

Indispensable Information —

DATA COLLECTION AND INFORMATION
MANAGEMENT FOR HEALTHIER COMMUNITIES

By Peter A. Tatian
March 2000

National Neighborhood Indicators Partnership

THE URBAN INSTITUTE

The nonpartisan Urban Institute publishes studies, reports, and books on timely topics worthy of public consideration. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders.

ACKNOWLEDGMENTS

This handbook was written at the request of the Accelerating Community Transformation (ACT) project to serve as the seventh in its series of Learning Modules. The development of the handbook was supported jointly by ACT and the National Neighborhood Indicators Partnership (NNIP).

ACT is a five-year initiative of the Health Forum to help increase the capacity of community partnerships to achieve and demonstrate measurable improvements in community well-being and quality of life. Through its Learning Modules, ACT provides its partner communities with information they can use to develop an outcomes-based approach to community development. For more information on ACT, please visit the project's Web site at <http://www.healthforum.com/thfnet/act/act.htm>.

NNIP is a collaborative effort by the Urban Institute and local partners to further the development and use of neighborhood-level information systems in local policymaking and community building. All NNIP local partners have built locally self-sustaining information systems with integrated and recurrently updated information on neighborhood conditions in their cities. These systems facilitate the direct use of information by local government and community leaders to build the capacities of distressed urban neighborhoods. Current NNIP activities are sponsored by the Annie E. Casey Foundation and the Rockefeller Foundation. For more information on NNIP, please visit the Urban Institute's Web site at <http://www.urban.org/nnip/>.

The author would like to thank Neil Bania of Case Western Reserve University's Center on Urban Poverty and Social Change for his excellent and thoughtful comments on an earlier draft of this document.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	1
ORGANIZATION OF THIS HANDBOOK	1
CHAPTER 2: USING DATA AS A TOOL FOR UNDERSTANDING AND CHANGE	3
ASSESSING CONDITIONS AND TRENDS	5
<i>Example: Broward Benchmarks (Broward County, Florida)</i>	7
PLANNING AND IMPLEMENTING SPECIFIC IMPROVEMENT INITIATIVES	8
<i>Example: Targeting a First-Time Parents Program (Milwaukee)</i>	10
PARTNERING IN COMMUNITY-WIDE OR REGIONAL INITIATIVES	11
<i>Example: Social Services and Public Schools (Oakland, California)</i>	12
<i>Example: Cleveland Community-Building Initiative</i>	14
REFERENCES	15
CHAPTER 3: USING EXISTING DATA	17
CATALOG OF SECONDARY DATA SOURCES	21
CHAPTER 4: COLLECTING DATA NOT AVAILABLE ELSEWHERE	39
SELECTED DATA COLLECTION METHODS	42
CHAPTER 5: MANAGING AND WORKING WITH DATA	64
DATA FILE STRUCTURE AND ORGANIZATION	65
GEOGRAPHY	67

TOOLS FOR WORKING WITH DATA	69
<i>Spreadsheet Software</i>	<i>70</i>
<i>GIS Software.....</i>	<i>72</i>
<i>Statistical Software</i>	<i>73</i>
DATA ENTRY.....	73
BASIC DESCRIPTIVE STATISTICS	74
<i>Frequency Table for SEX.....</i>	<i>75</i>
DOCUMENTING DATA	75
REFERENCES.....	78
CHAPTER 6: PRESENTING INFORMATION EFFECTIVELY.....	80
TABLES	81
CHARTS.....	85
MAPS	89
REFERENCES.....	92
ANNEX A: DATA COLLECTION AND INFORMATION MANAGEMENT PLAN TEMPLATE ...	94
ANNEX B: GLOSSARY OF STATISTICAL TERMS.....	97

*Chapter 1***INTRODUCTION**

The purpose of this handbook is to serve as a reference and how-to guide for collecting, using, and disseminating data for community decisionmaking and information-sharing. It describes resources and methods available for implementing an outcomes-based approach to developing healthier communities. It outlines the ways to use data effectively to bring about community change, provide secondary data source references, illustrate quantitative and qualitative data collection methodologies, and describe elements of a research and data collection plan.

ORGANIZATION OF THIS HANDBOOK

The handbook is divided into six chapters. **Chapter I** is this introduction. The remaining chapters describe ways to use data effectively, resources for obtaining existing data and collecting new data, how to work with and manage data, and techniques for effective data presentation.

Chapter II, “Using Data as a Tool for Understanding and Change,” presents a conceptual framework for using data effectively. It describes four different ways data can be used to effect community change and draws on real-world examples from actual communities. It

also introduces the concept of an **outcome**, which is a statement of a goal to be achieved through some project or initiative, and the concept of an **indicator**, which is a number intended to measure some characteristic of people, organizations, or communities.

Chapter III, “Using Existing Data,” provides a reference of nationally available data sources that communities can use to learn more about current conditions and past trends across a variety of indicators. Obtaining data from existing sources, referred to as **secondary** data, is often less costly than collecting original data, but secondary data may be limited in scope and content. These and other issues surrounding the use of existing data sources are explored further in this chapter.

If there are no suitable existing sources of data, one may need to collect original data through surveys or other methods. **Chapter IV**, “Collecting Data Not Available Elsewhere,” describes several different methods for collecting original, or **primary**, data. Cost and other issues concerning various data collection methods are discussed.

To be used efficiently, data must be organized and managed. **Chapter V**, “Managing and Working with Data,” provides a basic introduction to data file structure and data management issues. This chapter includes sections explaining the various geographical areas that are frequently used by the U.S. Census Bureau and other data providers, and describes different types of software that can be used to work with data.

Chapter VI, “Presenting Information Effectively,” explains how to “tell a story” with data and provides some guidance on the best ways to format tables, charts, and maps to make a good presentation.

At the end of this handbook, you will find a template for a **data collection and information management plan** and a **glossary of statistical terms**.

This handbook is not intended to be an exhaustive resource on all the subjects listed above. Entire books have been written on each of these topics. Rather, it is intended to be a guide and a reference to help communities understand the basic concepts behind collecting, managing, and using data. Each chapter includes a list of resources to which you can turn for more information on each of the topics described in this handbook.

*Chapter 2***USING DATA AS A TOOL FOR
UNDERSTANDING AND CHANGE**

The use of data is not an end in itself. Rather, the value of data is in helping people and community groups to better accomplish the objectives they have set for themselves. If an organization collects lots of information on a community, but cannot use it in an effective way to bring about desirable change, then the data serve no useful purpose. But how can data be used effectively?

The value that good data bring to the community improvement process is that they provide a way to bring about **understanding** and to promote **change**. Let's suppose that people are concerned about the problem of teenage pregnancy in their community. Having reliable, up-to-date information on this subject will enable them to better understand the extent and the nature of it and its possible causes. Armed with this information, they will be able to make clear to others in their community why this issue is important and to discuss what can be done to change the situation. Once the community has taken action to address the problem, data will allow those concerned to monitor whether the expected changes are taking place. If not, they can reassess the situation and modify their actions.

Data can also be valuable in promoting **stability** and **preserving** what is good about a community. Traditionally many indicators have been expressed negatively, focusing on

problems within a community (indicators such as *poverty rate* or *percentage of abandoned buildings*), but indicators can also focus on the positive aspects of a community. Positive indicators attempt to highlight **community assets**, such as the *number of community-based organizations* or the *percentage of persons with a college education*.

In this chapter, we present the ways that data can be used to bring about community change as four activities. These four activities are:

1. **Assessing conditions and trends.** Public policies and conditions in cities are changing rapidly. Today's community leaders increasingly see the need to keep abreast of a broader range of information on conditions and trends that will affect what they can do to improve their neighborhoods.
2. **Planning and implementing specific improvement initiatives.** This means working out concrete plans of action around individual issues (e.g., teen pregnancy, employment, crime) that have been selected as high-priority concerns within the community. The work focuses on planning and implementing specific action programs, but it may also include mounting campaigns and giving presentations to influence the actions of outside public agencies and private interests.
3. **Partnering in community-wide or regional initiatives.** This involves community and neighborhood leaders working with others to support city-, county-, or metropolitan-wide programs and policy changes. Such initiatives, while not necessarily focused on the interests of each individual community, will benefit the entire region directly or indirectly.
4. **Comprehensive strategic planning.** This entails working with residents to help them determine overall goals for their community's development—visions of what they want their community to become—and develop courses of action (strategies) by which they can expect to achieve those goals.

These four activities require somewhat different sorts of information and somewhat different ways of using information, although in many cases the methods overlap. They should not be thought of as mutually exclusive, distinct categories, but rather as general viewpoints on how data can be used. In actual practice, communities would probably not do only one of these activities, but rather some combination. For example, while the Broward Benchmarks project (described below) is a good example of using data to monitor conditions and trends, the project also involves using information to make connections between different groups and organizations and to facilitate strategic planning.

The rest of this chapter will discuss the four activities in more detail and give examples of how each approach has been applied to specific cases in actual communities.¹

ASSESSING CONDITIONS AND TRENDS

To be effective in addressing community needs, you must keep abreast of a broad range of information on conditions and trends and continue to monitor these conditions over time. You need a systematic and efficient means to get “early warnings” of trends that might indicate the emergence of new problems or the opening up of new opportunities. You also want to assess a community’s strengths and assets so that they can be promoted and built upon.

To accomplish these goals, a number of communities have created systems of key indicators that are updated periodically. An **indicator** is a number intended to measure some social, health, economic, or other characteristic of people, organizations, or communities. Indicators can be used by a wide variety of people within the community to evaluate local conditions, design interventions, and engage in informed dialogues concerning these issues.

There is a vast array of data that can be used to describe conditions in a society, but not all data are indicators. Indicators are distinguishable from other data in at least two respects. First, they are measures purposely selected for tracking because they relate to important societal values and goals. Second, indicators must be expressed in a consistent form that permits comparison over time and, normally, between places.

For making comparisons, indicators are often expressed as ratios, rates, or percentages rather than as absolute numbers. For example, if you have **data** indicating that \$800,000 has been invested in housing renovation in one neighborhood over the past year and \$2.6 million in another, you cannot tell which neighborhood has had the best renovation record because you cannot compare these numbers directly. To make meaningful comparisons, you might bring these data together with information on the total number of housing units in the two neighborhoods and divide the amount invested by the number of housing units to create an indicator: *the dollar value of housing renovation per 1,000 housing units*.

Indicators can serve different functions in different contexts. Some indicators can **raise awareness** of a problem or issue, while others provide information that is more useful for **developing a solution** to the problem. For example, a high proportion of low-birth-weight babies born in a community can be an indication of some underlying problem. While this indicator may help identify the existence of a problem, and even mobilize support for some kind

¹ This categorization of data activities and several of the examples used to illustrate them have been adapted from *Mapping Your Community*, a publication of the U.S. Department of Housing and Urban Development. October 1997.

of action, it does not provide much guidance on creating an effective intervention. Other indicators, such as the percentage of birth mothers who received prenatal care or the percentage of young women who smoke, however, may suggest ways to address the problem.

Another important issue is the *interpretation* of indicators—of either their current value or their change over time. Is a high value for a particular indicator good or bad? If an indicator is decreasing over time, does that mean things are getting better or worse? For example, if the amount of renovation per 1,000 housing units is very high, that might indicate something positive (more investment taking place in the community) or something negative (there is too much dilapidated housing).

Establishing benchmarks for your indicators can help you and your community interpret their meaning. A **benchmark** is a specific value of an indicator selected as a reference point. Suppose, for example, that the crime rate in a particular neighborhood last year was 5.6 crimes per 1,000 persons. Is that good? Bad? By having a benchmark for this indicator, you have a way of comparing the value in particular areas and at particular points in time against some standard.

The choice of an appropriate benchmark depends on how the indicator is intended to be used. If you are interested in comparing neighborhoods within a city, you might want to take the overall city average as a benchmark. Or, to measure the change in an indicator over time, you might select an earlier year as a benchmark and compare subsequent years to this value. To see how a community measures up to some national standard, you might use the average value of the indicator for other similar communities in the United States.

Since communities are often interested in assessing conditions and trends across a wide range of substantive issues, it is helpful to organize indicators into separate domains. A **domain** is a collection of indicators that all deal with related topics. For example, a fairly comprehensive set of indicator domains developed by the Urban Institute includes the following:

1. Family, Children, and Youth
2. Education
3. Public Health
4. Economics: Income and Wealth
5. Economics: Consumption
6. Economics: Neighborhood Economy
7. Civic Life
8. Social/Cultural Life
9. Public Safety
10. Housing
11. Physical Environment

12. Transportation
13. Mobility/Turnover

Example: Broward Benchmarks (Broward County, Florida)

Broward County is located in southern Florida and has a population of about 1.4 million people. It includes the coastal communities of Ft. Lauderdale and Hollywood. In 1998, the Quality of Life Committee of The Coordinating Council of Broward issued its first *Broward Benchmarks* report, which was a collaboration of the public, the business community, universities, constituency groups, and the government. A second, updated report was released in 1999, and the Council intends to issue revised reports every year.

According to the Coordinating Council, the purpose of *Broward Benchmarks* is “to accurately frame where we were, where we are, and where we’re going.” In developing these indicators, the Council was starting a process of identifying which factors were the most important to people in defining the quality of life in Broward County. The indicators would then be used as a baseline to measure the progress being made toward improving the quality of life and to uncover those areas where improvement is needed.

In its first report, the Council identified a set of preliminary priority issues for the county. The Council subsequently sponsored open public forums and focus groups in different parts of the county to solicit public reaction to its selected priority issues. The 1999 report contained a revised set of priorities grouped into three “tiers”: top community priority projects, institutional collaborative initiatives, and basic support/process projects. The Council has started convening stakeholder groups that will identify the most appropriate ways to respond to each priority.

The latest *Broward Benchmarks* report contains past and current values for over 100 county-level indicators. The information comes from a wide array of sources, including the U.S. Census Bureau, state and local agencies, and opinion polls. The indicators are organized into seven domains: families and communities, safety, learning, health, economy, environment, and government. In addition to the county values, the report presents the most recent state-level values for the same indicators and county-level goals for 2000 and 2010.

The next phase of the project will be identifying different stakeholder groups to take “ownership” of the different *Benchmarks* domains, that is, to help keep them up to date and to adopt the indicators as a measure of progress. For example, the School Board has made recommendations on changes to the “Learning” section of the *Benchmarks* report and is building the indicators into its strategic planning process.

According to Richard Ogburn of the South Florida Regional Planning Council, one of the biggest impacts of the *Broward Benchmarks* has been bringing organizations together across existing institutional frameworks and getting people to focus on outcomes. Collaboration is hard,

Ogburn points out, but very important. The new relationships that have been forged are building greater confidence among the stakeholders so that they can work together to accomplish a common set of goals.

PLANNING AND IMPLEMENTING SPECIFIC IMPROVEMENT INITIATIVES

While assessing conditions and trends can help a community monitor the types of changes that are going on, people often wish to directly affect conditions and influence trends. One way of using data to bring about change is in planning and implementing specific community initiatives. This usually means working with community organizations and residents to design concrete plans of action that will address particular high-priority concerns, such as reducing crime, preparing residents for employment and helping them find jobs, or improving schools.

In their book *Basic Methods of Policy Analysis and Planning*, Carl Patton and David Sawicki outline a sequence of steps that practitioners can go through to apply information and analysis to any policy issue. They emphasize that you do not always follow these steps in order. Rather, you often circle through them at a general level and then go back through them in more detail, sometimes altering the sequence as you proceed. These steps can be boiled down to four basic actions:

1. Verify, define, and detail the problem or opportunity to be addressed.
2. Identify and analyze alternative courses of action.
3. Evaluate the alternatives and select a course of action.
4. Monitor implementation.

In fact, community practitioners have always gone through a sequence something like this in a commonsense way. And they have always relied on information. But in the past, such information may have depended heavily on the judgments, perceptions, and guesses of the participants. The addition of more objective data can add substantially to the depth, richness, accuracy, and reliability of the information that will be the basis for decisionmaking.

For example, consider the first step in the process: “verify, define, and detail the problem or opportunity.” Some residents may believe that their community does not have access to sufficient child care facilities and feel that a program is needed to remedy the situation. This belief may be based entirely on anecdotal information and personal experiences. Applying some objective data to this problem can shed considerable light on the validity of these perceptions.

Following the suggestion of the first step, you could begin by actually mapping out the locations of all child care facilities in the community. Examining this new information could lead to one of three possible outcomes:

The initial perceptions are wrong (that is, the map shows that the number and capacity of the facilities are quite adequate by sensible standards). In this case, the map will have saved the time and effort involved in mounting a program that was not needed. The community can shift its resources to more pressing issues.

The perceptions are partially right (for example, the map shows that the facilities are adequate overall, but seriously deficient in one or more parts of the community). In this case, the map will have shown where efforts should be focused.

The perceptions are right (that is, there are serious deficiencies throughout the community). Even in this case, the map should be valuable. First, the fact that you have documented the case in this way will give it credibility. It should help motivate others (residents, city agencies, funders, etc.) to get behind the program, and it should help in determining exactly how much work needs to be done and in which areas.

If the perceptions are wholly or partially right, the next step would be to “identify and analyze alternative courses of action.” It would be helpful at this point to define the particular **outcomes** that you hope to see result from your intervention. An outcome is a statement of a desired goal to be achieved through some project or initiative. Outcomes are usually focused on a specific condition (such as reducing teenage smoking) and are often related to some larger, overall objective (such as promoting community health).

Clarifying and stating the outcomes you hope to achieve can be very useful in the next step, “evaluating alternatives and selecting a course of action.” By articulating the outcomes from the beginning, you have a clear set of criteria against which to decide which of an array of possible interventions will be the most effective. A public statement of desired outcomes also can focus different community organizations and leaders on the problem and help ensure that everyone is working toward the same goal.

Once you have agreed upon a set of outcomes and chosen an intervention strategy, you will want to put your plan into action and “monitor implementation.” For this purpose, you need to select a set of **indicators** that relate to your outcomes and observe them over time. This may require developing some new sources of data, such as administrative data from a local agency or data from a survey. You should also agree on how you expect the indicators to change as a result of your action. If you do not see these results, you may need to reevaluate your action plan and change your intervention strategy accordingly.

Example: Targeting a First-Time Parents Program (Milwaukee)

The Next Door Foundation is an innovative youth agency on Milwaukee's West Side. It had grown rapidly from its origin as a neighborhood, church-sponsored activity into an independent agency substantially supported by program grants. The Foundation saw a special need for service in the West Side neighborhood, which had experienced substantial turnover in the 1980s. Many lower income families with older children had moved in. By the early 1990s, those children had become young adults, and many had children of their own. The perception was that a large proportion were single parents and that this proportion was higher than in most other neighborhoods in the city.

