1

Statistics and Geography

Most of us encounter probability and statistics for the first time through radio, television, newspapers, or magazines. We may see or hear reports of studies or surveys concerning political polls or perhaps the latest advance in the treatment of cancer or heart disease. If we were to reflect on it for a moment, we would probably notice that statistics is used in almost all fields of human endeavor. For example, many sports organizations keep masses of statistics, and so too do many large corporations. Many companies find that the current production and distribution systems within which they operate require them to monitor their systems leading to the collection of large amounts of data. Perhaps the largest data-gathering exercises are undertaken by governments around the world when they periodically complete a national census.

The word "statistics" has another more specialized meaning. It is the methodology for collecting, presenting, and analyzing data. This methodology can be used as a basis for investigation in such diverse academic fields as education, physics and engineering, medicine, the biological sciences, and the social sciences including geography. Even traditionally nonquantitative disciplines in the humanities are finding increasing uses for statistical methodology.

DEFINITION: STATISTICS

Statistics is the methodology used in studies that collect, organize, and summarize data through graphical and numerical methods, analyze the data, and ultimately draw conclusions.

Many students are introduced to statistics so that they can interpret and understand research carried out in their field of interest. To gain such an understanding, they must have basic knowledge of the procedures, symbols, and *vocabulary* used in these studies.

No matter which discipline utilizes statistical methodology, analysis begins with the collection of data. Analysis of the data is then usually undertaken for one of the following purposes:

1. To help *summarize* the findings of some inquiry, for example, a study of the travel behavior of elderly or handicapped citizens or the estimation of timber reforestation requirements.
2. To obtain a better understanding of the phenomenon under study, primarily as an aid in generalization or *theory validation*, for example, to validate a theory of urban land rent.
3. To make a *forecast* of some variable, for example, short-term interest rates, voter behavior, or house prices.
4. To *evaluate* the performance of some program, for example, a particular form of diet, or an innovative medical or educational program or reform.
5. To help *select* a course of action among a set of possible alternatives, or to plan some system, for example, school locations.

That elements of statistical methodology can be used in such a variety of situations attests to its impressive versatility.

It is convenient to divide statistical methodology into two parts: *descriptive statistics* and *inferential statistics*. Descriptive statistics deals with the organization and summary of data. The purpose of descriptive statistics is to replace what may be an extremely large set of numbers in some dataset with a smaller number of summary measures. Whenever this replacement is made, there is inevitably some loss of information. It is impossible to retain *all* of the information in a dataset using a smaller set of numbers. One of the principal goals of descriptive statistics is to minimize the effect of this information loss. Understanding *which* statistical measure should be used as a summary index in a particular case is another important goal of descriptive statistics. If we understand the derivation and use of descriptive statistics and are aware of its limitations, we can help to avoid the propagation of misleading results. Much of the distrust of statistical methodology derives from its misuse in studies where it has been inappropriately applied or interpreted. Just as the photographer can use a lens to distort a scene, so can a statistician distort the information in a dataset through his or her choice of summary statistics. Understanding what descriptive statistics can tell us, *as well as what it cannot*, is a key concern of statistical analysis.

In the second major part of statistical methodology, *inferential statistics*, descriptive statistics is linked with probability theory so that an investigator can generalize the results of a study of a few individuals to some larger group. To clarify this process, it is necessary to introduce a few simple definitions. The set of persons, regions, areas, or objects in which a researcher has an interest is known as the *population* for the study.

DEFINITION: STATISTICAL POPULATION

A statistical population is the total set of elements (objects, persons, regions, neighborhoods, rivers, etc.) under examination in a particular study.

For instance, if a geographer is studying farm practices in a particular region, the relevant population consists of all farms in the region on a certain date or within a

certain time period. As a second example, the population for a study of voter behavior in a city would include all potential voters; these people are usually contained in an eligible voters list.

