected through routine information systems and population censuses (e.g. immunizations). However, more detailed data are only available through surveys.

**Living standards or socioeconomic status**

The premise of health equity analysis is to assess and understand how health outcomes or health related behaviors vary along some measure of socioeconomic status or living standards. Health related information must therefore be complemented by data on living standards. As discussed in detail in TN#4, 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. That said, the choice of measure may have an impact on analytical 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—e.g. income or consumption in terms of dollars or another currency—numbers convey comparable information about magnitude. This can be contrasted with ordinal measures, which can only be used to rank individuals or households, and do not permit comparisons across units. Some forms of health equity analysis require a cardinal measure of living standards. This is the case, for example, with multivariate analysis, analysis of financing progressivity, and poverty impact of health financing or health events. But in some cases, a ranking of households by some measure of living standards suffices. For example, both benefit incidence analysis and the construction of concentration curves and indices do not rely on ordinal measures of living standards.

**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 we are interested in performing multivariate analysis of health related variables to better understand why observed inequalities arise, we need data on *community, household, and individual characteristics*. This could include, for example, availability and characteristics of health care providers, environmental and climactic characteristics of the community, housing characteristics, education, etc. In some cases, it may be necessary to include seemingly unrelated variables in multivariate analysis to deal with potential endogeneity.

Complementary data are also required for benefit-incidence analysis. The primary requirement is data on *unit subsidies* on health services. This information tends to be based on public expenditure data, but in some cases, more detailed cost information is available. Benefit-incidence analysis may also be broadened if other complementary data—e.g. on gender, ethnic group, or geographic location—are available. Similarly, analysis of health financing fairness and progressivity depends on detailed data on *user payments* for health care.
Table: Health equity analysis and data needs

<table>
<thead>
<tr>
<th></th>
<th>Health variables</th>
<th>Utilization variables</th>
<th>Living standards measure (ordinal)</th>
<th>Living standards measure (cardinal)</th>
<th>Unit subsidies</th>
<th>User payments</th>
<th>Background variables</th>
<th>TYPICAL SURVEYS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health inequality</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSMS, DHS, MICS, etc.</td>
</tr>
<tr>
<td>Equity in utilization</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSMS, DHS, MICS, etc.</td>
</tr>
<tr>
<td>Multivariate analysis</td>
<td>✓ or</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSMS, Household Budget Surveys</td>
</tr>
<tr>
<td>Benefit-incidence analysis</td>
<td>✓</td>
<td>✓</td>
<td>(✓)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LSMS, DHS, MICS, etc.</td>
</tr>
<tr>
<td>Health financing</td>
<td></td>
<td></td>
<td></td>
<td>(✓)</td>
<td></td>
<td></td>
<td></td>
<td>LSMS, Household Budget Surveys</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data sources and their limitations

Household surveys and other non-routine 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 multi-purpose surveys, with a focus on a broad set of demographic and socio-economic issues, while other surveys focus explicitly on health. Although surveys only cover a subset of the population, they are representative of the population as a whole (or whatever target population that is defined for the survey) and have the benefit of permitting more detailed data collection than a comprehensive census. Although many surveys are conducted on an *ad hoc* basis, there are an increasing number of multi-round 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). The *Living Standards Measurement Study* is different from the other surveys in that it collects detailed expenditure and/or income data. In that sense, they are a form of *Household Budget Surveys*. Most countries implement *Household Budget Surveys* in some form or other on a semi-regular basis. A core objective of these surveys is to capture the essential elements of the

1 Some surveys, in particular the DHS and some budget surveys, are repeated on regular basis, and can in that sense be considered “semi-routine” data.

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

---

Data for health equity analysis
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—e.g. on health, nutrition, and other topics—making them ideal for detailed analysis of the relationship between economic status and health variables. Aside from large-scale household surveys, there is often a wealth of other non-routine 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 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, where specific providers are typically not identified, and where recall periods of up to four weeks can result in considerable measurement error. On the other hand, unlike a household survey, an exit poll only provides information about users of health services.