The Foundation began developing plans for a First-Time Parents Program in the neighborhood. The program would employ paraprofessionals to visit the homes of young parents, provide counseling on parenting skills, and offer friendship that would help diminish their isolation. In preparing their plans (and proposals to secure funding for the program), they recognized the value of having mapped information that would provide clear visual evidence of the need for the program. This same data could also be used to direct resources to areas where needs were particularly high.

The 1990 U.S. Census data include block group level data detailing the age of residents and family structure.² Maps were prepared showing the location of blocks with higher proportions of children under age seven and with children living in two-parent households. Each block group on the map was shaded to denote the proportion of different populations in that block. These maps clearly demonstrated the need for a First-Time Parents Program in West Side—there were high numbers of very young children per block compared with other areas in the city. They also indicated a subset of neighborhood blocks where there was the most critical need for focused outreach.

Using these maps as guides, the neighborhood was partitioned into a number of "service areas" by combining clusters of block groups for maximum flexibility. A selection of detailed demographic tables was prepared, customized for the service areas. Recognizing distinctions within the neighborhood helped make the program implementation process more sensitive to local conditions.

Program organizers felt that the First-Time Parents proposal and subsequent operating plans were much more solid because of the use of geographic information. The tables, maps, and additional graphics both demonstrated the need and suggested specific target areas to make implementation more effective. The proposal was funded, and the program has now been operating successfully for several years.

² For definitions of "block group" and other Census Bureau geographic terms, see **Chapter V, Geography**.

PARTNERING IN COMMUNITY-WIDE OR REGIONAL INITIATIVES

While implementing strategies for change in their own community or neighborhood will always be central to the work of many organizations, there is an increasing recognition that people must also think more broadly. The external environment constrains what they and their community's residents can accomplish, and they need to do something to alter those constraints. Accomplishing this usually means partnering with other groups (neighborhood-level as well as city- or county-wide and regional organizations) in larger-scale initiatives. Examples could include helping to establish a more effective metropolitan-wide job referral system, working on a broad commission to reduce the effects of racial discrimination, or helping to plan an overall, cross-service response to welfare reform.

In this context, information can be used to build bridges between different organizations and to unite them for action on a common set of goals. These might be groups working on the same issues in different parts of the community (inner city and suburban, for example) or on complementary activities in the same service area (preventing youth violence and reducing substance abuse). Oftentimes, these different groups may not be aware of other activities that are going on in their community or how their clients may overlap with other service providers. Information can provide a way of examining the common interests of different organizations and getting different actors around the same table.

The steps followed in this activity would be essentially the same as those outlined for planning and implementing specific initiatives. The character of the process would naturally be altered, however, by having a larger number of participants involved, often with less initial coherence in goals and interests. In addition, impacts must be considered across all neighborhoods or communities, rather than in just one.

When you begin looking at things from a wider perspective, you must consider how the different issues and actors overlap and interact with each other. This implies a **systems thinking** approach to understanding what is going on in the broader community. As more information is gathered about the different factors that might affect a particular issue, you can begin to formulate a model of how the different parts affect each other. Such a model can be very useful in helping people understand the need for cooperative effort in attacking problems on a community-wide basis.³

³ For more detail on systems thinking and its application to community initiatives, see *ACT Module 6: Systems Thinking for Community Improvement*.

Example: Social Services and Public Schools (Oakland, California)

In 1990, the Urban Strategies Council (USC) and the superintendent of the Oakland Unified School District recognized a common challenge. The school system and the city's array of social service agencies were not dealing with children holistically. Students' difficulties at school often emanated from problems at home, but the efforts of the schools and other agencies to help were fragmented and sometimes contradictory. They normally became involved at times of crisis, rather than working coherently to address root causes and to prevent crises.

The USC was able to secure, process, and link school and social agency data files for the students of one elementary school and their families. The results were presented to city and county agency representatives at a 1991 meeting called "The Same Client." The results on the overlap of service provision were striking and motivated agreement to conduct a similar study for an additional eight schools. In 1992, the USC published the results in the report *Partnership for Change*. The report showed that almost two out of three students used public services and that more than a third used at least two different services. It also documented that the system was investing much more in crisis services than in prevention.

The study's findings were presented to the County Board of Supervisors and other high-level officials, but their most important use was in the work of Oakland's Interagency Group (convened and facilitated by USC). The process established new working relationships between representatives of different agencies and forced them to recognize their common challenge. They had to "acquaint themselves with agencies outside of their normal scope of work" in defining questions they hoped the data-match would answer, and then, after the results were in, "discuss the kinds of joint action they might undertake, patterns of service use, relationships among agencies, and the ultimate effectiveness of different programs."⁴

The process resulted in the idea of redeploying staff from different agencies to form a "Family Support Team" around individual schools. The team would "develop new collaborative strategies for working with troubled families, taking on the crisis situations most taxing for schools, and leaving school resources to be focused on prevention, on establishing more positive activities, and on outreach to parents."⁵ This concept has since been tested in several schools and wider scale implementation is under way. USC continues to be involved in monitoring performance and in providing ongoing guidance and support.

⁴ Maria Campbell Casey, "Using Data as an Advocacy Tool: What it Takes," *Georgia Academy Journal*, Summer, 7-15.

⁵ Casey, 7-15.

COMPREHENSIVE STRATEGIC PLANNING

Comprehensive strategic planning requires getting community residents and organizations together in a process to think through where they want to go overall, and how they might best be able to get there. Comprehensive strategic planning would benefit from information about virtually all conditions that pertain to relevant neighborhood goals—thinking through problems, opportunities, and potential courses of action across all domains as a part of one process.

The strategy in this case is broader in scope, and takes a more long-term view. The sub-activities are basically the same as those noted above for planning an individual initiative. However, in this case each step is gone through much more quickly and in less detail than when designing concrete actions around a specific issue. You would rely on quick scans of a few key indicators in different domains and on rougher approximations.

While a comprehensive community strategy necessarily addresses many disparate issues, an important step in the process is explicitly **selecting priorities**. Today's community practitioners have recognized that, while they know they need to address a large number of interrelated issues that affect them over time, they cannot and should not try to do everything all at once. Making detailed plans for all sectors at the same time might well be so ambitious that it may be difficult to implement and could diminish the momentum of the initiative. It is therefore vital to tackle the highest priority issues first, and address other issues later.

As with implementing a specific initiative, from the chosen priorities one can develop specific **outcomes** to guide the strategic plan. These outcomes would describe conditions that the community would like to change (or preserve). The outcomes provide a more focused set of goals for the plan to describe. The outcomes can then be related to specific **indicators** that can be used to measure the community's progress toward achieving the outcomes it desires.

A comprehensive community strategy often results in a framework that can be used as the basis for developing **specific improvement initiatives** and for **partnering in community-wide or regional initiatives**. The strategic plan should enumerate broad goals for the community and indicate priority areas of action. These can naturally be refined further into specific initiatives for tackling priority problems in neighborhoods or communities or into partnerships among organizations and citizens to address common concerns or issues.

Example: Cleveland Community-Building Initiative

The Cleveland Community-Building Initiative (CCBI)⁶ was established in 1993 in response to a report on poverty in Cleveland's neighborhoods prepared by Case Western Reserve University's Center on Urban Poverty and Social Change. The report provided detailed analysis of poverty and other data and concluded that "bold action" was required to address Cleveland's persistent poverty problem. The CCBI was formed to implement a comprehensive community-building approach to address these concerns.

Several principles guided the formulation of the plan to address poverty in the city. The plan had to be comprehensive and integrated, but at the same time employ strategies that addressed the differences among neighborhoods. The emphasis for a neighborhood's strategy was to be an inventory of the community's assets, not its deficits. It was also necessary to involve local representatives in developing the plan and choosing strategies. Finally, the approaches needed to be tested in pilot areas first, before being deployed elsewhere in the city.

In developing a comprehensive plan for individual Cleveland neighborhoods, the CCBI employed a method known as **theories of change**. In this approach, stakeholders articulate a "theory of change" for their community; that is, they must describe in detail how and why particular initiatives will work and what outcomes are expected to be observed because of the initiative. By going through this process, the stakeholders can sharpen the planning and implementation of initiatives and gain a better understanding of how different initiatives may interact.

CCBI selected four geographic areas for testing its approach. In each of these areas, called "villages," CCBI identified groups of key stakeholders to articulate theories of change. Using a series of guided discussion questions, facilitators helped the stakeholders work through and develop their theories of change. The initial discussion focused on short-term objectives and outcomes, which then led to exploration of longer-term strategies and goals. This process was further refined until a consensus was reached by the key stakeholders.

The theories of change produced by the village stakeholders identified a series of early, intermediate, and long-term **outcomes** and related them to a set of observable **indicators** that can be used to monitor and evaluate the success of the community-building initiatives. The outcomes ranged from "attracting more businesses to the community" to "revitalizing the

⁶ Adapted from S. Milligan, et al., "Implementing a Theory of Change Evaluation in the Cleveland Community-Building Initiative: A Case Study," in K. Fulbright-Anderson, A. C. Kubisch, and J. P. Connell, eds. *New Approaches to Evaluating Community Initiatives, Vol. 2: Theory, Measurement, and Analysis*. Queenstown, MD. Aspen Institute, 1998, 45-85.

economy”; from “getting youth more involved in the community” to “increasing village pride and energy.”

The result was a focused agenda for bringing about change in these communities. Each village’s energies could thus be directed toward comprehensive change consistent with long-term strategies. As Milligan, et al., describe it:

The community’s growing capacity to achieve its goals allows advancement in areas it considers important, while stronger social structures sustain this movement. Thus neighborhood identity, security, service quality, economic opportunity, and family development are all promoted by the strengthened social organization within the community and the improved connections to the larger society.⁷

In this example, data was a key motivator in initiating the comprehensive community-building strategy in Cleveland. The initial data provided in the Case Western report provided the impetus for the community-building initiative. Data was also very important in informing and guiding the entire process of developing the initiative, especially the theories of change discussions. Furthermore, data collection and measurement strategies were incorporated into the final plans, including a list of indicators to measure outcomes.

REFERENCES

- Casey, Maria Campbell . “Using Data as an Advocacy Tool: What it Takes.” *Georgia Academy Journal*. Summer 1995, 7-15.
- Connell, James P, Anne C. Kubish, Lisbeth B. Schorr, and Carol H. Weiss, eds. *New Approaches to Evaluating Community Initiatives, Vol. 1: Concepts, Methods, and Contexts*. Washington, D.C.: Aspen Institute, 1995.
- Fulbright-Anderson, Karen, Anne C. Kubish, and James P. Connell, eds. *New Approaches to Evaluating Community Initiatives, Vol. 2: Theory, Measurement, and Analysis*. Washington, D.C.: Aspen Institute, 1998.
- Kingsley, G. Thomas, ed. *Building and Operating Neighborhood Indicator Systems: A Guidebook*. Washington, D.C.: The Urban Institute, 1999.
- Patton, Carl V., and David S. Sawicki. *Basic Methods of Policy Analysis and Planning*. Englewood Cliffs, N.J.: Prentice Hall, 1993.

⁷ Milligan, et al., 68.

The Coordinating Council of Broward. *The Broward Benchmarks*. Ft. Lauderdale, Fla.: February 1999. Available on the Internet at <http://www.sfrpc.com/current/menu536.htm>.

U.S. Department of Housing and Urban Development. *Mapping Your Community*. Washington, D.C., 1997. A complimentary copy can be obtained by calling the Community Connections Hotline 800-998-9999 and asking for Community 2020 Help Desk.

*Chapter 3****USING EXISTING DATA***

Once a community has decided that it wants to use data for one or more of the activities to affect community change, it then faces the daunting problem of acquiring the information it needs to look at current conditions and past trends. It will need data on different issues (health, housing, economics, etc.), at different levels of geography (county, city, neighborhood), and for different periods of time. Where will all of this information come from?

Fortunately, there are many sources of existing data that communities can use for a variety of purposes. The U.S. Census Bureau and other federal government agencies collect extensive data on the U.S. population and economy that are made available to the public. Much of this data is now available on CD-ROM or can be downloaded from the Internet. Local and state governments also collect a lot of data, although much of this may not normally be made available to the general public. For example, the state agency that administers welfare must collect information on its clients and keep track of people as they enter and leave the welfare system. Such data can be very valuable to community groups, if they can get access to it.

Finally, there are a number of private sources of data that can be useful to communities. There are several companies, for instance, that sell electronic business directories listing the names, addresses, and types of business establishments throughout the country. Other private

data vendors, such as Claritas, sell updated versions of U.S. Census data that include current and future population estimates.

Data that have been collected by someone else are often referred to as **secondary data**. This is distinct from primary data, which is data that you collect yourself (ways of collecting primary data will be discussed in the next chapter). The distinction is intended to signify that, with secondary data, you had no control over how the data were originally collected. The types of information obtained, the population sample from which it was obtained, the way that questions were asked—all of this was beyond your control.

Secondary data have a number of advantages over primary data. First, secondary data is usually less costly to obtain than data you collect yourself. Someone else has gone to the trouble and expense of collecting the data already. This often includes the process of “cleaning” the data, that is, making sure that they contain no errors. Second, using secondary data from a particular source may help create a demand for that data and thereby maintain the data source for future use. Finally, using secondary data can help build relationships between the data provider and the community—relationships that will benefit both groups. For example, if local police provide crime data to community groups, the community benefits by getting valuable information on public safety and the police could benefit by encouraging better reporting of crimes by citizens.

But there can also be disadvantages to secondary data. Because you have no control over how the information was collected, the data may not cover exactly the population or topics that you want. Data that are collected for administrative purposes, for instance, may not be entirely suitable for evaluating programs. The data also may not be very current or may contain errors. If the data are based on a one-time survey, then it may not be possible to get access to comparable data in the future. Finally, there may be restrictions placed on the use of the information by the collecting entity that make it difficult to use the data in practice.

There are several key issues that must be addressed when considering whether a particular source of existing data is appropriate for your use. The first issue is that of **ownership** of the data. Most data collected by the U.S. Census Bureau, for example, are in the public domain and, therefore, are free to be used by anyone. Data from some other sources, such as commercial vendors or local government agencies, may be private or proprietary. In this case, you might not be able to obtain all of the data that the provider collects (for instance, certain pieces of identifying information may be purged from welfare case records), or there may be a licensing fee to get access to the data. Furthermore, the provider may put certain restrictions on how the data can be used and whether they may be shared with others. Some private vendors, for example, allow only the licensing organization to access the data—the information cannot be provided or reported to others.

Related to the issue of ownership is the question of **confidentiality** of the data. As noted above, some information collected by government agencies for administrative purposes may be sensitive and, to protect the privacy of their clients, not releasable to the public. In this case, there are two options. The provider can simply omit the sensitive information (names, addresses, social security numbers) from the data before releasing them to others. Alternatively, the provider can summarize the data so that it is not possible to identify individual cases. For example, welfare caseloads could be summarized at the neighborhood level.

These approaches solve the confidentiality issue, but can make the data less useful to the community group. For instance, you might want to use social security numbers to match welfare case records with other types of social service records. If this information was removed by the data provider it will not be possible for you to do this. A second option that gets around this problem is for the community group to enter into a **confidentiality agreement** with the provider. In this case, the community group would get access to the complete, uncensored version of the provider's data, but certain restrictions would be placed on their use. The agreement might specify that the data can only be reported in a form where individual cases cannot be identified. It might also require the community group to keep the data in a secure location (such as on a password-protected computer).

Another issue to consider is the **timeliness** and **frequency** of the data. Are the data based on current information? Or are they out of date? Are new versions of the data collected on a regular basis? Of course, one always wants to have up-to-date data. But this may not always be possible. U.S. Census data, for example, are collected only once every ten years. While the Census Bureau does provide revised estimates of population characteristics between decennial censuses, these estimates are not available for small levels of geography (such as neighborhoods). So, to compare populations across cities, you may be able to use very recent data, but neighborhood comparisons may have to rely on older information.

This leads to the issue of **geography**. Data from some sources may be available for small levels of geography (blocks, Census tracts), while other data may only be provided for larger areas (cities, counties, states). Survey data are especially difficult to get for small areas, because to be able to produce accurate estimates one has to have many observations in the survey sample. Consequently, sources such as the Current Population Survey (CPS) or the American Housing Survey (AHS) cannot be used for very small areas.

A final issue regarding secondary data is the **format** in which the data are provided. Do the data come in a file format that can be readily read into a PC program (such as Excel or Access)? Or is it an ASCII file that must be converted? Is there documentation clearly identifying each of the data fields? Do the data come on diskettes, CD-ROM, or computer tape? How large are the data files? Do they require a lot of processing before they can be used (such as summarizing or combining)? All of these questions are important to answer up front so that

you do not invest a lot of time and effort obtaining data that you do not have the technical resources or expertise to use.

The remainder of this chapter consists of a catalog of secondary data sources that communities can use. This is not intended to be an exhaustive listing, but we have tried to include some of the major sources that cover an array of subject areas. We provide a name and description of each source, information on how to acquire it, details about its coverage (geography, timeliness, frequency), and other information.

More information on other sources of secondary data can be found on the “Tools” and “Networking” sections of the NNIP Web site (<http://www.urban.org/nnip/>).

CATALOG OF SECONDARY DATA SOURCES

Source:	U.S. Census Bureau
Domain:	Demography, Housing, Economy
Description:	<p>The U.S. Census Bureau is the major collector and provider of data on the United States. The Bureau carries out the decennial census, which is conducted every ten years to enumerate the entire population and to collect extensive information on the characteristics of people living throughout the nation at a very small level of geography.</p> <p>The Bureau also conducts many special purpose surveys on a regular basis. These include the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the Economic Census.</p> <p>Finally, the Bureau produces population estimates between decennial censuses. These estimates are available for states, counties, large cities, and some smaller geographic areas.</p>
How to Acquire:	Much Census data can be downloaded from the Bureau's Web site (http://www.census.gov). Other products can be ordered by calling the Census Bureau.
Cost:	Data can be downloaded for free from the Web site. For data on CD-ROM or tape, there is a charge depending on the item.
Timeliness and Frequency:	Decennial census data are collected only at the start of each decade, so the current information for 1990 is almost ten years old. Other surveys are conducted more frequently: the CPS and SIPP are conducted monthly; the Economic Census is done every five years.
Geography:	The decennial census data are collected using two survey forms—the short form and the long form. Short form information is collected on every person living in the United States and

includes basic characteristics, such as age, sex, and race. The long form is administered to only one out of every six households, and collects much more detailed information, such as employment status, income, and previous residence.

Because of the smaller sample sizes for the long form, this information is made available only down to the **block group** level. The short form information, however, is available at the **block** level.⁸

Accuracy:

Census data are generally regarded as being highly accurate. With regard to the decennial census, there are two important issues regarding accuracy. The first is the age of the Census data. Since the decennial census is conducted once every ten years, the information can become rather outdated after only a few years. This is especially important in areas that have experienced dramatic or rapid population changes since the last Census.