In many instances, the statistical population under consideration is *finite;* that is, each element in the population can be listed. The eligible voters lists and the assessment rolls of a city or county are examples of finite populations. At other times, the population may be *hypothetical.* For example, a steel manufacturer wishing to test the quality of output may select a batch of 100 castings over a few weeks of production. The population under study is actually the *future* set of castings to be produced by the manufacturer using this equipment. Of course, this population does not exist and may have an infinitely large number of elements. Statistical analysis is relevant to both finite and hypothetical populations.

Usually, we are interested in one or more characteristics of the population.

DEFINITION: POPULATION CHARACTERISTIC

A population characteristic is any measurable attribute of an element in the population.

A fluvial geomorphologist studying stream flow in a watershed may be interested in a number of different measurable properties of these streams. Stream velocity, discharge, sediment load, and many other characteristic channel data may be collected during a field study. Since a population characteristic usually takes on different values for different elements of the population, it is usually called a *variable.* The fact that the population characteristic does take on different values is what makes the process of statistical inference necessary. If a population characteristic does not vary within the population, it is of little interest to the investigator from an inferential point of view.

DEFINITION: VARIABLE

A variable is a population characteristic that takes on different values for the elements comprising the population.

Information about a population can be collected in two ways. The first is to determine the value of the variable(s) of interest for each and every element of the population. This is known as a *population census* or *population enumeration.* Clearly, it is a feasible alternative only for finite populations. It is extremely difficult, some would argue even impossible, for large populations. It is unlikely that a national decennial Census of Population in a large country actually captures all of the individuals in that population, but the errors can be kept to a minimum if the enumeration process is well designed.

DEFINITION: POPULATION CENSUS

A population census is a complete tabulation of the relevant population characteristic for all elements in the population.

The second way information can be obtained about a population is through a *sample.* A sample is simply a subset of a population, thus in sampling we obtain values for only selected members of a population.

DEFINITION: SAMPLE
A sample is a subset of the elements in the population and is used to make inferences about certain characteristics of the population as a whole.

For practical considerations, usually time and/or cost, it is far more convenient to sample rather than enumerate the entire population. Of course, sampling has one distinct disadvantage. Restricting our attention to a small proportion of the population makes it impossible to be as accurate about population characteristics as is possible with a complete census. The risk of making errors is increased.

DEFINITION: SAMPLING ERROR
Sampling error is the difference between the value of a population characteristic and the value of that characteristic inferred from a sample.

To illustrate sampling error, consider the population characteristic of the average selling price of homes in a given metropolitan area in a certain year. If each and every house is examined, it is found that the average selling price is $150,000. However, if only 25 homes per month are sampled and the average selling price of the 300 homes in the sample (12 months × 25 homes), the average selling price in the sample may be $120,000. All other things being equal, we could say that the difference of $150,000 – $120,000 = $30,000 is due to sampling error.

What do we mean by *all other things being equal*? Our error of $30,000 may be partly due to factors other than sampling. Perhaps the selling price for one home in the sample was incorrectly identified as $252,000 instead of $152,000. Many errors of this type occur in large datasets. Information obtained from personal interviews or questionnaires can contain factual errors from respondents owing to lack of recall, ignorance, or simply the respondent's desire to be less than candid.

DEFINITION: NONSAMPLING OR DATA ACQUISITION ERRORS
Errors that arise in the acquisition, recording, and editing of statistical data are termed nonsampling or data acquisition errors.

In order that error, or the difference between the sample and the population can be ascribed solely to sampling error, it is important to *minimize* nonsampling errors. Validation checks, careful editing, and instrument calibration are all methods used to reduce the possibility that nonsampling error will significantly increase the total error, thereby distorting subsequent statistical inference.

The link between the sample and the population is probability theory. Inferences about the population are based on the information in the sample. The quality of

these inferences depends on how well the sample reflects, or represents, the population. Unfortunately, short of a complete census of the population, there is no way of knowing how well a sample reflects the population. So, instead of selecting a *representative* sample, we select a *random* sample.

DEFINITION: REPRESENTATIVE SAMPLE
A representative sample is one in which the characteristics of the sample closely match the characteristics of the population as a whole.