**Box: Sampling and non-sampling bias in survey data**

When analyzing survey data, analysts must be aware of potential sources of sampling and non-sampling bias in analysis. Sampling bias refers to a situation when 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 sample drawn from users of health facilities. This is because different groups in the population use health facilities to different degrees—e.g. 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 non-sampling bias which can arise in the process of survey implementation. For example, non-response or measurement errors may be systematically related with variables of interest—e.g. non-response about utilization of health services may be higher among the poor. If this were the case, analysts must 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 non-response. Other potential sources of non-sampling bias include errors in recording or data entry.

**Routine data: Health Information Systems and Censuses**

There are different forms of routine data that may be amenable 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 primarily used for management purposes, e.g. for planning, needs assessments, resource allocation and quality assessments. However, in some contexts, HIS data include demographic or socio-economic 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 socio-economic 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* comprise another form of routine data. Censuses are an important source of data for planning and monitoring of population issues and socio-economic and environmental trends, both in 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 homeless and migrant population. Population censuses have been conducted in most countries in recent years. Census data often contain only limited information on health and living standards, but have sometimes been used to study health inequalities by linking it to HIS data. For example, socio-economic differences in disease incidence and hospitalization has been studies by linking cause-of-death or hospital discharge records with census data. In the US, there have also been efforts to link public health surveillance data with area-based socio-economic measures based on geo-coding. While 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 profile 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, due to spatial differences in the coverage of health facility infrastructure, routine data is 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, where routine data are used for management purposes, there may exist incentives for staff to record information inaccurately.

---

3 Information about dates of censuses in different countries can be found on http://unstats.un.org/unsd/demographic/census/cendate/index.htm
### Table: Data sources and their limitations

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

### 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. The design of the DHS drew on the experiences of the World Fertility Surveys⁴ (WFS) and the Contraceptive Prevalence Surveys (CPS), but included an expanded set of indicators in the areas of

---

⁴ The WFSs were a collection of internationally comparable surveys of human fertility conducted in 41 developing countries in the late seventies and early eighties. The project was conducted by the International Statistical Institute (ISI), with funding from USAID and UNFPA.
population, health, and nutrition.\(^5\) Demographic and Health Surveys (DHS) are nationally representative household surveys that provide on a broad range of issues relating to population, health, and nutrition. The sample size typically ranges from 5,000 and 30,000 households.

The standard DHS survey consists of a household questionnaire and a women’s questionnaire (ages 15-49). The core questionnaire of the DHS survey emphasizes basic indicators and is standardized across countries. The household questionnaire cover basic demographic data for all household members, household and dwelling characteristics, and nutritional status of young children and women age 15-49. 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 standard DHS surveys, interim surveys are sometimes implemented to collect information on a reduced set of performance monitoring indicators. These surveys have smaller sample size and are often conducted between rounds of DHS surveys. In addition, many of the DHSs have included tools to collect community level data (Service Availability Modules (SAM)). More recently, detailed facility surveys—Service Provision Assessments (SPAs)—have been implemented alongside household surveys with a view to provide 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 (LSMS)**

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 multi-topic surveys, designed to permit four types of analysis: (i) simple descriptive statistics on living standards; (ii) monitoring of poverty and living standards over time; (iii) describing the incidence and coverage of government programs; and (iv) measuring 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 LSMS survey. It typically includes a health module that provides information on (i) health related behavior; (ii) the utilization of health services; (ii) health expenditures; (iv) insurance status; (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 on http://www.worldbank.org/lsms/.

**UNICEF Multiple Indicator Clusters Surveys**

The multiple indicator cluster survey (MICS) were developed by UNICEF and others in 1998 to monitor the goals of the World Summit for Children. By 1996, 60 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.