The second issue is what is commonly referred to as the Census “undercount.” Although the Census Bureau attempts to collect data on every person living in the United States, many people are missed and some are counted twice. In the 1990 decennial census, the net undercount was estimated at 4.7 million people, or 1.8 percent of the population.⁹ The major problem is that not all population groups are equally represented in the undercount. The undercount varies according to several factors, including minority status and whether one lives inside or outside a central city.

Format:

Most Census data can be obtained on CD-ROM in dBase IV format. A data extraction program is usually included with CD-ROM data. Some products are available only as ASCII files.

⁸ See **Chapter V, Geography**, for an explanation of different Census geographical units.

⁹ U.S. Department of Commerce, Bureau of the Census, "Report to Congress —The Plan for Census 2000," originally issued July 1997, revised and reissued August 1997, accessed online at <http://www.census.gov/main/plans/plan2000.pdf> (November 30, 1998), 2-4.

Data downloaded from the Census Bureau Web site can be obtained in HTML and tab-delimited ASCII formats.

Technical Skills Needed: Data processing and management skills.

Confidentiality and Use: Census data are in the public domain and may be used freely. Most decennial census data are provided as summary tabulations, so it is not possible to identify individual respondents. For survey data, the Bureau removes sufficient information before the data are released to the public so that individual respondents cannot be identified.

Possible Uses: There are a wide variety of uses for Census data. They can provide a basic or detailed demographic profile of communities and neighborhoods, or serve as baseline information for collecting local data.

Where to Get More Information: The U.S. Census Bureau maintains an extensive and informative Web site at <http://www.census.gov>.

For information on data products, see the CenStore page at <http://www.census.gov/mp/www/censtore.html> or call Customer Service at 301-457-4100.

Census data can be downloaded directly from the following Web sites:

American FactFinder: The Census Bureau's new data extraction system. Currently contains 1990 data but will include 2000 Census data in the future.

<http://www.census.gov/dads/www>

1990 Census Lookup: 1990 Census summary tabulations.
<http://venus.census.gov/cdrom/lookup>

Census Data Extraction System: Allows you to create tabulations from SIPP, CPS, the Consumer Expenditures Survey (CES), the American Housing Survey (AHS), and the Decennial Census Public Use Microdata Samples (PUMS) data.
<http://www.census.gov/DES/www/welcome.html>

Source:	Bureau of Labor Statistics (BLS), Local Area Unemployment Statistics (LAUS)
Domain:	Economy
Description:	BLS provides employment and unemployment statistics for the United States. Note: Some states may provide more detailed information on employment than the BLS.
How to Acquire:	Data can be downloaded from the BLS Web site http://stats.bls.gov/lauhome.htm .
Cost:	Free
Timeliness and Frequency:	Employment and unemployment statistics are updated monthly.
Geography:	Employment and unemployment data are available at the regional, state, and metropolitan area level.
Accuracy:	Employment and unemployment measures reported by the BLS are model-based estimates and, therefore, are not 100 percent accurate. Details on the LAUS estimation procedure can be found at http://stats.bls.gov/laumthd.htm .
Format:	Data can be downloaded from the BLS Web site as HTML tables or as comma, space, or tab-delimited ASCII.
Technical Skills Needed:	Data processing and management skills.
Confidentiality and Use:	Data are not confidential and may be used without restriction.
Possible Uses:	LAUS data enable one to compare annual employment and unemployment trends in states and metropolitan areas.
Where to Get More	The Bureau of Labor Statistics

Information:

The LAUS home page, <http://stats.bls.gov/lauhome.htm>, includes a set of Frequently Asked Questions and contains more information about the data provided and the estimation methodology.

Source: Business Directories

Domain: Economy

Description: Selected listings of businesses. These directories are produced by several private companies. Information varies by directory. Many include some type of business category code—either a Standard Industry Classification (SIC) Code or a yellow page heading.

How to Acquire: Contact individual vendors (see below).

Cost: Varies. (**Note:** Cost may not include unlimited access to data.)

Timeliness and Frequency: Databases are updated continuously.

Geography: Street address

Accuracy: Accuracy varies depending on the source of information used by the data provider and its update practices. Not all businesses may be included, and the information may be out of date. Since these directories are intended for use by marketers, they may not be as complete as administrative or research data sources. Contact individual vendors for details.

Format: Data available on CD-ROM. File format varies.

Technical Skills Needed: Address-based data must be **geocoded** to be used in mapping software or to create tabulations for specific geographic areas.

Confidentiality and Use: Data are not generally confidential, but the licensing agreement with the data provider may restrict data access and dissemination. For example, some products are “metered” so that only a certain number of records can be read from the database. Once this limit is reached, the user must pay an additional fee to access more records. Contact individual vendors for details.

Possible Uses: These data can be used to locate and map businesses in particular neighborhoods. If used over time, they can reveal shifts in business patterns.

Where to Get More Information: The following is a list of business data providers:

Acxiom — <http://www.databyacxiom.com>

Bresser Company — <http://www.bressers.com>

Cole Publications — 901 W. Bond St, Lincoln, NE 68521, Tel: 800-228-4571

DeLorme — <http://www.delorme.com/PhoneSearchUSA/>

Dun & Bradstreet — <http://www.dnb.com>

Global Business Information — <http://www.bryceallen.com>

Hill-Donnelly — <http://www.hilldonn.com>

InfoUSA — <http://phonedisk.com>

Source: Public School Records

Domain: Education

Description: Most public schools maintain computerized files of individual student records. Files usually include the student’s address, school attended, school transfers or exits, scores on standardized tests, attendance and disciplinary records, free lunch eligibility, and family status.

How To Acquire:	Contact local school district.
Cost:	Usually no cost to obtain data from schools, but may require substantial processing costs to put the data in a usable format.
Timeliness and Frequency:	Schools should update these records regularly.
Geography:	Street address or by school
Accuracy:	Depends on school system and its record-keeping practices. Private school attendance in an area will also affect the usefulness of this data.
Format:	Varies
Technical Skills Needed:	Address-based data must be geocoded to be used in mapping software or to create tabulations for specific geographic areas.
Confidentiality and Use:	School records are confidential, but, with proper protection agreements, can be used to develop measures for small areas. Access to data usually must be negotiated with individual school districts.
Possible Uses:	Data can be used to calculate attendance rates and average achievement for students by school or neighborhood. School residential mobility can be determined by matching student records across years. School completion requires matching records for a cohort of students (usually from eighth grade onward) to determine which ones graduate. School entry records have also been used to determine immunization status and school readiness.
Where to Get More Information:	Contact local school district.

Source: Vital Records

Domain:	Health
Description:	States and localities register births, deaths, fetal deaths, and other vital events. The civil laws of every state provide for a continuous, permanent, and compulsory vital registration system.
How to Acquire:	Contact the state or local vital statistics office, which typically compiles these records.
Cost:	Usually no cost to obtain data, but may require substantial processing.
Timeliness and Frequency:	Records should be updated at least annually by the vital statistics office.
Geography:	Street address
Accuracy:	Depends on local record keeping system
Format:	Varies
Technical Skills Needed:	Address-based data must be geocoded to be used in mapping software or to create tabulations for specific geographic areas.
Confidentiality and Use:	Portions of the vital records may be confidential and not releasable to the public. For instance, birth information is available in two sections. The first section contains a unique birth certificate number, mother's name, address, and other demographic and identifying data (such as age, race, and education of the parents). This section, also called the index portion , is confidential and released only when a special request justifying the need for such information is made. The second section, called the statistical portion , has information about prenatal care, congenital anomalies, and birth weight. The statistical portion is normally readily available for public health research.
Possible Uses:	There are many small area indicators that can be calculated from birth certificate data. Recorded birth weights can be analyzed to

arrive at the number of low-birth-weight infants. Information about the mother's prenatal care visits are also recorded. Kessner's adequacy of prenatal care index can be calculated from these data.¹⁰

The death file can be used to examine causes of death to see if leading causes are different in small areas than in the community as a whole, or in the community compared with the nation. Excess mortality can be calculated by comparing age-specific deaths in the community with expected deaths based on a standard population.¹¹

Where to Get More Information:

State or local vital statistics agency

Source:

Public Assistance Files

Domain:

Economy

Description:

Various forms of cash and in-kind assistance are given to eligible persons who qualify under means tested criteria. These programs operate under state or federal law but are delivered locally. Data on public assistance benefits and beneficiaries are kept by state or local departments of human services. Computerized individual records include name, address, case number, program participation, eligibility status, and benefit amount. A few states maintain longitudinal records, which track multiple programs used by the same household. In most places, however, cross-program records like this must be created by merging record systems from different agencies.

¹⁰ D. M. Kessner, J. Singer, C. E. Kalk, and S. Schlesinger, *Infant death: An analysis by maternal risk and health care*. Washington, D.C.: Institute of Medicine and National Academy of Sciences, 1973.

¹¹ C. McCord and H. P. Freeman, "Excess Mortality in Harlem," *The New England Journal of Medicine*, 322 (3), January 18, 1990, 173-177.

How to Acquire:	Contact state or local human service agency
Cost:	Usually no cost to obtain data, but may require substantial processing.
Timeliness and Frequency:	Varies. Most states maintain this information on a monthly basis.
Geography:	Street address of beneficiary
Accuracy:	Depends on local record keeping system
Format:	Varies
Technical Skills Needed:	Address-based data must be geocoded to be used in mapping software or to create tabulations for specific geographic areas.
Confidentiality and Use:	Public assistance files are confidential and can be released only for valid purposes with proper protection agreements in place. Some departments have geocoded their monthly files and can provide data with Census tract codes rather than names and addresses, reducing the confidentiality problems. Without recipient identifiers, however, longitudinal or matched files cannot be created.
Possible Uses:	Monthly files can be used to calculate participation in various public assistance programs for community residents. Longitudinal files can be used to calculate rates of long-term and short-term welfare participation.
Where to Get More Information:	State or local human services department

Source: Municipal Police

Domain: Public Safety

Description:	Police departments maintain records for each incident of reported crime occurring in their jurisdiction. These records contain a significant amount of information about the crime (including type, date, time, and location), the victim, and sometimes the suspect or arrestee.
How to Acquire:	Contact local police department.
Cost:	Usually no cost to obtain data, but may require substantial processing.
Timeliness and Frequency:	Data are updated regularly.
Geography:	Street address
Accuracy:	Depends on police reporting practices and record-keeping system
Format:	Varies
Technical Skills Needed:	Address-based data must be geocoded to be used in mapping software or to create tabulations for specific geographic areas.
Confidentiality and Use:	Detailed crime data are confidential. Availability of these data will vary by jurisdiction depending on the policy of the local police department.
Possible Uses:	There are several small-area indicators that can be developed with these data. The number and rates of crime by neighborhood or community can be calculated and compared. When making such counts, many researchers make use of only serious crimes (called Part I crimes under the Federal Bureau of Investigation's Uniform Crime Reporting (UCR) Program). Examining the less serious Part II crimes can also be of interest. Crimes can be broken out by the race and sex of victims and assailants or by the victim-assailant relationship. For example, if the information is available, it can be determined whether the victim and assailant live in the same Census tract or are of the same race.

Where to Get More Information: Local police department

Source: Property Tax Assessor or Auditor Records

Domain: Housing

Description: Information about every parcel of property in a community is maintained by the local auditor's or assessor's office for the purposes of levying taxes. There are three types of data about a property: (1) The **tax billing record** includes parcel number, parcel size, address of the property, owner's name and address, land and building assessed values, land use codes, gross taxes, special assessments, and delinquency status. (2) The **characteristics data** include parcel number, number of rooms, year built, and roof type. (3) The **deed transfer data** include information about property sales and transfers, names of buyers and sellers, sales amount, date of sale, and deed type.

How to Acquire: This information is in the public record and can be obtained from the local tax assessor's or auditor's office. The ease in obtaining this information in a usable format varies from one community to another, however.

There are also two national sources of property data. The Property Data Research Center collects information about more than 53.5 million properties nationwide, including vacant land and residential and commercial properties. Some of the same property information is contained in a commercial software product from Transamerica Intellitech called MetroScan.

Cost: To obtain data directly from local source, there is usually a charge per record. The Property Data Research Center also charges per record.

Timeliness and Frequency:	Data are usually updated frequently.
Geography:	Street address
Accuracy:	Quality of data depends on local assessor's office.
Format:	Varies
Technical Skills Needed:	Address-based data must be geocoded to be used in mapping software or to create tabulations for specific geographic areas. Complete parcel data for a county usually contain thousands of records and may require working with magnetic tape.
Confidentiality and Use:	Property records are in the public domain and so no confidentiality restrictions apply. Data obtained from private vendors may have restrictions as to use and dissemination.
Possible Uses:	There are several measures that can be developed from property data. The market and assessed values of homes can be computed. Median and average sales prices can be calculated and compared from year to year. The number of tax delinquent properties can be determined, along with the volume of property sales and transfers. Land-use patterns within the community (residential verses commercial) can be displayed. All of this information can be mapped for small geographic areas.
Where to Get More Information:	Local property assessor's or auditor's office Property Data Research Center — Experian RES, 5601 E La Palma Ave, Anaheim CA 92807, Tel: 800-421-1052 Transamerica Intellitech — http://ta-intellitech.com

Source: Home Mortgage Disclosure Act (HMDA)

Domain: Housing

Description:	<p>The Home Mortgage Disclosure Act (HMDA), enacted in 1977, is implemented by the Federal Reserve Board. This act requires certain mortgage lending institutions to compile and disclose data about loan applications and approvals. Institutions required to file HMDA data include commercial banks, savings and loans, credit unions, and mortgage companies that meet specific criteria.</p> <p>The data reported in HMDA include the institution's <i>Loan Application Register (LAR)</i> and <i>Transmittal Sheets (TS)</i>. These are referred to as the "LAR & TS Raw Data." LAR data contain loan and application information such as type of loan, purpose, amount, and action taken. There are also applicant characteristics such as race, gender, and gross annual income. TS data include information on the lending institution, such as name, address, parent company name and address, and tax identification number.</p>
How to Acquire:	HMDA data can be purchased from the Federal Financial Institutions Examination Council (FFIEC).
Cost:	A CD-ROM containing one year's worth of data for the United States costs \$50.
Timeliness and Frequency:	HMDA data are issued every year. There is a one- to two-year delay in the release of the data.
Geography:	LAR data contain census tract, state, county, and metropolitan area identifiers. TS data contain street addresses of lending institutions.
Accuracy:	HMDA data are most accurate in urban areas, where there is now a high proportion of institutions that are required to report. Data in past years may not be as complete, however.
Format:	Data since 1992 are available on CD-ROM and retrievable in ASCII format.
Technical Skills Needed:	Data processing and management skills.

Confidentiality and Use: All personal identifiers have been removed from HMDA loan data and so there are no confidentiality restrictions. Data may be used and disseminated freely.

Possible Uses: HMDA data can be used to determine the total number of loans applied for in an area and whether they were approved, denied, or withdrawn. The reasons for denial can also be examined. Loans can be broken down by purpose (home purchase, improvement, or refinancing), by amount of the loan, and by characteristics of the borrower or lender.

The approval and denial rates of financial institutions can be computed for small areas. HMDA data have been used by fair housing groups to look for possibly discriminatory lending patterns in communities. They are also useful to determine whether financial institutions are meeting the housing credit needs of their communities.

Where to Get More Information: Federal Financial Institutions Examination Council (FFIEC) — <http://www.ffiec.gov/hmda>

HMDA data order form can be downloaded at <http://www.ffiec.gov/order.pdf>

Source: U.S. Environmental Protection Agency (EPA), Envirofacts

Domain: Environment

Description: The Environmental Protection Agency (EPA) created the Envirofacts Warehouse to provide the public with direct access to the wealth of information contained in its databases. The Envirofacts Warehouse allows you to retrieve environmental information from EPA databases on Air, Chemicals, Facility Information, Grants/Funding, Hazardous Waste, Spatial Data, Superfund, Toxic Releases, and Water Permits and Drinking Water. You may retrieve information from several databases at once, or from one database at a time. You may use online

queries to retrieve data from these sources and create reports, or you may generate maps of environmental information by selecting from several mapping applications available through EPA's Maps On Demand service.

How to Acquire:	Data can be downloaded from the Envirofacts Web site at http://www.epa.gov/enviro/index_java.html .
Cost:	Free
Timeliness and Frequency:	Envirofacts data are updated monthly.
Geography:	Envirofacts permits users to limit their searches to locations in a particular state, city, county, or zip code. The locational information in the database includes the street address, city, state, county, and zip code, as well as the latitude and longitude of each site.
Accuracy:	Contact EPA for details about specific information in the database.
Format:	Data are retrieved as HTML tables through the Envirofacts Web site.
Technical Skills Needed:	None, but a Java-compatible browser is needed to access Envirofacts.
Confidentiality and Use:	The data are in the public domain, and there are no restrictions on dissemination or use.
Possible Uses:	The data contained in the Envirofacts database can be used to monitor compliance with various environmental regulations. Communities can identify specific pollution sources and identify which are meeting EPA standards and undergoing enforcement actions. Users can also identify locations of toxic chemical releases, and Superfund and hazardous waste sites in or near neighborhoods.

Where to Get More Information: Environmental Protection Agency (EPA) —
http://www.epa.gov/enviro/index_java.html

Source: National Center for Health Statistics (NCHS)

Domain: Public Health

Description: The NCHS maintains several data sets from a series of national health care surveys. These include the National Health Care Survey and the National Health Interview Survey. The surveys collect information on the provision and use of medical services, characteristics of patients discharged from hospitals, surgical procedures performed, use of home and hospice care, use of nursing homes, and the prevalence and effects of illnesses and disabilities.

How to Acquire: Download from NCHS Data Warehouse —
<http://www.cdc.gov/nchswww/datawh/datawh.htm>

Cost: Free

Timeliness and Frequency: Varies according to the survey. Some data are available annually, but other data are only available for a single year. The earliest data are for 1992.

Geography: Primarily national. Some data sets can produce estimates at the state level.

Accuracy: NCHS data are based on surveys and so are affected by both sampling and non-sampling error. Documentation provided by NCHS describes possible problems with the data.

Format: Complete public-use data sets can be downloaded in ASCII format. Summary tables and graphs for some data can be obtained in spreadsheet or HTML format.

Technical Skills Needed: Need to be able to download ASCII data files and convert them to a database or statistical software format.

Confidentiality and Use: Identifying and most geographic information have been suppressed in the data files made available to the public. NCHS prohibits users from attempting to link its data to other sources of individually identifiable data.

Possible Uses: Since the NCHS data are available only at the national or state level, they are generally not suitable for building community-based indicators. However, they can be used for setting benchmarks against which local data can be compared. Local users may be able to duplicate some of the NCHS survey methods to obtain comparable data in their own communities.