DEFINITION: RANDOM SAMPLE
A random sample is one in which every individual in the population has the same chance, or probability, of being included in the sample.

Basing our statistical inferences on random samples ensures unbiased findings. It is possible to obtain a very unrepresentative random sample, but the chance of doing so is usually very remote if the sample is large enough. In fact, because the sample has been randomly chosen, we can *always determine the probability* that the inferences made from the sample are misleading. This is why statisticians always make probabilistic judgments, never deterministic ones. The inferences are always qualified to the extent that random sampling error may lead to incorrect judgments.

The process of statistical inference is illustrated in Figure 1-1. Members, or units, of the population are selected in the process of sampling. Together these units comprise the sample. From this sample, whereas inferences about the population are made. In short, sampling takes us from the population to a sample, statistical inference takes us from the sample back to the population. The aim of statistical inference is to make statements about a population characteristic based on the information in a sample. There are two ways of making inferences: *estimation* and *hypothesis testing.*

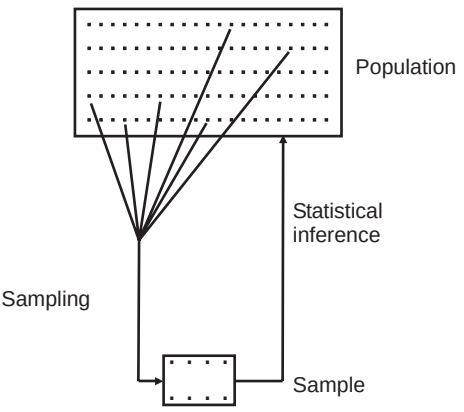

FIGURE 1-1. The process of statistical inference.

DEFINITION: STATISTICAL ESTIMATION
Statistical estimation is the use of the information in a sample to estimate the value of an unknown population characteristic.

The use of political polls to estimate the proportion of voters in favor of a certain party or candidate is a well-known example of statistical estimation. *Estimates* are simply the statistician's best guess of the value of a population characteristic. From a *random sample* of voters, we try and guess what proportion of *all* voters will support a certain candidate.

Through the second way of making inferences about a population characteristic, *hypothesis testing*, we hypothesize a value for some population characteristic and then determine the degree of support for this hypothesized value from the data in our random sample.

DEFINITION: HYPOTHESIS TESTING
Hypothesis testing is a procedure of statistical inference in which we decide whether the data in a sample support a hypothesis that defines the value (or a range of values) of a certain population characteristic.

As an example, we may wish to use a political poll to find out whether some candidate holds an absolute majority of decided voters. Expressed in a statistical way, we wish to know whether the proportion of voters who intend to vote for the candidate exceeds a value of 0.50. We are not interested in the actual value of the population characteristic (the candidate's exact level of support), but in whether the candidate is likely to get a majority of votes. As you might guess, these two ways of making inferences are intimately related and differ more at the conceptual level. The relation between them is so intimate that, for most purposes, both can be used to answer any problem. No matter which method is used, there are two fundamental elements of any statistical inference: the inference itself and a measure of our faith, or confidence in it. A useful synopsis of statistical analysis, including both descriptive and inferential techniques, is illustrated in Figure 1-2.

1.1. Statistical Analysis and Geography

The application of statistical methods to problems in geography is relatively new. Only for about the last half-century has statistics been an accepted part of the academic training of geographers. There are, however, earlier references to uses of descriptive statistics in literature cited by geographers. For example, several 19th-century researchers, including H. C. Carey (1858) and E. G. Ravenstein (1885), used statistical techniques in their studies of migration and other interactions. Elementary methods of descriptive techniques are commonly seen in the geographical literature of the early 20th century. But for the most part, the three paradigms that dominated academic