---

\(^5\) In 1997 DHS changed name to *DHS+* to reflect the integration of DHS activities under the MEASURE programme. Under this mandate, *DHS+* is charged with collecting and analysing demographic and health data for regional and national family planning and health programs.
The MIC surveys 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 have varied somewhat across countries. Household questionnaires often cover education, child labor, maternal mortality, child disability, water and sanitation, and salt iodization. The women’s questionnaire 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, care 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 on http://www.childinfo.org/index2.htm.

**WHO World Health Surveys**

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

In the first stage, WHSs target 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. The sample design varies somewhat across countries, but in most contexts, the sample consists of approximately 5000 individuals.

The content of the questionnaires also vary across contexts. However, in general, the household questionnaire covers general household information, geocoding, malaria prevention, home care, health insurance, income indicators, household expenditure (including on health). In additions, a specific module is administered to household members that 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/work, reasons for not working (if applicable), professional training. The individual questionnaire includes sections on socio-demographic characteristics, health state descriptions, health state valuations, risk factors, mortality, coverage, health system responsiveness, and health goals and social capital.

Further information, including survey reports, questionnaires, and information on how to access the data can be found on http://www.who.int/whs.

**WHO Multi-Country Evaluation of IMCI**

Currently, the WHO is coordinating a multi-country evaluation (MCE) of the Integrated Management of Childhood Illnesses (IMCI). The 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 five 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 the improvement of 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.

---

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 (1976-77, 1988-89), Guatemala (1995), and Bangladesh (1996). Further information about these surveys, and information on how to access the data can be found on http://www.rand.org.

The Indonesia Family Life Survey
The Indonesia Family Life Survey (IFLS) is an ongoing, multi-topic longitudinal survey. It aims to provide data for the measurement and analysis of a range of individual- and household-level behaviors and outcomes. The survey has collected data at individual and household level, including indicators of economic well-being, education, migration, labor market outcomes, fertility and contraceptive use, health status, use of health care and health insurance, intra-household 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 (IFLS1) was conducted in 1993/94, covering approximately 7000 households. The IFLS2 and IFLS2+ were conducted in 1997 and 1998, and a further wave (IFLS3) was planned for 2000.

Malaysian Family Life Surveys (MFLS)
The MFLS was 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 the subsequent waves; in the second wave, a refreshment sample was added.

Matlab Health and Socio-Economic Survey (MHSS)
The MHSS was implemented in 1996, in Matlab, a rural region in Bangladesh. The general focus of the survey was on issues relating to health and well-being for rural adults and the elderly, including the effects on health status and health care utilization of socio-economic characteristics; health status, social and kin network characteristics and resource flows; community services and infrastructure. The study included a survey of individuals and households, a specialized outmigrant 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 (EGSF)
The EGSF is a single cross section survey which 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 (CPC) at The University of North Carolina at Chapel Hill has been involved in a range of different data collection exercises. Much of the data is publicly available. Information can be found on http://www.cpc.unc.edu/dataarch/iprimary/.

Cebu Longitudinal Health and Nutrition Surveys
Surveys have been conducted by a team of researchers from the United States and the Philippines. The Cebu Longitudinal Health and Nutrition Survey is an ongoing study of a cohort of Filipino women who gave birth between May 1, 1983 and April 30, 1984 and have been re-interviewed periodically since then.

China Health and Nutrition Survey
The China Health and Nutrition Survey was conducted in 1989 and 1991 in 8 provinces in China and 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 1994/95. New entrants were interviewed and a subsample of out migrants were followed.
Sample design and the analysis of survey data

On overview of sample design issues

The premise of surveys is that data are collected from a subset of a population (a sample). If the sample is appropriately selected, it provides the basis for drawing inferences about the sample’s target population. The members of this target population—e.g. all children under five in a particular country—comprise the larger universe that the sample is chosen to represent. The basis for selecting a sample is a sampling frame, which is a list of sampling units. Probability sampling design means that every single element in the sampling frame has a known and non-zero chance of being selected into the survey sample. This can be contrasted with non-probability methods such as quota or convenience sampling and random walks.