Where to Get More Information: National Center for Health Statistics (NCHS) —
<http://www.cdc.gov/nchswww/default.htm>

*Chapter 4***COLLECTING DATA NOT
AVAILABLE ELSEWHERE**

Making use of existing data is one strategy that communities can follow to build up a comprehensive information system. But, despite the wealth of information to be mined from existing sources, such sources will only go so far. As a community becomes more sophisticated in its understanding and use of data, it will undoubtedly wish to investigate new areas that are not covered by any preexisting data sources. In this case, the community must consider methods by which it can collect data that cannot be obtained from secondary sources.

As mentioned in the previous chapter, data that you collect yourself is referred to as **primary data**. Unlike secondary data, with primary data you have almost complete control over how the information is collected. You specify the overall goal of the data collection effort, what questions are asked, who is included in the sample, and so forth.

Surveys are one of the most common methods of primary data collection. One might conduct a survey of neighborhood residents to find out their opinions or attitudes on a variety of issues, such as crime, public transportation, or civic involvement. Surveys usually consist of a very structured series of questions that are asked of all participants, and the responses given may be restricted to a set of discrete choices (“very important,” “somewhat important,” etc.) or very brief statements.

At the other end of the spectrum are more open forms of data collection, such as focus groups. In this method, a set of neighborhood residents might be assembled and asked to give their opinions on a set of topics. The focus group would be led by a moderator, who asks a series of questions to guide the discussion, but generally allows the conversation to proceed in a free-flowing manner. Focus groups, like surveys, have their own special methods and procedures and there are particular ways of analyzing the resulting information.

Primary data have a number of advantages over secondary data. Original data can be tailored exactly to the program's or community's needs. As a result, original data can provide the most critical indicators by conforming more precisely to the chosen objectives. Furthermore, the community is more likely to "buy in" and support a program with original data that irrefutably characterize their unique situation. With sufficient resources, primary data can be made as precise as needed (such as to produce estimates for individual neighborhoods) and can be collected as often as needed.

But, of course, primary data also have their disadvantages. The major disadvantage is the cost and effort that are required to collect primary data. This is especially true for neighborhood-level data, where a large number of observations are needed to obtain sufficiently precise estimates for small areas. Collecting new data can absorb valuable resources that might otherwise be devoted to other efforts, and, once collected, primary data create a recurring demand for more data. Finally, certain primary data collection methods require technical expertise or resources that are not readily available in some communities.

There are several key issues that must be addressed when considering whether a particular method of collecting data is appropriate for your use. The first issue, as already discussed, is that of **cost**. Collecting primary data can sometimes be very expensive. There is a great deal of time involved in doing, for example, door-to-door interviews. If you need to pay the interviewers, then the cost will be greater the more "doors" you need to visit. To do a telephone survey, you may need to pay an organization with a computer-assisted telephone interview (CATI) system to conduct the survey for you. Even focus groups can be expensive, if they involve mailing out large numbers of invitations to prospective participants and hiring professional facilitators.

A second issue is that of the **technical difficulty** in doing primary data collection. Most data collection methods require specialized knowledge and skills to carry them out properly. For example, in surveys you must decide what population you are trying to collect data on and then choose a proper sample that will give you useful information on that population. Knowing how to construct a proper sample and interview the right number of people requires some expertise in statistics. Furthermore, you must design survey questions that will accurately obtain information on the issues in which you are interested and carefully train interviewers so that they administer the survey properly.

You must also be concerned with the unit of **geography** when collecting primary data. If you want data for an entire city, then you must cover more area than if you need only data for a single neighborhood. But, if you want data both for the entire city **and** for individual neighborhoods, then you will need to structure your sampling and data collection differently.

Finally, as with secondary data, there can be issues of **confidentiality** associated with primary data. If you are asking people to reveal personal information or to respond to sensitive questions, you will probably need to provide assurance that this information will not be revealed publicly in a way that would allow someone to associate specific answers with a particular person. In the case of focus groups, it is usual practice to have participants sign an “Informed Consent Agreement,” which indicates that they understand the purpose of the focus group and that they agree to allow the answers they provide to be used in the manner specified in the agreement.

The remainder of this chapter consists of a catalog of primary data collection methods that communities can use. Like the catalog of secondary data sources, it is not intended to be an exhaustive listing, but rather representative of the most commonly used methods. We provide a name and description of each method, information on when and how to use it, details about special issues (geography, accuracy, technical skills needed), suggestions of possible uses, and sources for more information.

SELECTED DATA COLLECTION METHODS

Name: Surveys

Description: In a survey, you ask a series of questions of a representative group of a particular population, with the goal of being able to make statements about the characteristics of that population. For example, a survey may allow you to make statements such as, “35 percent of the adults in our city drive their cars to work,” or “families in this neighborhood spent an average of \$12,000 on health care in 1998.” The representative group that responds to your survey is called the survey **sample**. Each individual or family in the survey is referred to as a **respondent**.

In most surveys, you ask the same set of questions of each respondent. The questions are listed on a form called a **questionnaire**, also referred to as a **survey instrument**. Normally, the responses in a survey are limited to very short or structured answers, such as “Yes” or “No,” a single number (such as \$12,000), or some kind of multiple-choice response (such as “Very important,” “Somewhat important,” or “Not important”). Some surveys do allow more open-ended answers, however, so that the respondents can elaborate on a question in their own words.

Many surveys are administered by an **interviewer** or **enumerator**—a person who asks the questions on the questionnaire and records the respondent’s answers. Such surveys are often done **in person** or by **telephone**. Other surveys are self-administered and do not use an interviewer. The most common form of self-administered survey is a **mail survey**, where the questionnaire is sent to the prospective respondent, completed, and returned by mail. More recently, many self-administered surveys are being conducted by e-mail or through the Internet.

When to Use: Surveys are generally used when the following three conditions apply:

- 1) You need information that is not available from a secondary source.
- 2) The information you require can be obtained by asking a series of structured questions that mostly require brief responses.
- 3) It would be too costly or impractical to obtain this information from the entire population.

If the information is available from a preexisting source, then it will almost always be easier and less expensive to obtain the data in this way. In the previous chapter, we listed several existing data sources that should be tried before undertaking a survey.

If you cannot obtain the information you need by asking a series of structured questions, then you will probably need to use **key informant interviews** or **focus groups** to collect the information.

Surveys are normally used in situations where you understand in advance what you need to ask the respondents and what types of answers you will be getting. For example, if you want to examine the programs and activities of organizations concerned with teenage pregnancy, you may be able to come up with a fairly structured set of questions to ask each group: *What types of activities do you engage in (counseling, medical referral, contraceptive distribution, etc.)? How much do you budget for each activity (\$/year)? How many clients do you see in a month?*

If, on the other hand, you are interested in understanding the perspectives of these groups on the broader issues surrounding preventing teenage pregnancy, you may need to do some key informant interviews that are less structured and allow more open-ended discussion. Questions asked during these interviews might include *What are the main obstacles to reducing teenage pregnancies? Which interventions are the most effective? Why?*

Finally, if you were interested in learning about a small group, such as the board of a neighborhood association, you could easily interview each member of the board and not bother with creating a sample. This is not possible if the population is an entire county, city, or even a neighborhood. You simply cannot interview every person in these groups, so you must select a smaller, representative sample.

How To Use (Action Steps):

- 1) Decide what population will be covered in the survey.
- 2) Write out questions that will be asked of each respondent.
- 3) Determine how large a sample will need to be surveyed.
- 4) If interviewers are being used, train them on proper administration of the questionnaire.
- 5) Test questions on small group of potential respondents (**pre-test**) to be sure that the questions are clear and will not be misunderstood by any of the respondents.
- 6) Correct errors or problems with questionnaire design.
- 7) Select sample for survey.
- 8) Conduct survey on selected sample.
- 9) Enter data from questionnaire into a computer database.
- 10) Check data for errors.
- 11) Analyze results.

Costs:

The costs of doing a survey depend on the size of the sample, the length of the questionnaire, the geographic area being covered, and the survey type (mail, phone, in-person).

The cost of a survey is usually expressed in terms of the cost per respondent. In this way, you can see how much it will cost to add an additional respondent to the sample, such as the cost to photocopy and mail out one more questionnaire or the cost to visit and interview one more household. As discussed below under “Accuracy,” the more respondents in the sample the more precise the results are. But the cost of additional respondents must be balanced against the need for more precise estimates.

Geography:

Surveys can be designed to cover any area of geography. For small areas, it is often practical to do any of the three survey types—in-person, mail, or telephone. The larger the area, the less cost-effective it can become to do in-person surveys.

Accuracy:

There are two types of error commonly discussed in surveys. The first, called **sampling error**, deals with the uncertainty in the survey results because you are only collecting data on a fraction of the total population in which you are interested.¹²

If you could interview every person in the population, there would be no sampling error because you would know for certain everyone's characteristics. Since a survey is only administered on a portion of the population, the exact responses you get depend on who ends up being in the sample. If you repeated the same survey a second time, but on a different sample, you would get slightly different answers. Nevertheless, if your sample is representative of the entire population, your survey results should be fairly close to the results you would get if you could interview everyone in the population.

The sampling error can be calculated mathematically based on the characteristics of the sample and is often expressed as a **confidence interval**. For example, when a newspaper reports that an opinion poll is accurate to "plus or minus 3 percentage points," this is an expression of the sampling error. So, if the poll reports that 25 percent of the respondents answered "Yes" to the question, "Do you like chocolate ice cream?" you know that the "true" percentage lies between 22 and 28 percent (25 plus or minus 3).

There is another piece of information you need to interpret this confidence interval. You need to know the **confidence level** you have in this estimate, that is, the probability that the true value lies inside the confidence interval. The confidence level is chosen by the person designing the sample. It indicates how certain you want to be in the answer obtained from the survey and is usually set at some high value, such as 90 or 95 percent. Once you have selected the confidence level, you can then calculate an appropriate confidence interval to reflect this value.

¹² The discussion of sampling that is presented here is extremely simplified to give you an idea of the general concepts. In reality, constructing a proper survey sample can be very complicated, and you will probably need to consult with a professional statistician for advice.

The **higher** the confidence level, the **larger** the confidence interval will be.

Returning to the polling example, if the confidence level was set at 95 percent, you would interpret the result “25 percent plus or minus 3 percentage points” as meaning “we are 95 percent certain that the true percentage of people in the population who like chocolate ice cream is between 22 and 28 percent.”

Sampling error is primarily a function of the size of the sample.¹³ When you are constructing your sample, you want it to be large enough so that the confidence intervals will be sufficiently small for the confidence level you have chosen. The larger the sample size, the smaller the sampling error will be. In a simple example, if you were going to conduct a survey in a neighborhood and wanted the results to be accurate to within plus or minus d percentage points, the following formula gives you the minimum sample size that you need:¹⁴

$$\frac{z^2}{4d^2}$$

The value z in this formula is a statistical function of the confidence level you have chosen. For a confidence level of 90 percent, $z = 1.64$; for 95 percent, $z = 1.96$. So, if you want your survey results to be accurate to within plus or minus 2 percentage points at a confidence level of 90 percent, you must have a sample size of at least 1,681 persons.¹⁵

¹³ The sampling error is solely a function of the sample size only in the case of a simple random sample (that is, where everyone in the population has an equal chance of being in the sample). For more complicated sample designs, such as clustered samples, calculating the sampling error is not so straightforward.

¹⁴ R. J. Larsen and M. L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, 2nd edition, Englewood Cliffs, N.J.: Prentice-Hall, 1986, 280-282.

¹⁵ This is the number of people who must **respond** to your survey. In actuality, you may have to attempt to ask many more people to get this level of response. See the information on **response rate**, below.

$$\frac{1.64^2}{4 \times 0.02^2} = \frac{2.69}{0.0016} = 1,681$$

Sampling error is only one type of error that can affect survey results. Other types of errors are referred to as **non-sampling errors**. There are numerous sources of potential non-sampling errors, and often these can have a larger impact than sampling error. Non-sampling errors are often hard to quantify—we cannot easily determine a confidence interval for these errors. In some cases you may be able to correct for non-sampling errors, but in others you will only be able to acknowledge the possible problems.

The first type of non-sampling errors are from respondents giving **wrong answers** to the survey questions. This may happen because the survey questions were not clear or were misunderstood, or because the respondent did not know or did not want to give the correct answer. It may be possible to include some cross-checks in your questionnaire to try to catch inconsistencies in the respondent's answers. For example, you can ask people what their total income is and then have them give their income from different sources. If the two amounts are very different, you probably have incorrect data.

A second type of non-sampling error is caused by **coverage problems** in the sample or by **incomplete data**. Both of these errors can lead to a situation where the survey sample is not truly representative of the entire population. When this happens, the survey results may be **biased** and not accurately reflect the characteristics of the population.

Coverage problems occur when you are not able to include all segments of the population in your sample. A typical example is a telephone survey, which necessarily excludes families without telephones. If this is a significant segment of the population in which you are interested, then a telephone survey will not give

you good results.¹⁶

When people refuse to answer certain questions in your survey, or refuse to respond to the survey at all, you end up with incomplete data. One measure of incompleteness is the survey's **response rate**, which is the proportion of people in the original survey sample who actually respond. If, for instance, you have a mail survey where you send out 1,000 questionnaires but only 430 come back, then you have a response rate of 43 percent.

Actually, the response rate by itself is not a very good indicator of whether you have a problem with incomplete data. The issue is not how many people in your original sample respond to your questionnaire, but rather whether those who respond are very different from those who do not; that is, whether the respondents are truly representative of the total population.

If you know certain characteristics of the population beforehand (such as demographics from the Census), then you can collect some of this same information in your survey and compare the characteristics of your respondents to the whole population. For example, if your target population is 32 percent African American, but only 8 percent of your respondents are from this group, you will probably have a problem with bias in your responses. If the characteristics of respondents and population match, however, then you **may** be able to make the case that your sample is representative.¹⁷

Technical Skills Needed: You need to have sufficient substantive knowledge of the subject area your survey is about to be able to formulate good questions. You will need an expert in sampling theory and statistics to help with issues of sample design. Someone experienced with

¹⁶ Another potential problem with telephone surveys is that households with multiple telephone lines are more likely to be selected for the sample. This can also cause bias in the survey results because, in principle, each possible survey respondent should have an equal chance of ending up in the sample.

¹⁷ Even if your sample matches the existing population according to an **observed** set of characteristics, the sample might still not be representative. If the sample may differs from the population by some important **unobserved** characteristics, then you may still have a significant non-response errors in your results.

questionnaire design (that is, how to ask questions in a survey so that you get accurate responses) would also be very helpful.

Confidentiality:

If your survey asks for personal or confidential information, it is important to let your respondents know that they will remain anonymous and their privacy will be respected. You may need to take precautions in your handling and storage of the questionnaires and data to ensure that unauthorized persons do not have access to it.

Possible Uses:

Surveys are used for many purposes. The federal government, particularly the Census Bureau, conducts many different types of surveys to measure characteristics of various populations. These surveys include the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Housing Survey (AHS). These surveys are designed to provide updated estimates of U.S. population and housing characteristics and are used for program planning and evaluation, budgeting, and allocation of funds to localities.

State, county, and local governments conduct surveys to get information on their localities that is more up-to-date than the U.S. Census. They may use these surveys to help them determine current community needs and to plan future spending and programs to address these needs. They may also use surveys to measure the effectiveness of the services they provide.

Nonprofits, neighborhood associations, and other nongovernmental groups may conduct surveys to learn more about their constituencies. Such information may not be available in published data or may be difficult to obtain for smaller areas. They may also use surveys to learn more about people's attitudes and concerns.

Where to Get More Information:

The American Statistical Association's Survey Research Methods Section (SRMS) publishes a series of pamphlets under the title "What Is a Survey?" For more information and to download selected pamphlets, see the SRMS's Web site at <http://www.stat.ncsu.edu/info/srms/srms.html>.

Customer Surveys for Agency Managers: What Managers Need to Know, by Harry P. Hatry, et al. Washington, D.C.: The Urban Institute, 1998. Tel: 877-847-7377, <http://www.urban.org/>.

Guide to the Design of Questionnaires. University of Leeds Computing Service, 1996. Available as a PDF document at <http://www.leeds.ac.uk/ucs/documentation/top/top2.pdf>.

How Effective Are Your Community Services? Procedures for Measuring Their Quality, 2nd Edition, by Harry P. Hatry, et al. Washington, D.C.: The Urban Institute, 1992. Tel: 877-847-7377, <http://www.urban.org/>.

Mail and Telephone Surveys, by Don A. Dillman. New York: John Wiley & Sons, 1978.

Survey Questions: Handcrafting the Standardized Questionnaire, by J. M. Converse and S. Presser. Quantitative Applications in the Social Sciences Series. Newbury Park, Calif.: Sage Publications, Inc., 1986. Tel: 805-499-9774, <http://www.sagepub.com/>.

Name:	Focus Groups
Description:	A focus group is a moderated discussion among people who share some common characteristics, such as living in a certain geographic area or belonging to a certain socioeconomic group. The purpose of a focus group is to get information about what people think and why they think that way.
When to Use:	You use a focus group when you want to understand what a particular segment of the population thinks and why it thinks that way. You would not use focus groups if you needed to be able to make statistical statements comparing different populations.

Unlike surveys, focus groups are not designed to produce statistical measures. Rather, the idea is develop a more in-depth understanding of how different people perceive particular issues. Focus groups allow you to answer questions such as *What issues are most important to homeowners who live in this neighborhood? Why are these issues important to them? How do their feelings about their neighborhood affect their quality of life?*

How to Use (Action Steps):

- 1) Decide which segments of the population you want to include in your focus groups.
- 2) Write out a set of questions to guide the focus group discussion. The questions are intended to start a conversation among the focus group members, so they should be fairly open-ended.
- 3) Recruit focus group participants. This can be done by sending invitations through the mail, calling people on the telephone, posting signs in the community, or taking out advertisements in the local newspaper. Many focus groups offer some kind of incentive (such as a cash payment, gift certificate, or other gift) to encourage people to participate.
- 4) Among the prospective participants responding to your invitation, select those who will participate in the focus groups. You want to select participants who reflect a fairly good representation of the population in which you are interested. You will probably want to get some information from the prospective participants (such as their age, race, employment status) to help you make a selection. If there are more people than you need to select in a particular group, you will probably want to make a random selection among the possible participants. The focus groups should be constructed to encourage a good, open discussion. If, for example, you are concerned that members of one racial group might not speak as freely in a mixed racial setting, you may want to separate your groups by race.
- 5) Select locations for your focus groups and inform the participants. You will want to select locations that will be easy for people to get to and will be a comfortable environment. You should avoid locations that may be perceived as being “not neutral” by some members of your group.