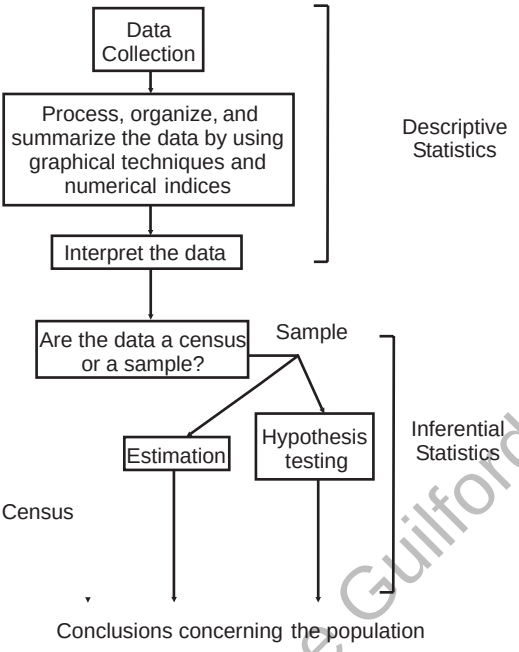

FIGURE 1-2. Statistical analysis.

geography in the first half of the 20th century—exploration, environmental determinism and possibilism, and regional geography—found few uses for statistical methods. Techniques for statistical inference were emerging at this time but were not applied in the geographical literature.

Exploration

This paradigm is one of the earliest used in geography. Unexplored areas of the earth continued to hold the interest of geographers well into the current century. Explorations, funded by geographical societies such as the Royal Geographical Society (RGS) and the American Geographical Society (AGS), continued the tradition of geographers collecting, collating, and disseminating information about relatively obscure and unknown parts of the world. The research sponsored by these organizations helped lead to the establishment of academic departments of geography at several universities. But, given only a passing interest in generalization and an extreme concern for the unique, little of the data generated by this research were ever analyzed by conventional statistical techniques.

Environmental Determinism and Possibilism

Environmental *determinists* and *possibilists* focused on the role of the physical environment as a controlling variable in explaining the diversity of the human impact on

the landscape. Geographers began to concentrate on the physical environment as a control of human behavior, and some determinists went so far as to contend that environmental factors drive virtually all aspects of human behavior. Possibilists held a less extreme view, asserting that people are not totally passive agents of the environment, and had a long, and at times bitter debate with determinists. Few geographers studied human–environment relations outside this paradigm; and very little attention was paid to statistical methodology.

Regional Geography

Reacting against the naive lawmaking attempts of the determinists and possibilists were proponents of regional geography. Generalization of a different character was the goal. According to this paradigm, an *integration* or *synthesis* of the characteristics of areas or regions was to be undertaken by geographers. Ultimately, this would lead to a more or less complete knowledge of the areal differentiation of the world. Statistical methodology was limited to the systematic studies of population distribution, resources, industrial activity, and agricultural patterns. Emphasis was placed on the data collection and summary components of descriptive statistics. In fact, these systematic studies were seen as preliminary and subsidiary elements to the essential tasks of regional synthesis. The definitive work establishing this paradigm at the forefront of geographical research was Richard Hartshorne's *The Nature of Geography*, published in 1939.

Many of the contributions in this field discussed the problems of delimiting homogeneous regions. Each of the systematic specializations produced its own regionalizations. Together, these regions could be synthesized to produce a regional geography. A widely held view was that regional delimitation was a personal interpretation of the findings of many systematic studies. Despite the fact that the map was considered one of the cornerstones of this approach, the analysis of maps using quantitative techniques was rarely undertaken. A notable exception was Weaver's (1954) multiattribute agricultural regionalization; however, his work was not regarded as mainstream regional geography at the time.

Beginning in about 1950, the dominant approach to geographical research shifted away from regional geography and regionalism. To be sure, the transition took place over the succeeding two decades and did not proceed without substantial opposition. It was fueled by the increasing dissatisfaction with the regional approach and the gradual emergence of an acceptable alternative. Probably the first indication of what was to come was the rapid development of the systematic specialties of geography. The traditional systematic branches of physical, economic, historical, and political soon were augmented with urban, marketing, resource management, recreation, transportation, population, and social geography. These systematic specialties developed very close links with related academic disciplines—historical geography with history, social geography with sociology, and so forth. Economic geographers in particular looked to the discipline of economics for modern research methodology. Increased training in these so-called parent disciplines was suggested as an appropriate means of improving the quality of geographical scholarship. Throughout the 1950s and 1960s,

the teaching of systematic specialties and research in these fields became much more important in university curricula. The historical subservience of the systematic fields to regional geography was reversed during this period.