The most straightforward way of selecting a sample is to select sample units from the sampling frame with equal probability. In this case, each sampling unit has an equal probability of being included in the sample. In many cases, single stage random sampling design is impractical. This may be 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 sub-populations, or because of high costs and logistical constraints in visiting a randomly selected sample of sampling units. Because of these and other concerns, many surveys have what is referred to as a complex survey design. There are three factors that arise from the sample design that have important implications for data analysis.

Stratification

Stratification is the process by which the population is divided into subgroups or subpopulations, and sampling is then done separately for each sub-population. Stratification can be done on the basis of geography, level of urbanization, socio-economic zones or administrative areas, etc. Stratification is used when there is an expectation of heterogeneity between different sub-populations. Stratification is also done to reduce sampling error, and to ensure that representative estimates can be produced for each strata.

Cluster sampling

A cluster is a naturally occurring unit or grouping within the population (e.g. enumeration areas). Cluster sampling entails selecting a number of clusters, and then including all or a random selection of units within the cluster. In multi-stage cluster sampling, further clusters are selected within the first cluster. For example, enumeration areas may be the primary sampling unit, followed by households as secondary sampling units, and individuals as final unit. Cluster sampling is useful because it reduces the informational requirement in the sampling process (a complete list of sampling units is only required for selected clusters), and because it can significantly reduce the costs of survey implementation. However, if there is a lot 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 consequences of stratification or other sample design decisions. In this case, it is necessary to weight each observation in the analysis in order to generate unbiased estimates of variables 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.

---

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 five in any country, it may be more convenient to consider households as the sampling units, and then to include all children under five from the selected households in the sample.

8 Simple random selection can be done with or without replacement…
The importance of sample design: an illustration

Many software packages have pre-programmed features for the analysis of complex survey data. This 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 post-estimation testing (e.g. `svytest`). Issues in the analysis of complex survey data is discussed in greater detail in Technical Note #10 and #11. Here, we simply provide an illustration of the importance of taking sample design into account.

The following example is based in the 1997 Mozambique household survey (IAF). The IAF sample was selected through a three-stage process, with stratification by province (11 provinces) and area (urban/rural), selection of primary sampling units (locality variable), followed by selection of households within each locality. The resultant sampling weights are recorded in the variable `wgt`. In surveys where 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, 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 using the `group` function of the `egen` command.

\[\text{egen strata = group(provin urban)}\]

We now have the three variables—`strata`, `wgt`, and `strata`—required to take sample design fully into account in the analysis. Here, we consider how child immunizations varies across consumption quintiles. Four different cases are considered:

(A) sample design not taken into account
   \[\text{svymean vacc, by(quint)}\]

(B) sample weights taken into account
   \[\text{svymean vacc [pw=wgt], by(quint)}\]

(C) sample weights and stratification taken into account
   \[\text{svymean vacc [pw=wgt], by(quint) strata(strata)}\]

(D) sample weights, stratification, and clustering taken into account
   \[\text{svymean vacc [pw=wgt], by(quint) strata(strata) psu(locality)}\]

As can be seen from the table, the use of weights have a substantial impact on both point estimates and standard errors. As expected, the stratification does not affect the point estimates, but reduces standard errors somewhat. In contrast, taking clustering into account increases the standard errors substantially. In other words, only using weights is not sufficient. It ensures that proper design-based point estimates are achieved, but standard errors, confidence intervals, and test statistics are most likely wrong.

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 we would have obtained if we had carried out a similar survey using simple random sampling.

---

9 For many survey commands, it is possible to use standard stat commands with `pweight[]` option to account for unequal sampling probabilities and `cluster()` or robust to reflect the cluster design of the sample. The resulting point estimates will correspond. Non-survey commands do not handle stratified sampling. As stratification tends to reduce standard errors.