- 6) Conduct the focus groups. The focus group will be run by a **moderator** who will ask the participants the questions that have been prepared in advance and guide the group discussion. The moderator should be sensitive to the dynamics of the group and make sure that everyone participates in the discussion. He or she should also be careful not to lead the discussion too much, but rather to let it develop naturally. Also present will be a **recorder** who takes notes during the session. In addition to taking notes, you may wish to record the session on audio- or video-tape.
- 7) Analyze the focus group results. The recorder's notes and the audio or video recordings made during the focus group provide the basis for this analysis. You want to try to summarize the proceedings of the discussion and look for common themes, such as "Five of the 12 focus group members agreed that crime was a serious concern in the neighborhood." It is also common practice to highlight some key quotes that illustrate these views.

Costs:

Recruitment is probably the largest cost of doing a focus group. If you do a mailing, you will probably have to send out a large number of invitations to get an adequate response. Telephone or mass media recruitment (such as advertisements in newspapers) are probably cheaper, but may not be as effective.

In addition, if you are offering incentives to your participants, you must factor in this cost. You may also need to rent a meeting space for your focus group, provide light refreshments, and obtain an audio or video recorder. Finally, since it is essential to have a skilled and experienced focus group moderator, you may need to hire a professional moderator to conduct your groups.

Geography:

Focus groups usually draw people from a fairly small geographic area, such as a neighborhood or town. It is difficult to have a small group be representative of a larger community. If you want to examine the attitudes in a large city, you will need to have a series of groups drawing from residents in different areas.

Accuracy:

Focus groups, unlike surveys, do not generate estimates that have a statistical level of certainty. Nevertheless, the goal of both

methods is to have observations that reflect the views of some larger population. The extent to which the focus group participants are representative of this larger population, and the degree to which they feel that they can openly and honestly express their views, will determine how accurately the group represents the population as a whole.

Technical Skills Needed: Moderating a focus group is a fairly specialized skill and having an effective, experienced moderator is crucial to the success of the group. You may need to hire a professional moderator to conduct your focus groups. You may also need to work with U.S. Census or other demographic data to help you select focus group participants that are representative of the larger population.

Confidentiality: Since the goal of a focus group is to get people to freely share their thoughts and opinions on a subject, you must create an environment where they feel comfortable doing this. This will mean ensuring the participants that any comments they make or views they express will be reported **anonymously**, so that it will not be possible for someone to later attribute statements to particular individuals.

When you report the results of your focus groups, you should not reveal the names, addresses, or any other identifying information about the participants. It is acceptable to report some aggregate statistics on their characteristics, however, such as, “This focus group was made up of 12 adults: 2 African-American males, 4 African-American females, 3 white males, 2 white females, and 1 Hispanic male.”

You may quote statements made by the participants, but you should not attribute these statements to a particular person. For example, you might say, “One of the participants in the focus group conducted in the Wooded Glenn neighborhood said, *‘People don’t clean up after themselves. There’s always litter left on the sidewalk and street, and it blows into my yard!’*”

It is standard practice to have focus group members sign an **informed consent agreement** prior to the start of the session. By signing this agreement, the participants will be stating that

they agree voluntarily to take part in the focus group and that they understand what the purpose of the group is and how the information collected during the session will be used.

Possible Uses:

As mentioned above, focus groups allow you to get a more in-depth understanding of how people view particular issues. They can be used to discover people's attitudes and opinions toward a variety of topics.

Focus groups are sometimes used prior to a survey—to give some background information that would allow you to develop a more comprehensive survey—or in conjunction with a survey—to help you better interpret the meaning of the survey results.

Where to Get More Information:

Focus Groups as Qualitative Research, by D. L. Morgan. Newbury Park, Calif.: Sage Publications, Inc., 1988. Tel: 805-499-9774, <http://www.sagepub.com/>.

Focus Groups: A Practical Guide for Applied Research, 2nd Ed., by R. Krueger. Newbury Park, Calif.: Sage Publications, Inc., 1994. Tel: 805-499-9774, <http://www.sagepub.com/>.

Focus Groups: Theory and Practice, by D. Stewart and P. Shandasani. Newbury Park, Calif.: Sage Publications, Inc., 1990. Tel: 805-499-9774, <http://www.sagepub.com/>.

The Handbook for Focus Group Research, by T. Greenbaum. Newbury Park, Calif.: Sage Publications, Inc., 1998. Tel: 805-499-9774, <http://www.sagepub.com/>.

Successful Focus Groups: Advancing the State of the Art, edited by D. Morgan. Newbury Park, Calif.: Sage Publications, Inc., 1993. Tel: 805-499-9774, <http://www.sagepub.com/>.

What Are Focus Groups? From the American Statistical Association's Survey Research Methods Section (SRMS) series *What Is a Survey?* <http://www.stat.ncsu.edu/info/srms/srms.html>.

Name:	Community Forums
Description:	<p>Community forums are public meetings where information is exchanged and where everyone, within a prescribed process, is free to express their opinions. A forum can be used to educate the public about a particular topic, or to get feedback from a community about issues that are important to them.</p> <p>Forums are usually gatherings of people who are all physically in the same place. With teleconferencing or videoconferencing technology, however, it is now possible to have forums that include people in different locations. It is also becoming more common to have “virtual” forums through the Internet.</p>
When to Use:	Community forums are used when you need to address a broad audience in a community, and when you do not need to discuss issues that people would not talk about in public. You can use forums, like focus groups, to try to understand what people think and why they think that way.
How to Use (Action Steps):	<ol style="list-style-type: none">1) Decide what issue the community forum will address, such as “Access to health care in our community.”2) Specify the goal for the forum. Is it to learn people’s opinions about the issue? Is it to present differing expert views on the issue and begin a public debate?3) Set the format for the forum. Will there be a formal presentation? Will an expert panel discuss the issue and take questions from the audience? Do you want to limit the number of participants?4) Define the population to be covered by the forum. Is it an entire county? City? Neighborhood? Is there some segment of the population that you wish to focus on, such as the elderly or parents with school-aged children?5) Given the population to be covered and the size and format you have selected, determine the number of forums you need to hold and the locations where they should be held.6) Secure meeting places for the forums. These should be facilities large enough for the participants you expect and easily accessible by your target population. They should also be held in “neutral” locations that will not dissuade anyone

from attending. Public buildings, such as libraries or civic auditoriums, are usually a good choice.

- 7) Find appropriate moderators, speakers, panelists, etc. for your forums.
- 8) **Publicize** the forums to the target population. You can use posted signs and advertisements in local newspapers for this purpose. You can also make use of neighborhood associations or other community groups to publicize the forums—particularly if the group deals with the issue or population that is the focus of the forum.
- 9) **Conduct** the forums. You may want to tape-record or videotape the proceedings. At a minimum, someone should take thorough notes on what happens.
- 10) **Compile** and **analyze** the results of the forum. You may want to write a report of the forum results.

Costs:

Costs for publicizing, renting meeting spaces, and providing refreshments. You may also need to hire moderators or pay for people to participate in panels or expert discussions.

Geography:

Because a forum is (usually) a physical gathering of people, you are limited by the size of the meeting space, so it is difficult to have a single forum that covers a large area. Depending on the turnout that can be expected to your forums, you will need to hold several to cover a large area like a city or county.

Accuracy:

Forums are not intended to provide accurate statistical measures of the attitudes of a population. They only tell you about the views of the people who come to the forum. Ideally, you would like participants in the forum to represent a broad spectrum of people in your target population. But with forums, unlike focus groups, you have a very limited ability to preselect participants according to certain characteristics. Consequently, your forum may disproportionately represent the views of particular segments of the population.

Technical Skills Needed:

Forums require good organization and planning. Moderating a forum is a special skill, and it is best to find someone who is experienced. If you are inviting a panel of experts to your forum, you will be able to identify and select people who are appropriate

for the issue being addressed.

Confidentiality:

Forums are, by definition, public meetings, and so confidentiality issues do not apply.

Possible Uses:

Georgia Health Decisions used community forums as part of its process for understanding how Georgian citizens felt about health care. “The purpose of the community forums was to educate participants about the health care crisis, to create the opportunity for dialogue about health care among participants, to engage participants in activities that required thinking beyond the current health system, and to build a network of concerned citizens interested in health issues.”¹⁸

The National PTA promotes the use of community forums as part of its strategy for combating community violence. It has outlined 10 steps for conducting community forums.¹⁹

Sustainable Racine has conducted a series of community forums at different sites so that people can discuss what needs to be done to build on their neighborhood’s strengths and improve their community. Brief summaries of the forum results have been posted on the Internet so that each site can share its views with the others.²⁰

Where to Get More Information:

Collaborating to Improve Community Health: Workbook and Guide to Best Practices in Creating Healthier Communities and Populations, by Kathryn Johnson, Wynne Grossman, and Anne Casidy, editors. Jossey-Bass Publishers. August 1997.

¹⁸ *Community Forums Bring Georgians’ Values to Health Debate*, Georgia Health Decisions. Available on the Internet at <http://www.cpn.org/sections/topics/health/stories-studies/ghd.html> (accessed 18 June 1999).

¹⁹ *National PTA Community Violence Prevention Kit*. Available on the Internet at <http://www.pta.org/events/violprev/violforum.htm> (accessed 18 June 1999).

²⁰ *Sustainable Racine Community Forums*. Available on the Internet at <http://www.johnsonfdn.org/comments/index.cfm> (accessed 18 June 1999).

Name:	Key Informant Interviews
Description:	<p>Key informant interviews are used when you want to get information from a select group of individuals who have a certain level of knowledge or experience on a particular subject. People commonly selected for key informant interviews are policymakers, staff of community-based organizations or nonprofits, government professionals, politicians, and so on. Unlike surveys, but like focus groups, key informant interviews are relatively unstructured, allowing for more open explanations by the respondent.</p>
When to Use:	<p>Key informant interviews are used to give you background on an issue from the point of view of those who are intimately involved. You use them when you need to get a more “informed” assessment of a situation than could be provided by someone in the general public.</p>
How to Use (Action Steps):	<ol style="list-style-type: none">1) Decide what issue or topic will be covered in the interviews.2) Determine what types of key informants would best be able to provide insight on the issue or topic.3) Write out questions that will be asked of each respondent, to create an interview guide. These should be fairly open-ended questions designed to provoke a discussion with the respondent. You may need to vary the questions for different types of respondents, such as health care providers versus government health agency officials.4) Create a list of key informants to interview.5) If interviewers are being used, train them on proper administration of the interview guide. Normally, you will use only a few interviewers to do key informant interviews. Since the interview is relatively unstructured compared with a formal survey, it is important that each interviewer understand clearly the goals of the interviews. For example, they must know when to probe the respondents for more details and how to follow up on certain responses. You will probably want to have practice interviews to go over these points with the interviewers.6) Since the number of respondents will be relatively small compared with a formal survey, pretesting the interview

guide may not be practical. Nevertheless, you should be sure that the questions are clear and will not be misunderstood by any of the respondents. Reviewing the interview guide after the first few interviews is a good idea.

- 7) Correct errors or problems with interview guide design.
- 8) Select a set of key informants to interview.
- 9) Conduct interviews. You may wish to record the interviews on audiotape.
- 10) If you have done a lot of interviews, you will probably want to enter the responses into a computer database, a spreadsheet or a word-processed document. This will help you organize and examine the responses to the same questions by different key informants.
- 11) Analyze results.

Costs:	Key informant interviews are fairly low cost. There are usually no mass mailings to do, and you do not need large numbers of interviewers.
Geography:	Any level of geography can be covered with key informant interviews. It depends only on the areas the informants are competent to discuss.
Accuracy:	The accuracy of key informant interviews depends on how knowledgeable the informants are. Even then, different “experts” may disagree on the interpretation of facts, the description of past events, or other details. It is therefore wise to get as broad a sampling as possible of different types of respondents so that you will get different perspectives on the issue in question.
Technical Skills Needed:	The interviewer should be fairly knowledgeable about the topics being discussed during the interviews so that he or she can ask good follow-up questions and probe for more detailed answers.
Confidentiality:	Confidentiality restrictions can vary with your needs and the comfort level of the respondents. For some issues, respondents may not wish to have their opinions revealed publicly. You will probably get franker responses if you do not quote named sources when reporting on key informant interviews, but rather identify respondents as, “a health care professional” or “a staff

member of a nonprofit organization.” In any case, you should clearly inform your respondents, prior to the start of the interview, as to how the information collected from them will be used and whether they will be identified in any reports or materials released to the public.

Possible Uses:

Key informant interviews can be used to get more background on particular policy issues. For example, to understand the issues relating to health care access in a community, you might interview health care providers, government health regulators, hospital administrators, and residents.

This method can also be used to learn about events that occurred in the past and relate them to current situations. For example, the Urban Institute used key informant interviews in a study about the effects of subsidized housing on neighborhoods.

Key informants, such as nonprofit developers and housing agency staff, were questioned about past controversies surrounding the siting of subsidized housing in particular communities. This information was used as background to understanding the current policy environment surrounding this issue.

Where to Get More Information:

Questions and Answers in Attitude Surveys, by Howard Schuman and Stanley Presser. New York: Academic Press, Inc., 1981.

A Summary of Studies of Interviewing Methodology. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Public Health Service, 1977, Series 2, Number 69.

Name:

Behavior Mapping

Description:

Behavior mapping is a technique in which you observe the behavior of individuals or groups in some setting and then characterize that behavior according to some predefined classifications. A **behavior mapper** will begin making observations at a specified time and place. Over a given period,

the mapper will record the numbers and types of people observed and may also record information about what they are doing (sitting, talking, playing checkers, etc.). The mapper enters this information onto a form that has been set up to allow easy recording of the data.

The mapper may remain stationary, if the entire area of observation can be seen from a fixed point, or the mapper may move about, if it is necessary to cover an area that cannot be seen all at once.

A series of these observations will be made at the same locations over the course of the survey. It may be necessary to vary the starting times of the observation periods (such as morning, afternoon, and evening, or weekdays and weekends) so that a representative sample of times is recorded.

When to Use:

This method is used when you want to observe patterns of use of particular facilities or services or other behaviors of people. You can obtain information on the numbers of persons at particular times, the amount of time they spend doing particular types of activities, and the visible characteristics of people.

How to Use (Action Steps):

- 1) Determine what **types of information** should be collected by the behavior mappers. Is it simple counts of people? Do you want to know characteristics of these people (race, age, sex)? Do you want information on what they are doing (running, playing, loitering)? Collecting large amounts of detailed information is very difficult in behavior mapping, so the requirements must be kept fairly simple.
- 2) Decide what **types of locations** are to be observed. Some examples might be public parks, recreation centers, and health clinics.
- 3) Create a **recording form** for use by the mappers in recording the information. The form may consist of a grid on which the mapper can make tick marks for each person of a given type who is performing a given activity. Or the form may be more of a log on which the mapper writes out what is observed: "2 African-American boys playing ball."
- 4) Choose **specific locations** where the observations will take

place.

- 5) Set up a ***schedule*** for the observations. The schedule should cover different times during the day (morning, afternoon, and evening) and different days of the week, as appropriate.
- 6) ***Pretest*** the recording form at one or more locations to ensure that it is adequate for recording the information needed.
- 7) ***Train the mappers*** in proper procedures and use of the recording form. It is a good idea to have some tests for the mappers so that they can practice making observations. It is important that the mappers all perform their observations the same way so that the information from different mappers will be comparable.
- 8) ***Analyze the results*** of the mapping data. If the observation schedule is constructed properly, the data collected may be considered a random sample and can be analyzed statistically.

Costs:

Behavior mapping is a very time-consuming activity, and so the largest cost is paying mappers to make the observations.

Geography:

Geography is limited by the territory that can be covered by a single mapper. If it is a small enclosed space or a large open space that can easily be observed by one person, then more territory can be covered. If, on the other hand, it is a park with trails that cannot be observed from one point, you may need several mappers walking about to cover sufficient area.

Accuracy:

The accuracy depends on the quality of the observations made by the mappers and the representativeness of the sample of times when observations were made.

Technical Skills Needed:

No special technical skills are necessary for behavior mapping, but the mappers must be well trained and procedures and forms must be tested beforehand. If the results are to be analyzed statistically, then you will need the help of a statistician in designing the observation schedule and analyzing the data.

Confidentiality:

Since the information collected in behavior mapping contains only general characteristics and no identifying information, there are

usually no confidentiality issues. It may be necessary to get permission, however, to observe people in certain settings, such as private locations.

Possible Uses:

Behavior mapping can be used to study the usage patterns of public facilities, such as commercial areas or public parks. The Urban Institute is conducting a multiyear study of how people use major urban parks by observing the numbers and types of persons in different park areas and identifying the activities they are engaged in.

Behavior mapping was used by Bell and Smith to assess the efficacy of change in special care units for nursing home residents. Residents in one newly opened and one well-established Alzheimer care unit were observed over three months. The researchers were able to make comparisons of performance outcomes between the two facilities.

Where to Get More Information:

“A Behavior Mapping Method for Assessing the Efficacy of Change on Special Care Units,” by Paul A. Bell and Jeffrey M. Smith. *American Journal of Alzheimer’s Disease*. Vol. 12 (4), July/August 1997, 184-189.

Inquiry by Design: Tools for Environment-Behavior Research, by John Zeisel. Monterey, Calif.: Brooks/Cole Publishing Co., 1981, 111-136.

*Chapter 5***MANAGING AND WORKING
WITH DATA**

Once the data have been collected, either from a secondary source or using one of the primary data collection methods described in the previous chapter, they must be stored in some format for use. This can be an enormous organizational task, and the way it is done can greatly influence the quality of the data and the ease with which you can analyze them. The process of organizing information into computer files for later use and analysis is known as ***data management***. This chapter will present a basic introduction to the topic and describe some of the important principles that you should understand.²¹

Data management includes the following activities:

- Creating an organization and structure for computer files to hold your data
- Entering data into these files

²¹ This chapter deals primarily with issues surrounding the management of ***quantitative*** data, such as from a secondary source like the U.S. Census or from a survey. Many of these same issues also apply to handling ***qualitative*** data, such as from a focus group or key informant interviews.

- Verifying that the data have been correctly entered and contain no errors
- Combining and transforming original data files to create supplementary files for analysis
- Performing preliminary or exploratory analysis of data
- Documenting the data sources and files, including a list of known problems.

Having good, well-thought-out data management practices is crucial to being able to get the most out of your data. If the data are improperly organized, they will be difficult to work with and may even cause you to analyze the information improperly. If the data are not correctly entered and tested, they may contain errors that will spoil your analysis. Finally, if the data are not thoroughly documented, it may be difficult for you or others to make use of them later and much of the effort that went into collecting them will be wasted.

The rest of this chapter will describe the basic concepts behind the structure and organization of data files. It will present some basic geographic definitions that are used by the U.S. Census Bureau and are common for many types of data. Next, it will describe some of the different types of software tools that are useful for managing and working with data. The last three sections will discuss techniques for entering data, conducting preliminary analysis, and documenting data.