The Scientific Method and Logical Positivism

The new paradigm that took root at this time focused on the use of the *scientific method.* This paradigm sought to exploit the power of the scientific method as a vehicle to establish truly geographical laws and theories to explain spatial patterns. To some, geography was reduced to pure spatial science, though few held this rather extreme view. As it was applied in geography, the scientific method utilized the deductive approach to explanation favored by *positivist* philosophers.

The deductive approach is summarized in Figure 1-3. The researcher begins with a perception of some real-world structure. A pattern, for example, the distance decay of some form of spatial interaction, leads the investigator to develop a model of the phenomenon from which a generalization or hypothesis can be formulated. An experiment or some other kind of test is used to see whether or not the model can be verified. Data are collected from the real world, and verification of the hypothesis or speculative law is undertaken. If the test proves successful, laws and then theories can be developed, heightening our understanding of the real world. If these tests prove successful in many different empirical applications, then the *hypothesis* gradually comes to be accepted as a *law.* Ultimately, these laws are combined to form a theory.

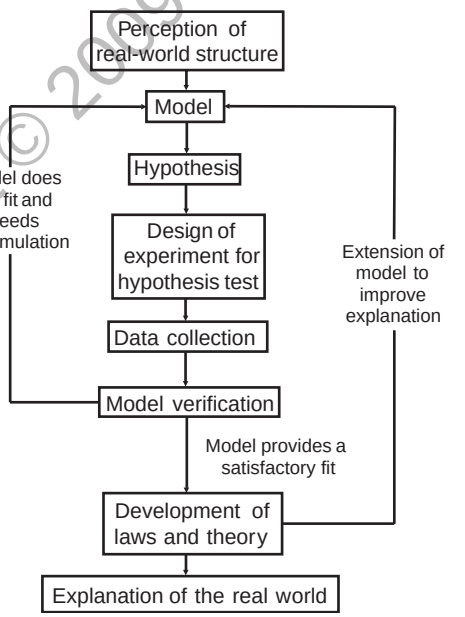

FIGURE 1-3. The deductive approach to scientific explanation.

This approach obviously has many parallels to the methodology for statistics outlined in the introduction to this chapter.

The deduction-based scientific method began to be applied in virtually all fields of geography during the 1950s and 1960s. It remains particularly important in most branches of physical geography, as well as in urban, economic, and transportation geography. Part of the reason for this strength is the widespread use of the scientific method in the physical sciences and in the discipline of economics.

Quantification is essential to the application of the scientific method. Mathematics and statistics play central roles in the advancement of geographic knowledge using this approach. Because geographers have not viewed training in mathematics as essential, the statistical approach has been dominant and is now accepted as an important research tool by geographers. That is not to say that the methodology has been accepted throughout the discipline. Historical and cultural geographers shunned the new wave of quantitative, theoretical geography. Part of the reason for their skepticism was that early research using this paradigm tended to be long on quantification and short on theory. True positivists view quantification as only a means to an end—the development of theory through hypothesis testing. It cannot be said that this viewpoint was clear to all those who practiced this approach to geographic generalization. Too often, research seemed to focus on what techniques were available, not on the problem or issue at hand. The methods *themselves* are clearly insufficient to define the field of geography.