10 The respective stratification variables (provin and urban) have to be in the format of numerical variables, with a unique value for each category.
### Table: Sample design, estimation, and inference

**(A)**

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Mean</th>
<th>s.e.</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>poorest</td>
<td>0.545</td>
<td>0.014</td>
<td>1.000</td>
</tr>
<tr>
<td>2</td>
<td>0.659</td>
<td>0.014</td>
<td>1.000</td>
</tr>
<tr>
<td>3</td>
<td>0.708</td>
<td>0.013</td>
<td>1.000</td>
</tr>
<tr>
<td>4</td>
<td>0.805</td>
<td>0.011</td>
<td>1.000</td>
</tr>
<tr>
<td>richest</td>
<td>0.892</td>
<td>0.008</td>
<td>1.000</td>
</tr>
<tr>
<td>Total</td>
<td>0.728</td>
<td>0.006</td>
<td>1.000</td>
</tr>
</tbody>
</table>

n = 6447

No. strata = 1

No. PSUs = 6447

**Notes:**
- Pweight: `-`
- Strata: `-`
- Psu: `-`

***(B)***

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Mean</th>
<th>s.e.</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>poorest</td>
<td>0.531</td>
<td>0.017</td>
<td>1.694</td>
</tr>
<tr>
<td>2</td>
<td>0.629</td>
<td>0.019</td>
<td>2.196</td>
</tr>
<tr>
<td>3</td>
<td>0.621</td>
<td>0.019</td>
<td>2.117</td>
</tr>
<tr>
<td>4</td>
<td>0.708</td>
<td>0.024</td>
<td>3.416</td>
</tr>
<tr>
<td>richest</td>
<td>0.843</td>
<td>0.014</td>
<td>1.488</td>
</tr>
<tr>
<td>Total</td>
<td>0.654</td>
<td>0.009</td>
<td>2.138</td>
</tr>
</tbody>
</table>

n = 6447

No. strata = 1

No. PSUs = 6447

**Notes:**
- Pweight: `wgt`
- Strata: `-`
- Psu: `-`

***(C)***

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Mean</th>
<th>s.e.</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>poorest</td>
<td>0.531</td>
<td>0.017</td>
<td>1.630</td>
</tr>
<tr>
<td>2</td>
<td>0.629</td>
<td>0.019</td>
<td>2.164</td>
</tr>
<tr>
<td>3</td>
<td>0.621</td>
<td>0.019</td>
<td>2.075</td>
</tr>
<tr>
<td>4</td>
<td>0.708</td>
<td>0.024</td>
<td>3.366</td>
</tr>
<tr>
<td>richest</td>
<td>0.843</td>
<td>0.014</td>
<td>1.456</td>
</tr>
<tr>
<td>Total</td>
<td>0.654</td>
<td>0.008</td>
<td>1.942</td>
</tr>
</tbody>
</table>

n = 6447

No. strata = 21

No. PSUs = 6447

**Notes:**
- Pweight: `wgt`
- Strata: `strata`
- Psu: `-`

***(D)***

<table>
<thead>
<tr>
<th>Quintile</th>
<th>Mean</th>
<th>s.e.</th>
<th>Deff</th>
</tr>
</thead>
<tbody>
<tr>
<td>poorest</td>
<td>0.531</td>
<td>0.028</td>
<td>4.469</td>
</tr>
<tr>
<td>2</td>
<td>0.629</td>
<td>0.033</td>
<td>6.577</td>
</tr>
<tr>
<td>3</td>
<td>0.621</td>
<td>0.026</td>
<td>4.014</td>
</tr>
<tr>
<td>4</td>
<td>0.708</td>
<td>0.029</td>
<td>5.092</td>
</tr>
<tr>
<td>richest</td>
<td>0.843</td>
<td>0.018</td>
<td>2.485</td>
</tr>
<tr>
<td>Total</td>
<td>0.654</td>
<td>0.017</td>
<td>8.313</td>
</tr>
</tbody>
</table>

n = 6447

No. strata = 21

No. PSUs = 273

**Notes:**
- Pweight: `wgt`
- Strata: `strata`
- Psu: `locality`