DATA FILE STRUCTURE AND ORGANIZATION

A **data file** is a computer file that contains related information (“data”) that has been structured in a way that makes the information easily accessible. There are actually many different systems for organizing data files, some more useful than others. The most common arrangement is the **rectangular data file**. A rectangular file closely resembles a two-dimensional table, with rows (also called **records**) and columns (or **variables**):

<i>Record number</i>	<i>Variable names</i>			
	ID	NAME	AGE	WEIGHT
1	101	Alice Young	27	128
2	102	Bob Smith	45	245
3	103	Cecil Jones	36	162

In a rectangular file, each column represents a variable—the basic type of information stored in the file. A person’s name, age, and weight are examples of information that can be entered into variables. Variables often have a name that is used to identify them within the file. In the example above, there are four variables: ID, NAME, AGE, and WEIGHT. Each variable in a rectangular file is used to hold only one type of information. In other words, the variable that stores a person’s age will not suddenly switch in the middle of the file and become the person’s weight.

The rows of a rectangular file represent different records or observations in the file. If, for example, you have a data file containing information on 200 persons and each person is represented by a different row in your file, then the file will contain 200 records. The example file above contains three records with information on three different persons.

Each record in a rectangular file commonly represents the *level* of the file. In our example, we have a *person level* file because each record represents a single person. Another way of putting this is to say that a person is the *unit of observation* of this file. Having all of the records represent the same unit of observation is important for statistical analysis. Most software packages assume that each record represents that same unit of observation when calculating descriptive statistics like averages. If you wanted to calculate the average age and weight of each person in your sample, it is easier if your file only contains person level data.

Each file should contain one or more *key variables* that uniquely identify each record. In the above example, the variable ID is the key variable as it is a unique identifier for each person. Key variables may relate to some external source (such as social security numbers or case numbers generated by a hospital’s intake system) or it may simply be a sequential numbering of sample cases.

Key variables are important because they allow you to be sure that you are not confusing data from one respondent with another. For instance, NAME would not be a good key variable in our sample file if we had two different people named “Bob Smith.” In this case, if you wanted to compare this information with data from another part of the survey stored in a second file, you would not be able to tell which Bob Smith was which. With a unique identifier, however, this is not a problem.

A final note on the issue of data file *format*, which is distinct from file structures: the file format is the way in which the data are actually stored in a computer file and is specific to the type of software being used to create and access the file. Although many different software packages work with rectangular files, they use different formats for storing the information. Two of the most common formats are *dBase* format, which is used by the dBase database package,

and **Excel** format, which is used by the Microsoft Excel spreadsheet package.²² Other software packages, such as Microsoft Access, SPSS for Windows, and SAS for Windows, have their own proprietary formats that cannot be used directly by most other software. It is possible to convert files from one format to another, however, and most packages (including Access, SPSS, and SAS) have the ability to read and write Excel or dBase files.

GEOGRAPHY

As was stated above, it is important to be able to identify the unit of observation in your data. One component of this is the **geographical unit**, such as state, county, or neighborhood, represented by the data. Such information is a key characteristic of data.²³ By knowing the different geographical units your data represent you can summarize the data in different ways to get statistics for smaller and larger areas.

The U.S. Census Bureau has defined an entire hierarchy of geographical units for collecting, tabulating, and reporting data. Since these definitions are widely used by many providers and users of data, we will present them here. The hierarchy of basic Census geographical units is (from largest area to smallest):

State
 County
 Census tract/Block numbering area
 Block group
 Block

Each of these units is made up of combinations of the units listed below it. States, for instance, are made up of combinations of counties, while census tracts/block numbering areas are made up of combinations of block groups.

The two highest levels of geography, **states** and **counties**, correspond to the political boundaries for these two jurisdictions.

²² In the Microsoft Windows environment, dBase files are usually identified by the .DBF extension added to the end of their names, while Excel files have a .XLS extension.

²³ Another key characteristic is **time**, that is, the period of time the data represent. Since units of time (hours, days, months, etc.) are much more familiar than geographical units, they will not be discussed here.

Census tract/block numbering areas are the next smallest areas in the geographic hierarchy. As of the 1990 census, every part of the U.S. is covered by either a census tract or a block numbering area. **Census tracts** are small, relatively permanent statistical subdivisions of a county. Census tracts are delineated for all metropolitan areas and other densely populated counties by local census statistical areas committees following Census Bureau guidelines. **Block numbering areas (BNAs)** are small statistical subdivisions of a county for grouping and numbering blocks in nonmetropolitan counties where census tracts have not been established. BNAs are often referred to as census tracts, since they are functionally equivalent to tracts.

Census tracts/BNAs usually have between 2,500 and 8,000 persons and are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. The land area of census tracts varies widely depending on their population density. Their boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons can be made from census to census.

The smallest areas used by the Census Bureau for tabulations are **block groups** and **blocks**. Census blocks are roughly equivalent to street blocks. They are defined as small areas bounded on all sides by visible features such as streets, roads, streams, and railroad tracks, and by invisible boundaries such as city, town, and county limits. A block group is a cluster of blocks within a census tract/BNA. Block groups generally contain between 250 and 550 housing units, the ideal size being 400 housing units.²⁴

In addition to these geographical units, there are a number of other geographies defined and used by The Census Bureau that do not fit so neatly in the hierarchy listed above. **Metropolitan Areas (MAs)** consist of a large population nucleus together with adjacent communities that have a high degree of economic and social integration. The MA classification is a statistical standard designated and defined by the federal Office of Management and Budget (OMB), following a set of official published standards. Outside of New England, MAs are comprised of one or more whole counties; in New England, MAs are composed of cities and towns rather than whole counties. Like tracts, MA definitions are updated between censuses—new areas may be added to the MA definition or, less commonly, areas may be removed from the definition. Therefore, care must be taken when comparing MA data from different censuses.

The census has several different categories of geographic entities that fall under the general classification of **place**. Simply put, a place is a city or town, which may or may not be incorporated. Places may cross over other geographic boundaries, such as counties or census tracts/BNAs. Places are identified by a four-digit census code that is unique within a state.

²⁴ Not all U.S. Census data are available down to the block level. The Census short form, which includes only basic demographic and housing characteristics, is provided in block level tabulations, while the Census long form information, which includes a much more extensive set of characteristics, is provided at higher levels of geography.

Each place is also assigned a five-digit FIPS code that is unique within its state. Both the census and FIPS codes are assigned based on alphabetical order of place names.

The final geographic areas that are commonly used are **ZIP codes**. Zone Improvement Plan (ZIP) codes are administrative units established by the United States Postal Service (USPS) for the distribution of mail. ZIP codes are designed for the purpose of efficiently delivering mail, and therefore generally do not respect political or census area boundaries.

There are a number of difficulties with using ZIP codes as geographic units for data analysis. ZIP codes usually do not have clearly identifiable boundaries, often serve a continually changing area, and do not cover all the land area of the United States. Furthermore, ZIP codes are changed fairly frequently to meet postal requirements, and it is difficult to obtain a current “list” of ZIP code areas in the country.²⁵

Nevertheless, ZIP codes are sometimes the only smaller geographic unit available for certain types of data. ZIP codes are identified by five-digit codes assigned by the USPS. The first three digits identify a major city or distribution center, and the last two digits generally signify a specific post office’s delivery area or point.

TOOLS FOR WORKING WITH DATA

Like any good craftsman, a data analyst must depend on a proper set of tools when working with data. For the data analyst, these tools include the different software packages that can be used to carry out the various data management and analysis tasks. Many different types of software have been created to handle these tasks, and, as any craftsman knows, you need the right tool for the job.

This section will describe some of the types of software tools available to the data analyst and explain what they do. Some types of packages are good for basic data entry, but not so good for analysis. Other software can present data graphically, but may not be the best at combining and transforming data files. By becoming familiar with the different types of software that exist, you can make better choices as to which tools best meet your needs.

²⁵ In preparation for the upcoming decennial census, the Census Bureau is currently developing ZIP Code Tabulation Areas (ZCTAs). This will enable the Bureau to provide ZIP-code-level tabulations of 2000 Census data. For more information, see the Census Bureau Web site at <http://www.census.gov/geo/ZCTA/zcta.html> (accessed 19 June 1999).

Spreadsheet Software

Spreadsheet software packages are probably the most familiar to people, as some form of spreadsheet software is included with most PCs that are sold today. The most popular spreadsheet program is Microsoft Excel, which is included in the Microsoft Office software suite, but other spreadsheet programs include Corel Quattro Pro and Lotus 1-2-3.

Spreadsheets basically consist of two dimensional tables of cells. Each cell is a place where a piece of information can be stored, such as a number or a line of text. Cells are identified by their row and column positions. Creating a rectangular data file in a spreadsheet is simply a matter of using separate columns for the variables, and rows for the different observations. Most spreadsheets do not support the concept of “variable names,” however, and identify their columns by a sequence of letters (“A”, “B”, “C”, etc.) or numbers (“1”, “2”, “3”, etc.) One way to better identify the columns is to use the first row in the spreadsheet to enter the “name” of the variable:

	A	B	C	D
1	ID	NAME	AGE	WEIGHT
2	101	Alice Young	27	128
3	102	Bob Smith	45	245
4	103	Cecil Jones	36	162

Spreadsheet programs are useful for data entry of relatively small files. Since they are very familiar to most computer users, they require no special training or setup to be used for this purpose. Spreadsheets do not, however, possess extensive data checking and validation abilities that database programs or special data entry software would have (see section on “Data Entry,” below). In addition, while spreadsheets allow you to mix numeric and text data in the same column, most other types of software do not have this kind of flexibility. This could be a problem if you need to transfer the data from the spreadsheet software to some other type of program.

Beyond creating simple rectangular data files, spreadsheets are very useful for creating formatted tables and charts. Most spreadsheets have formatting tools for drawing lines and justifying text and numbers. Spreadsheets also now come with a wide array of graphic capabilities and can produce good quality charts, including pie, line, and bar charts.

Spreadsheets were originally designed for doing straightforward calculations of numeric data. They have functions for computing sums, averages, and other basic statistics on numbers in groups of cells. Many spreadsheets also now include more sophisticated statistical tools, such as linear regression, and certain database functions, such as table lookup. Nevertheless,

doing complicated data analysis or data file manipulation with spreadsheets is difficult, and spreadsheets are notoriously difficult to “debug.” It is not recommended that spreadsheets be used for anything but the most rudimentary data tasks.

Database Software

Database software packages go beyond the open format of spreadsheets to provide a more structured approach for creating a database—an organized system of related data files. Most database packages use rectangular data files and have incorporated important database concepts such as variables, records, and key variables. They also generally have more sophisticated reporting capabilities than spreadsheets. Among the more popular database packages today are Sapphire’s Dataease, Microsoft’s FoxPro, Microsoft’s Access, and FileMaker’s FileMaker Pro.

Most database packages work directly with rectangular data files. To create a data file, you define the variables in your file, including the variable name and the type of information that it will hold. The types of variables will vary from software to software, although all packages support at least a distinction between numeric and text (or character) variables. In databases, unlike spreadsheets, once a variable’s type is set, you cannot enter information of another type in that variable.

Database programs are useful for data entry. Many allow you to design special data entry forms, rather than entering the data into a spreadsheet-like table. If you are entering data from survey questionnaires, you can design the data entry forms to resemble the questionnaires, which will make data entry easier and help reduce entry errors. Most database programs also allow you to set up data validation rules that will prevent entering incorrect information. For example, if the valid values for a response are the numbers 1 through 4, then a rule can be defined to prevent you from entering a number outside that range.

The most important characteristic of database packages is that they allow you to define **relationships** between your data files and access related information from different files in a consistent way.²⁶ Access to the data is often achieved by executing a **query**—a request for specific information from the database. For example, if you have a survey where you have collected both household-level and person-level data, you can enter this information into two separate files and define a relationship between them. Then you can query the database for a list of all persons who live in households of a certain type, or for a list of all households that have certain types of people living in them.

²⁶ Technically, this is a characteristic of **relational database packages**. Some simpler database programs are not relational and lack this capability.

Standard with most database packages is a report generator, which allows you to summarize the data in many different ways. Many of these are now based on a WYSIWYG design (“What You See Is What You Get”), which means you simply arrange the information on the report creation form exactly as you want it to appear. Once set up, these report formats can then be saved and rerun whenever the data are changed or new data are obtained. Like spreadsheets, database packages can also perform simple statistical operations and create charts and graphs. They are not so good at more sophisticated data analysis, however.

Most database packages now come with some kind of programming language that allows users to manipulate and transform the data in a variety of ways. This gives the user enormous flexibility in terms of how the data can be accessed and analyzed. Nevertheless, programming requires special expertise and can be time-consuming.

GIS Software

A relatively new type of software that is now available for the PC is **geographical information systems (GIS)** software. At the simplest level, GIS software produces maps. More conceptually, a GIS package is a database system that allows you to associate a geographical object with a database record. What makes a GIS system different from an ordinary database that just happens to have geographic information in it, is that GIS software knows how to **display** and **manipulate** geographic data. GIS software can displaying data geographically, for example, by using points to represent locations or shaded regions to represent the characteristics of geographic areas (states, counties, tracts, etc.) It can also access and analyze the data geographically, for example, by selecting all of the points within one mile of a given point or by computing the number of households per square mile in a region.

Several popular GIS packages exist for the PC. The most well-known are MapInfo’s MapInfo Professional, ESRI’s ArcView, and Caliper’s Maptitude.²⁷ All these packages have roughly the same capabilities.

While GIS programs have added more database-like functions, they still do not have all the functionality that can be found in database software. They work with rectangular data files and have some rudimentary data management functions. GIS software is not very good for data entry, because it lacks both data entry forms and automatic data verification. Data manipulation tools are also rather limited. It is often best to manipulate your data in some other package before importing it into GIS.

²⁷ The U.S. Department of Housing and Urban Development (HUD) is currently marketing a modified version of Maptitude under the name Community 2020. This program contains all of the functionality of Maptitude, but includes Census and HUD program data and a series of predefined maps for most metro areas. For more information, see the HUD Web site at <http://www.hud.gov/cpd/2020soft.html/> (accessed 20 June 1999).

What GIS is best at, of course, is the geographical display of data. There are two basic types of maps that can be produced by GIS. **Thematic maps** show areas with different shadings to indicate particular characteristics. For example, a map of census tracts in a city may be shaded with increasing intensities of blue to indicate areas with higher crime rates. A second type of map is a **point map**, which shows individual locations as points on a map. The points can be given different symbols (such as circles, squares, stars, etc.) to represent different types of things. For example, if one has addresses of crimes it is possible to plot them as points on a map and use different symbols for violent versus property crimes, crimes occurring during the day versus those occurring at night, and so on.

Statistical Software

The last type of packages that will be discussed is **statistical software packages**. These are specialized software designed for more complicated manipulation and analysis of data. They include the ability to generate basic descriptive statistics as well as more sophisticated modeling. The most common statistical packages for the PC are SAS Institute's SAS, SPSS Inc.'s SPSS, Stata Corp.'s Stata, and SPSS Inc.'s Systat.

While in the past most statistical software seemed to require a degree in computer science to operate, today's packages are primarily menu driven and have a much more user-friendly design. They allow you to do interactive analysis, rather than creating large programs that have to be run all at once. Nevertheless, to get full benefit from the capabilities of statistical software it is best to be comfortable with some computer programming techniques and concepts.

Most statistical software packages have extensive **data manipulation** capabilities. They allow you to create new variables, select observations based on specified criteria, combine different data sets, and summarize data according to chosen subgroupings. This gives you enormous power to transform your data into more useful forms.

Statistical software can also, of course, produce basic and advanced statistics on your data. You can do simple descriptive statistics, such as means, standard deviations, minimums, and maximums—both for the entire population and for comparison subgroups. You can also do statistical tests to generate confidence intervals, as well as more sophisticated modeling with your data, such as linear regression and cluster analysis. Many packages now have good graphing capabilities for looking at your data visually.

DATA ENTRY

If you have collected data from some primary source, or if you have obtained secondary data in printed form, then you will probably need to enter this information into a computer data

file so that it will be easier for you to access and manipulate. The process of putting information into computer files is called **data entry**. While it sounds like a fairly simple task, if it is not organized properly, you can introduce many errors into your data through typing mistakes. Furthermore, the data entry process provides a good point to check your data for other types of errors, such as those caused by improper or inconsistent responses.

As mentioned above, database packages often have the ability to perform validation of data as it is entered, preventing invalid data from being put into data files. In addition, there are some special data entry packages, such as SPSS Data Entry, that can perform the same functions. The most basic check that is done on data when it is entered is the **range check**, which simply verifies that the entered data falls within an acceptable range. The ranges must be programmed into the system in advance.

A second type of automatic check is the **consistency check**, which tests the relationships between different items. Normally, this is most useful when some redundant information is included on the questionnaire, such as entering a household's total income and then the separate sources of that income. If the total does not equal the sum of the sources, then there may be a data entry error, or a reporting error, or both.

Even if these automatic checks do not reveal any errors, it is still possible that some information was entered incorrectly from the survey questionnaires. For example, an entered value might be in the correct range for a variable, but it may still not match the value that was written on the questionnaire. It is therefore recommended that you make a **visual inspection** of the entered data and compare them with the information on the questionnaires. The person who makes the visual comparison should not be the one who originally entered the data. Alternatively, you can use a procedure called **double entry**, in which two different people enter the same questionnaires into different data files, and then the results are compared. If there are any discrepancies between the two versions, you can make the appropriate correction.

BASIC DESCRIPTIVE STATISTICS

Once the data have been entered into files, it is helpful to begin with some preliminary analysis to learn more about your data. There are some basic descriptive statistics that you should obtain on all your variables, such as minimums, maximums, means, and standard deviations.²⁸ Below is sample output from showing these statistics for two variables, AGE and WEIGHT:

²⁸ These statistical terms are defined in Annex B.

	N	Minimum	Maximum	Mean	Std. Deviation
AGE	200	18	64	39.18	13.82
WEIGHT	200	97.83	279.66	186.2041	60.4427

From this information you can see that there are 200 records in the data file (“N”), as well as the smallest, largest, and mean values for each variable. You should examine these values and see if they seem correct according to what you know about the variables. For example, if the minimum value of weight were 20, this would probably be an error if the file consisted only of adults.

The descriptive statistics shown above apply to numeric variables where the values correspond to some unit on a scale. In this case, age is expressed as the number of years and weight as the number of pounds. For non-numeric or categorical variables, however, these statistics do not apply. For these variables, you may want to do a **frequency table** (also called a contingency table), which shows the number of observations having different values for that variable:

Frequency Table for SEX

Value	Frequency	Percent	Cumulative Percent
F	70	35.0	35.0
M	130	65.0	100.0
Total	200	100.0	

In the above example, you can see the values for the variable SEX are “F” and “M” and that 70 out of the 200 records have the value “F” and 130 have the value “M.” The percentage of these cases out of the total is given in the column labeled “Percent.” The “Cumulative Percent” column keeps a running total of the percentages as you move down the list of values. If some unexpected value were to appear in the frequency table, it would indicate a data problem.

DOCUMENTING DATA

With all of the types of data sources that have been described in this handbook, you can imagine that it can become very complicated to keep track of all the different files that you may

accumulate over time. You need to have a system for organizing all of your data files and remembering important information about them, such as when they were created and where they came from. In other words, you must have good **documentation** for your files.

Having good documentation is an important part of effective data management. Without proper documentation, you may be unable to remember what information all your different data files contain, what variables mean, and what potential problems might exist in the data. Here is a sample format for documenting data files:

File name:	AGEWGT97.DBF
Description:	Age and weight data from adult health survey Data entry file
File format:	dBase IV
Location:	c:\data\survey
Source:	1997 adult health survey
No. records:	200
No. variables:	5

<u>Variable</u>	<u>Width</u>	<u>Type</u>	<u>Description</u>
ID	3	Numeric	Unique person ID
NAME	25	Character	Name of person
AGE	3	Numeric	Age of person (years)
SEX	1	Character	Sex of person: F = Female M = Male
WEIGHT	6.2	Numeric	Weight of person (pounds)
HEIGHT	3	Numeric	Height of person (inches)

Comments:

Height and weight are based on actual measurements made by survey staff.

The first part of the documentation form contains the basic information on the entire data file. The name of the file is identified, followed by a brief description of the file's contents. From this file's description you can see that this is a data entry file created from the survey questionnaires. The file format is given, as well as its location on the computer. The documentation also tells you the source of the data file. In this case, it was an adult health

survey done in 1997, but the source might also be a program (if it is a file created by transforming other data files), or a secondary source (such as the U.S. Census). Finally, the documentation indicates how many records and variables there are in the file.

The second part of the documentation form lists all of the individual variables in the file and provides some basic information about them. Following each variable's name, the form lists the width of the variable. The "6.2" specification for WEIGHT indicates that this variable is six spaces wide and has two digits after the decimal point. Next, the type of variable is given—either numeric or character. In the last column, there is a brief description of the variable. For variables whose values are coded responses, like SEX, the description includes a list of possible values and their meanings.

The last part of the documentation form is where you can enter comments about the data file. This might include clarifications or definitions of terms used in the other parts of the documentation, explanations of how the data were collected or manipulated, or a list of problems that are known to exist in the data.

All of this information is very important to someone who would like to use and analyze this data for a particular purpose. Good documentation will also help you to remember important facts about your data and to avoid wasting time trying to recall what is what or producing invalid results.

REFERENCES

- Bourque, L. B. and V. A. Clark. *Processing Data: The Survey Example*. Quantitative Applications in the Social Sciences Series. Newbury Park, Calif.: Sage Publications, Inc., 1992. Tel: 805-499-9774, <http://www.sagepub.com/>.
- Hartwig, F. and B. E. Dearing. *Exploratory Data Analysis*. Quantitative Applications in the Social Sciences Series. Newbury Park, Calif.: Sage Publications, Inc., 1980. Tel: 805-499-9774, <http://www.sagepub.com/>.
- Tatian, Peter A. . *Designing a Data Entry and Verification System*. Microcomputers in Policy Research Series. Washington, D.C.: International Food Policy Research Institute (IFPRI), 1992. Tel: 202-862-5600, <http://www.cgiar.org/ifpri/>.
- University of Leeds Computing Service. *Overview of Database Systems*. 1996. Available on the Internet as a PDF document at <http://www.leeds.ac.uk/ucs/documentation/ove/ove1.pdf>.

University of Leeds Computing Service. *Overview of Spreadsheets*. 1994. Available on the Internet as a PDF document at <http://www.leeds.ac.uk/ucs/documentation/ove/ove4.pdf>.

U.S. Census Bureau Geographic Areas Reference Manual. Available on the Internet at <http://www.census.gov/geo/www/garm.html>.

*Chapter 6****PRESENTING INFORMATION
EFFECTIVELY***

Now that you have a lot of data, have them organized into files, and have maybe even done some preliminary analysis using them, you need figure out what you will do with them. Presumably you had some ideas about issues that were important to you before you started collecting data. ***But how will you use this information to try to bring about better understanding and possibly even action on these issues?***

Just having good data is not enough—you need to be able to use them and present them effectively. You need to consider the audience you are trying to reach and then use appropriate means to get your message across to that audience. Your message must be clear and easy to understand, and it must be accessible to your audience. If people do not hear your message, then you will not have any impact.

It can often be overwhelming to try to wade through a large quantity of data and figure out messages they may be trying to convey. Extracting key insights from data is a bit of an art form, and something that takes a lot of time and effort to gain proficiency in. The main question to keep in mind as you are sifting through your data is, ***“What story can I tell from this information?”*** Are there any patterns to the data? If you have a series of observations, such as for different points in time, do the values increase or decrease in some direction? Is this what

you would expect? How do the data vary across different places? Across different types of populations? Can you explain these differences?

When examining your data to find stories, you must keep in mind the **audience** that you are trying to address. Presumably, you have had some contact with this audience already and know something about what their interests are. Do they have any preconceived ideas about the issue you are examining? Do the data confirm or contradict these views? If you can tell people something interesting that they did not know before, this will do a lot to pique their interest in what you have to say.

You should also take into account how familiar your audience is with looking at and understanding data. If it is not an audience that is very familiar with data presentation, then you may need to spend more time explaining basic concepts and use more graphical presentations (such as charts and maps), which can be more visually appealing and easier for people to grasp.

Whatever format you choose, your message must be **clear** and **easy to understand**, and your audience should **remember** it. Depending on the length and type of presentation, you should pick no more than two or three main points for your story and stick to those throughout your presentation. Every table, map, or chart should relate somehow to your main points. To help reinforce your message, it is best to state your points at the beginning of your presentation and then repeat them again at the conclusion.

The rest of this chapter will describe three formats for presenting data—tables, charts, and maps—and explain how they should be employed.

TABLES

A table is an ordered presentation of data. The typical organization is rows and columns of cells containing numbers. Different types of tables can be used for various purposes. As Lowry puts it:

One [use of tables] is simply to store information in a compact, readily accessible, and self-documenting form. Another is to persuade a reader that an argument presented in the text of an article, report, or book is valid. Another—regrettably common in research reports—is to demonstrate that the author has done a lot work.²⁹

The first use of tables—storing information in a self-documenting form—is fairly straightforward. It is most often used in situations in which you want to document large amounts

²⁹ Ira S. Lowry, *Designing Readable and Persuasive Tables*, Santa Monica, Calif.: The Rand Corporation, 1983, 1.

of data, such as in an annual indicator compilation or in the annex of a substantive report. These types of tables are useful for people who may want to look up a specific number or piece of information, but they are not very effective for telling stories.

In this section, we will focus mainly on the second purpose—tables intended to persuade someone. An effective and persuasive table is one that presents only the information that is needed to make your point, and does so in a format that allows your audience to see clearly how the data relate to the argument you are making. Remember, your goal is not to show people that you have “done a lot of work.” You want people to understand the story you are trying to tell.³⁰

In fact, you should **consider at first whether you need a table at all**. If your point can be made by using only three or four numbers, then it may be better simply to include them in bullet points or in plain text. In live presentations using overheads or slides projected on a screen, tables are particularly problematic because it is difficult to fit large amounts of numbers in the space available. To fit more numbers, you will have to make the text smaller and, at some point, people will no longer be able to read your table. Consequently, tables are usually best for **printed or electronic** media.

Tables are most useful when you need to show some **pattern** in your data or to illustrate the **relationships** between different numbers. Since it is difficult for people to grasp many relationships at once, you should try to limit yourself to one main point per table. The data in the table should be organized to make it easy for your audience to see the pattern that you are emphasizing. For example, if you want people to make a comparison between two sets of numbers, those numbers should be in adjacent columns or rows in the table.

Here is an example of a fairly good table:³¹

³⁰ A third important use for tables is to help **you** understand your data better. To detect patterns in the data, you may need to create a lot of preliminary tables showing relationships between different variables. These tables would not have to be as well formatted as tables prepared for a presentation. Nevertheless, organizing your data thoughtfully and sticking to only one topic per table will help you analyze this preliminary information more easily.

³¹ This table was adapted from information presented in C. Hayes and M. A. Turner, *Poor People and Poor Neighborhoods in the Washington Metropolitan Region*, Washington, D.C.: The Urban Institute, Publication no. 7425, 1997, <http://www.urban.org/neighborhoods/dcpov.htm> (accessed 24 June 1999).

**Poor Persons in the Washington, DC Metropolitan Area
By Race and Location, 1990**

Jurisdiction	Blacks	Whites
	<i>Number of Poor Persons</i>	
ENTIRE METRO AREA	126,628	88,814
	<i>Share of Metro Area Poverty (%)</i>	
ENTIRE METRO AREA	100.0	100.0
Washington, DC	61.1	14.1
Maryland Suburbs	28.4	44.0
Prince George's County	18.5	14.4
Montgomery County	6.1	18.8
Charles County	1.9	2.8
Calvert County	1.0	1.5
Virginia Suburbs	10.6	41.8
Fairfax	3.5	19.3
Alexandria	2.7	3.2
Arlington	1.7	7.5
Prince William	1.5	5.0
Loudoun	0.5	2.2
Stafford	0.3	2.2
Manassas	0.2	0.8
Fairfax City	0.1	1.0
Manassas Park	0.1	0.2
Falls Church	0.0	0.4

Source: 1990 U.S. Census, Summary Tape File 3

What makes this a good table?

- **The table can stand alone.** There is enough information in the table so that the reader can interpret all of the data.
- **The title is clear and succinct.** The title adequately describes the contents of the table, without providing excessive detail, so that you can easily grasp what the table is about. The title also indicates the geographic area and the time period covered by the table's information.

- **Only essential information is shown.** The table deals with one topic—poor persons in the Washington, D.C., metropolitan area. The author wants you to make comparisons between the shares of Black and White poor persons, which is easy since these are the only two columns. The table does not include a lot of redundant information. For instance, it does not show the total numbers of poor persons in the different locations, which could be calculated from the percentages provided. Such information might be interesting, but if it does not help you get your point across, it is better left out.
- **Totals are placed at the top of the table.** In a table containing percentages, it is often helpful to provide the totals upon which the percentages are based. The first row in the body of this table tells you how many Black and White poor persons there are in the metro area. These numbers are the “100 percent” counts for the columns. (If this information were included in a previous table, then it might be omitted here.) The next row may seem to be superfluous, but it serves a very useful purpose: it tells you that the percentages add up to 100 percent in each of the two columns.
- **The row captions are indented and spaced to show groupings.** Indentation used in the row captions parallels the geographic hierarchy of the areas in the table. The caption for the entire metro area is flush left. At the next level are the three main subareas of the metro area—Washington, D.C., the Maryland Suburbs, and the Virginia Suburbs. Finally, the table lists the individual suburban jurisdictions indented under the main subarea captions. Capitalization and changes in the typeface (bold text) are used to further distinguish these categories. The formatting of the row captions allows you to easily see the relationships of the different geographies, even if you are unfamiliar with the Washington metro area.
- **The rows are sorted in a meaningful order.** The subareas and the suburban jurisdictions are sorted in descending order by the values in the first column—the share of poor Black persons. You can thus quickly see that most poor Blacks in the metro area live in Washington, D.C., and that the largest share of poor Blacks in the Maryland suburbs are in Prince George’s County. (If, on the other hand, you wanted to focus the reader’s attention on the locations of White poor persons, then you would list that column first and sort by that percentage.) If this were a table intended only for documentation purposes, you might list the jurisdictions in alphabetical order. If, however, you are trying to show some particular pattern, then you should sort the rows in some meaningful way based on the data, not alphabetically.
- **Table columns are lined up and formatted neatly.** The numbers in the table columns are right-justified and have the same number of places after the decimal point to make

them easier to compare. For the larger population numbers, there is a comma separating the thousands position, which improves readability. The percentages do not have a percent sign (“%”) after each number, which would be redundant and make the numbers harder to read.

- **Values do not have too many places after the decimal point.** The percentages all show only one place after the decimal point, which is quite adequate for this data. You should avoid the tendency to have excessive precision by showing lots of digits after the decimal point. For example, it is much easier for the reader to compare the numbers “10.9” and “34.6” than it is to compare “10.890034” and “34.62762819”. Furthermore, the additional digits provide little useful information. If you are dealing with very small proportions or very large numbers, then consider changing the scale by converting from percent to per thousand or per million, or from whole units to thousands or millions.
- **The use of lines is limited.** The only lines appearing in the table are those marking the beginning and end of the table body and separating the column headings from the data. This gives a very neat appearance. You should avoid cluttering up your table with lots of extra lines or excessive formatting. Lines should be used only to divide major sections of your table. Note that there are no vertical lines separating the two columns for Blacks and Whites. Such separation is unnecessary and would put a visual barrier between pieces of information that you want the reader to compare.

The source for the table information is given. The last item in the table is an indication of the source of the table’s data. Not only is this a good means of documenting where your information comes from, but showing your sources adds an air of trustworthiness to your table. If this table is part of a written report that more thoroughly documents your sources, then you may only need a very brief description here. If, on the other hand, the table needs to be more independent, then further explanations about sources and manipulations of data may be required. You may also need to add a “Notes” section to provide any other important information the reader may require to understand your table, such as definitions of terms or admissions of problems.

CHARTS

Charts are a graphical means of presenting data, using different symbols, lines, and colors to provide an abstract representation of your information. The most common types of charts are pie charts, bar charts, line charts, and scatter plots. We will not discuss all of the different types of charts here, but rather present some general guidelines and principles on using charts.

The basic principles for creating effective charts are the same as those for making good tables. There should be a “story” behind your chart, and the presentation should make this message clear. The chart should include only those elements needed to make your point, and avoid extraneous information or flashy graphics. In addition, there are more aesthetic issues about the visual appearance of charts—such as the use of appropriate colors or shading. The correct choices can make all the difference in the world in allowing people to easily see the patterns and relationships in the data.

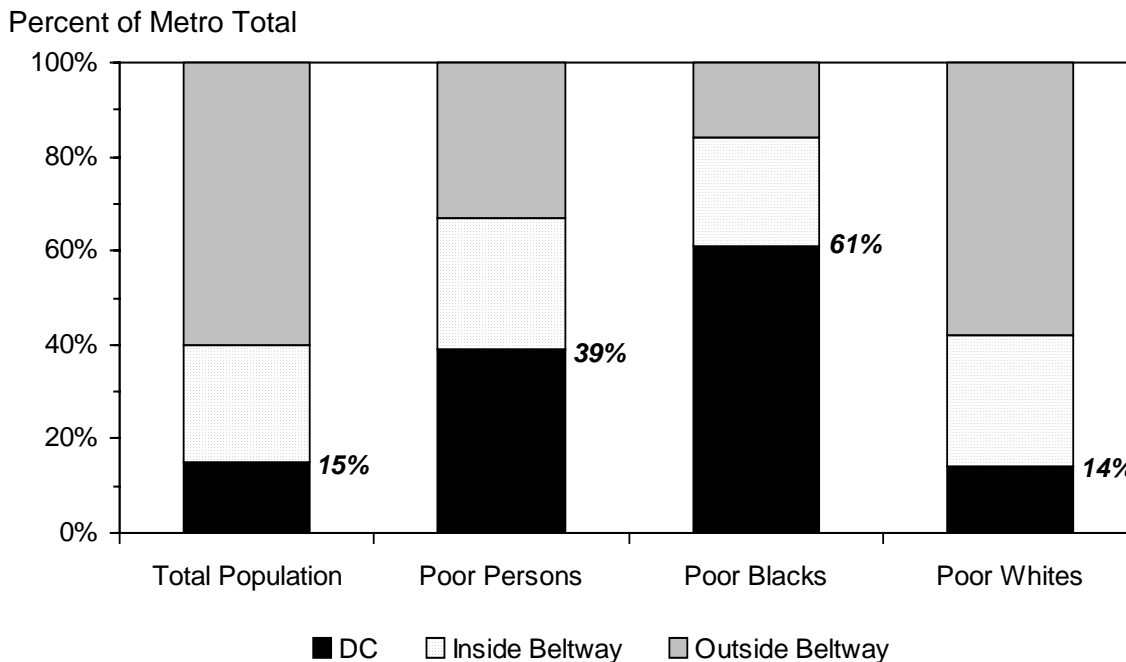
It is important to select the correct type of chart for your purpose. The three major types of charts are:

- **Pie chart.** This is a circular chart in which different “pie slices” represent the shares of different categories of the whole. They should be used rarely, as bar charts are often more effective, and only to show a limited number of categories. You should never use “3-D” pie charts, as they can distort the relationships between the different categories.
- **Bar chart.** This is a series of horizontal or vertical bars divided into sections to represent different categories of data. The individual bars may represent data for different periods of time or different population subgroups. Here, too, “3-D” effects should be avoided as they may distort the relationships between categories.
- **Line chart.** This type of graph shows the relationship between two variables on two axes. It is most useful for showing the change in some indicator over time, where time would be represented on the horizontal axis and the indicator values on the vertical axis. Straight line segments connect the individual indicator values. It is possible to show more than one indicator on the same line chart by using different colors or patterns of lines to distinguish between the indicators. The same technique can be used to represent the values of the same indicator from different populations. You should avoid overloading your chart with too much information, however.

Below is an example of an effective chart. It happens to be a bar chart:³²

³² This chart was adapted from Hayes and Turner, 1997.

Locations of Poor Persons in the Washington, DC Metropolitan Area, 1990



Source: 1990 U.S. Census, Summary Tape File 3

What makes this a good chart?

- **The chart can stand alone.** There is enough information in the chart so that the reader can interpret all of the data.
- **The title is clear and succinct.** The title adequately describes the contents of the chart, without providing excessive detail, so that you can easily grasp what the chart is about. The title also indicates the geographic area and the time period for the chart's information. If this chart were being used for a live presentation, then you might want to replace the descriptive title shown here with a title that conveys your *message*, such as "Most Poor Blacks Live in City; Poor Whites in Suburbs."
- **Only the essential information is shown.** The chart deals with one topic—the location of poor persons in the Washington, D.C., metropolitan area. Only three categories are shown within each bar—the percentages of people living in Washington, D.C., living inside the Beltway, and living outside the Beltway—making the chart very clear. Explicit

values are inserted to emphasize the category of primary interest—the percentage of people living in Washington, D.C.—rather than for all categories.

- ***The heights of the bars are identical.*** In this table we are trying to show the relationships between shares of populations living in different parts of the metro area, so all of the bars are the same height and represent 100 percent of their respective populations. Of course, we could have done this chart by having the height of each bar represent the total number of **people** in each group. But since there are 3.5 million people living in the metro area and only 89,000 poor Whites, the “Total Population” bar would have been about 40 times larger than the “Poor Whites” bar. This would have made it very difficult to compare the shares of these two groups living in different parts of the metro area.
- ***The bar sections are sorted in a meaningful order.*** The category that we are most interested in, the percentage of persons living in Washington, D.C., is on the bottom of each bar. This makes it very easy to compare this percentage across the four populations. The next two sections are ordered geographically, starting with the area closer to the city (“Inside Beltway”) and then moving further out.
- ***The bar sections are visually distinct.*** The bottom category, which we want to emphasize, is colored black so that it will stand out. A lightly shaded pattern (“Inside Beltway”) is inserted between the black and the darker gray (“Outside Beltway”) to give good visual separation between the categories.
- ***Values do not have too many places after the decimal point.*** The percentages are all shown in whole numbers, which is quite adequate for this data. As with tables, you should excessive precision in the form of too many digits after the decimal point.
- ***The use of lines is limited.*** The chart does not have a lot of extra lines, which would only add useless clutter. It is possible to add horizontal **grid lines** extending across the chart to demarcate the different values on the vertical axis. This would make it easier to measure the size of the bar sections. It might also, however, detract from the visual pattern of the bars, and so grid lines have been omitted here.
- ***The source for the chart information is given.*** As with tables, showing your sources is a good means of documenting your work and adding more credibility to your chart.

MAPS

Maps are a very useful way to present geographic data. Most people are familiar with maps, so they are easy to interpret. Maps not only allow you to display values of indicators for different areas, but also permit you to show the geographic relationships between indicator values. For example, are all of the high crime rate census tracts located in a few areas in a city? Do the locations of low-weight births cluster in particular parts of a county?

Maps represent different geographical features as points, lines, or areas. A **point** is a single location, usually representing a street address (“201 North Capitol Street”) or a geographic coordinate expressed in latitude and longitude. A point might be the location of a house, a landmark, or the center of a larger geographic area (often called a **centroid**). A **line** is a series of points connected together in sequence. Lines are used to represent boundaries, streets, railroad tracks, streams, and other linear objects. Finally, an **area** consists of some enclosed region bounded on all sides by lines. States, counties, and census tracts are examples of areas that can be represented on a map.

Each of these three categories of map features can have different visual properties to represent different types of objects. Points can be various **sizes** and use an array of **symbols** to correspond to larger cities versus smaller towns or addresses of public versus private hospitals. Lines can vary in **thickness** and use different **patterns** (solid, dashed, dotted, etc.) to represent major highways, local streets, railroad tracks, city boundaries, and so forth. Areas can be distinguished by using boundaries of different thicknesses and patterns, as well as by using **fill patterns** (solid, shaded, cross-hatched, etc.) Of course, if you have the capability to use **color** in your maps, this can be a highly effective way of representing different types of objects.

The same basic principles for creating effective tables and charts apply for making good maps—having a clear “story” being the most important. You should provide appropriate geographic references on your maps, but avoid cluttering them with too much information. The number of references you must provide will depend on your audience and on the level of geographic precision you need.

You also need to consider the **scale** of your maps. The scale is the degree of reduction used to represent actual places and distances on the map. A good way to express the scale is to relate smaller map units to some larger real-world units, such as “1 inch equals 10 miles.” As an alternative, you can put a graphic **scale bar** on the map (see the example map below), which has the advantage that it will remain correct even if the map is enlarged or reduced.

The scale of your map should be sufficient to show both the extent of the **area** you wish to cover and the level of **detail** that you need to present your data. The larger the area covered by the map, the less detail you can show. If your map has many symbols that need to be

distinguished from each other, then you may need to reduce the area shown to allow that level of detail. In that case, a larger area may need to be covered by several separate maps.

There are two main types of maps used to display data:

Thematic map. In this type of map, areas are shaded or colored according to some varying scale to represent different values of an indicator for each area. For example, a map of census tracts in a city may be shaded with increasing intensities of blue to indicate areas with higher crime rates. A **dot density map** is a particular type of thematic map in which a dot is used to represent certain numbers of things in an area. For example, if there are 2,000 people in a census tract, a dot density map of population showing one dot for every 100 people would display 20 dots scattered throughout this tract.

Point map. A second type of map is a point map, which shows individual locations as points on a map. The points can be given different symbols (such as circles, squares, stars, etc.) to represent different types of things. For example, if you have addresses of crimes, it is possible to plot them as points on a map using different symbols for violent versus property crimes, crimes occurring during the day versus those occurring at night, and so on.

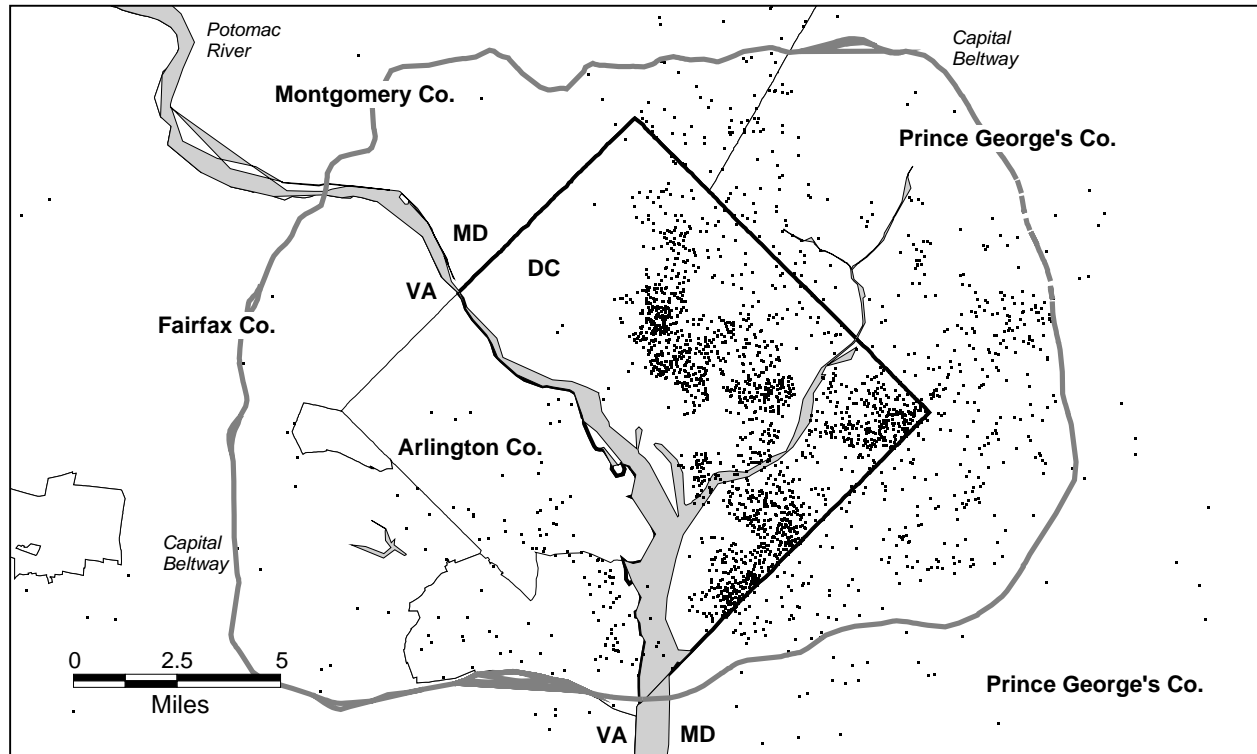
It is also possible to combine thematic and point displays on the same map. For example, you could overlay a point map of crime locations with a thematic map showing average property values in census tracts.

Below is an example of a dot density map showing the locations of poor Black persons in the Washington, D.C., area:³³

³³ This map was adapted from Hayes and Turner, 1997.

Concentrations of Poor Black Persons in the Washington, DC Metropolitan Area, 1990

(1 dot = 40 persons)



Source: U.S. Census, Summary Tape File 3A

Note: Dots do not represent actual locations of people, but are used to indicate the densities of persons in different census tracts.

What makes this a good map?

- **The map can stand alone.** There is enough information in the map so that the reader can interpret all of the data. Sufficient geographic landmarks are given (the state and county boundaries, the Potomac River, the Capital Beltway) to help the reader identify general locations.
- **The title is clear and succinct.** The title adequately describes the contents of the map, without providing excessive detail. The title also indicates the geographic area and the time period covered by the map's information. The subtitle explains what each dot on the map represents.

- **Only the essential information is shown.** The map deals with one topic—the concentration of poor Black persons in the Washington, D.C., metropolitan area. The dot density format chosen is very effective for showing clusters of poor Blacks in certain parts of the city.
- **Selection of geographic features is limited.** Only major geographic features are provided. The river and city boundaries are sufficient for the reader to be able to generally identify key locations and to grasp the main point—that poor Blacks are concentrated in the northeastern and southeastern parts of the city. Including extra features would have added more clutter to the map without giving much additional information, although showing one or two major roads might be helpful in providing additional reference points.
- **The scale and resolution of the map are appropriate.** The map scale is sufficient to show the area of interest. The dots are small enough so that their numbers can be reasonably distinguished.
- **Changes in visual characteristics are used to distinguish different types of objects.** The labels use different capitalization styles and typefaces for identifying different areas. State names are bold and all capital letters, while county names are bold and mixed case. Geographic features—the Potomac and the Beltway—are labeled with italics in a slightly smaller font. The river and highway are also different shades of gray.
- **A map scale is provided.** The scale of the map is indicated by the scale bar in the lower left-hand corner. Since the map is oriented with north at the top, it is not necessary to provide a compass arrow indicating that direction.
- **The source for the map information is given.** As with tables and charts, showing your sources documents your work and adds credibility. An additional note explains an important point for interpreting the map information—that the dots do not represent actual locations of persons but only the numbers of poor Blacks in each census tract. (The census tract boundaries, which would have obscured the map detail, have been left out.)

REFERENCES

Lowry, Ira S. *Designing Readable and Persuasive Tables*, RAND Paper Series, No. P-6945. Santa Monica, Calif.: The Rand Corporation, 1983. Tel: 310-393-0411, <http://www.rand.org/>.

Monmonier, Mark. *How to Lie with Maps*, 2nd ed. Chicago: University of Chicago Press, 1996.

Monmonier, Mark. *Mapping It Out*. Chicago: University of Chicago Press, 1993.

Tufte, Edward R. *Envisioning Information*. Cheshire, Conn.: Graphics Press, 1990.

Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press, 1992.

Tufte, Edward R., Bonnie Scranton, and Dimitry Krasny. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, Conn.: Graphics Press, 1997.

ANNEX A: DATA COLLECTION AND INFORMATION MANAGEMENT PLAN TEMPLATE

The following outlines some questions to guide you in developing a data collection and management plan. The questions start by having you define the issue you are concerned with and describe what you know about it. They then ask you to explore alternatives for obtaining more data (secondary and primary sources) and evaluate the costs and benefits of each source. Finally, you are encouraged to think about how you will use and present the information you collect.

1. Explain what you know about the problem or opportunity to be addressed.

- What is the issue that you wish to know more about?
- What do you know about this problem or opportunity? What data do you already have available to you? Do you have any reason to doubt the accuracy or validity of this information?
- What do you *not* know about this problem or opportunity? What information would you like to have to help you learn more about this issue? To help you evaluate alternative courses of action?

2. Describe the types of data that would help you understand your problem or opportunity better.

- What kinds of information would help you understand your issue better?
- What specific pieces of data would you ideally like to have?
- How would you use these data to analyze your problem or opportunity?
- How important would each of these pieces of information be to helping you understand your problem or opportunity? Which are the most crucial data?
- What would be possible sources for each of these data items?

3. List the data that can be obtained from existing sources.

- Review existing sources of national and local data. Which sources can provide some of the data that you will need?
- How much will it cost to acquire these data?

- What format will the data be in? Will the data require extensive processing for you to be able to use them? What steps will you need to follow for processing the data? How much will the data processing cost? Do you have the necessary expertise and resources for this? If not, can you get them from somewhere else?
- What is the quality of the data from these sources? Are the data reliable enough for your purposes?

4. Describe how to collect data that cannot be obtained from existing sources.

- Which data cannot be obtained from existing sources? Which data are available from existing sources but may not be reliable enough for your use?
- Can you collect any of these data yourself? Will you need to conduct a survey? Focus groups? Key informant interviews? Behavior mapping?
- What steps will you have to follow to collect these data? Do you have the necessary expertise and resources for this? If not, can you get them from somewhere else?
- How much will it cost to collect these data?
- What are the potential quality and accuracy problems that you may face in trying to collect these data? Can you overcome these problems in some way? Will the data be reliable enough for your purposes?

5. Evaluate your options for obtaining data.

- Review the data that you want to obtain—both from existing sources and data that you would have to collect yourself. Are you able to obtain the most important data? What is the total cost of obtaining these data? Can you afford these data?
- If you cannot afford to obtain all the data you want, which data would you choose to get?
- How important are any data you will not be able to obtain to understanding your problem or opportunity? Will the data you can get be sufficient?

6. Describe your plan for using and analyzing data.

- How will you use the data to analyze your problem or opportunity?
- What relationships do you hope to be able to uncover in the data?

- What kinds of tables, charts, and maps will help you display these relationships?
- Who would you like to see and use your data? What different audiences will you be presenting them to? What type of information would each audience be most interested in? What are the best ways to reach these audiences?

ANNEX B: GLOSSARY OF STATISTICAL TERMS

This glossary is intended to provide definitions of some basic statistical terms. Knowledge of this terminology will help you in communicating with statisticians and technical data people, and will be useful in understanding data analysis reports provided by researchers and others.³⁴

Average — See *mean*.

Clustered sample — A sample in which sample members are selected only from certain geographic subregions or sites within the population area. Clustered sampling is used when you need to reduce the costs of data collection by focusing only on certain locations. For example, a clustered sample may select patients from a randomly chosen subset of health clinics in a county (rather than from all sites), or may select households from a randomly chosen subset of street blocks in a city (rather than from all blocks).

Coefficient of variation (CV) — A measure of the variance in a set of data values. The coefficient of variation is equal to the **standard deviation** divided by the **mean**, and is usually expressed as a percentage. The higher the coefficient of variation, the more the values tend to deviate from the mean. Since the coefficient of variation is a percentage or proportion, the variations between different variables can be compared against each other.

Confidence interval — A range based on a sample that has a specified probability (the **confidence level**) of including the true value from a population. A “95 percent confidence interval” of 46 to 62 indicates that the true value has a 95 percent chance of being between these two values.

Confidence level — The probability that a population parameter (such as the mean) falls within a given **confidence interval**. For example, a “90 percent confidence level” indicates that the probability that the true population value falls within the computed confidence interval is at least 90 percent.

Contingency table — See *frequency table*.

Frequency table — Also called a **contingency table**, a frequency table shows the numbers and percentages of different values occurring in data from a population or sample. For

³⁴ Several of these definitions are adapted from Harry P. Hatry, et al., *Customer Surveys for Agency Managers*, Washington, D.C.: The Urban Institute, 1998, 107-108.

example, a frequency table might show the numbers and percentages of people who answered “Yes” and “No” to a given survey question. See also **Chapter V, Basic Descriptive Statistics**.

Margin of error — A amount representing the precision of a given estimate. The margin of error is equal to one half of the size of a corresponding **confidence interval**. For example, a mean value estimate of 20 with a margin of error of 5 would be equivalent to a confidence interval of 15 to 25.

Mean — A measure of the central tendency of a set of data values, the mean usually refers to the **arithmetic mean**, which is calculated as the sum of a set of values (x_i) divided by the total number of values in the population or sample (N):

$$\frac{\sum_{i=1}^N x_i}{N}$$

The mean is generally somewhere “in the middle” of the set of values and is used to represent the **average** or typical value in the population or sample. The mean can be heavily influenced, however, by a few extremely low or high values in the data set. An alternative specification of central tendency is the **median**.

Median — A measure of the central tendency of a set of data values, the median is defined as the point above which and below which half of a set of values falls. Unlike the **mean**, the median is not affected by a few extremely high or low values in the data set.

Population — Also referred to as the **universe**, the population is all the members of a well-defined group. Groups can be defined geographically or by other characteristics, such as ethnicity or gender. For example, a population may be defined as the population of Maryland or all women over 40 who live in California.

Standard deviation — A measure of the variance in a set of data values. The standard deviation describes how a set of values (x_i) are spread around the mean value (μ):

$$\sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N - 1}}$$

The higher the standard deviation, the more the values tend to deviate from the mean. The standard deviation is expressed in the same units of measurement as the data values (miles, hours, kilograms, etc.). This means that you can compare standard deviations of different

variables directly only if those variables are expressed in the same units. See also **standard error** and **variance**.

Standard error — An alternative to **standard deviation** as a measure of variance in a set of data points, the standard error is equal to the standard deviation divided by the square root of the number of values in the population or sample. The higher the standard error, the more the values tend to deviate from the mean. Unlike the standard deviation, standard errors are **not** expressed in the same units of measure as the data values, but in standardized units. Therefore, it is possible to compare standard errors for different types of variables. Standard errors are most often used for constructing statistical tests and creating **confidence intervals**. See also **standard deviation** and **variance**.

Statistical significance — The probability that the sampling error of a statistical estimate is less than a preset level. The preset level is equivalent to 100 percent minus the **confidence level**. For example, a significance level of 5 percent corresponds to a 95 percent confidence level.

Simple random sample — Sometimes just referred to as a **random sample**, a simple random sample is one in which all members of the population have an equal chance of ending up in the sample. The U.S. Census long-form data, for example, are (roughly) a simple random sample of one out of every six households in the United States.

Sample — A subset of a population, usually selected for the purposes of collecting survey data. A sample is generally chosen randomly and is intended to be representative of the population. See also **clustered sample**, **simple random sample**, and **stratified sample**.

Stratified sample — A sample in which members of different subgroups in the population have different chances of being chosen for the sample. Stratified samples are used when you want to be sure to have an adequate number of observations for a particular subgroup. For example, a stratified sample may randomly select equal numbers of Whites and Blacks in a community, even though Blacks represent only 20 percent of the population. Blacks would therefore be overrepresented in the sample relative to their proportion in the population.

Variance — A measure of how much data values vary from the mean value. The variance is equal to the square of the **standard deviation**. It is not very commonly used in data presentations but is found more often in certain types of data analysis.

Regression — Also referred to as **linear regression**, regression is a method of data analysis that attempts to explain the variation in a particular variable in terms of other, related variables. For example, a regression analysis may attempt to explain the number of times

women have received breast cancer screenings in the past five years in terms of personal characteristics (age, education, income, health insurance), family background (health history), and community characteristics (number of health clinics, distance to nearest clinic).

Universe — See ***population***.