Many researchers advocating the use of the scientific method also defined the discipline of geography as *spatial science*. Human geography began to be defined in terms of *spatial* structures, *spatial* interaction, *spatial* processes, or *spatial* organization. Distance was seen as the key variable for geographers. Unfortunately, such a narrow view of the discipline seems to preclude much of the work undertaken by cultural and historical geographers. Physical geography, which had been brought back into geography with the onset of the quantitative revolution, was once again set apart from human geography. Reaction against geography as a spatial science occurred for several reasons. Chief among these reasons was the disparity between the type of model promised by advocates of spatial science and what they delivered. Most of these theoretical models gave adequate descriptions of reality only at a very general level. The axioms on which they were based seemed to provide a rather poor foundation for furthering the development of geographical theory.

By the mid-1960s, a field now known as *behavioral geography* was beginning to emerge. It was closely linked with psychology and drew many ideas from the rich body of existing psychological research. Proponents of this approach did not often disagree with the basic goals of logical positivism—the development of theory-based generalizations—only with how this task could be best accomplished. Behavioral geographers began to focus on individual spatial cognition and behavior, primarily from an inductive point of view. Rather than accept the unrealistic axioms of perfect knowledge and perfect rationality inherent in many models developed by this time, behavioral geographers felt that the use of more realistic assumptions about behavior might provide deeper insights into spatial structures and spatial behavior. Their inductive approach was seen as a way of providing the necessary input into a set of

richer models based on the deductive mode. Statistical methodology has a clear role in this approach.

Postpositivist Approaches to Geography

Although statistics and quantitative methods seemed to dominate the techniques used during the two decades in the period 1950–1970, a number of new approaches to geographical research began to emerge following this period. First, there were approaches based on *humanistic philosophies.* Humanistic geographers take the view that people create *subjective* worlds in their own minds and that their behavior can be understood only by using a methodology that can penetrate this subjectivity. By definition then, there is no *single, objective* world as is implicit in studies based on positivist, scientific, approaches. The world can only be understood through people's intentions and their attitudes toward it. Phenomenological methods might be used to view the diversity and intensity of experiences of place as well as to explore the growing "placelessness" in modern urban design, for example. Such approaches found great favor in cultural and historical geography.

Structuralists reject both positivist and humanistic methodologies, arguing that explanations of observed spatial patterns cannot be made by a study of the pattern itself, but only by the establishment of theories to explain the development of the societal condition within which people must act. The structuralist alternative, exemplified by Marxism, emphasizes how human behavior is constrained by more general societal processes and can be understood only in those terms. For example, patterns of income segregation in contemporary cities can be understood only within the context of a class conflict between the bourgeoisie on one hand and the proletariat, or workers, on the other. Understanding how power and therefore resources are allocated in a society is a prerequisite to comprehending its spatial organization.

Beginning as *radical geography* in the late 1960s, much of the early effort in this subfield was also directed at the shortcomings inherent in positivist-inspired research. To some, Marxist theory provided the key to understanding capitalist production and laid the groundwork for the analysis of contemporary geographical phenomena. For example, the emergence of ghettos, suburbanization, and other urban residential patterns was analyzed within this framework. More recently, many have explored the possibilities of geographical analysis using variants of the philosophy of structuralism. Structuralism proceeds through an examination of dynamics and rules of systems of meaning and power.

Interwoven within these views were critiques of contemporary geographical studies from feminist geographers. The earliest work, which involved demonstrating that women are subordinated in society, examined gender differences in many different geographical areas, including cultural, development, and urban geography. The lives, experiences, and behavior of women became topics of legitimate geographical inquiry. This foundation played a major role in widening the geographical focus to the intersection of race, class, and sexual orientation, and to how they interact in particular spaces and lives under study.

Human geography has also been invigorated by the impact of *postmodern* methodologies. Postmodernism represents a critique of the approaches that dominated geography from the 1950s to the 1980s and that are therefore labeled as *modernist.* Postmodern researchers stress textuality and texts, deconstruction, reading and interpretation as elements of a research methodology. Part of the attraction of this approach is the view that postmodernism promotes differences and eschews conformity to the modern style. As such its emphasis on heterogeneity, particularity, or uniqueness represents a break with the search for order characteristic of modernism. A key concern in postmodern work is *representation*—the complex of cultural, linguistic, and symbolic processes that are central to the construction of meaning. Interpreting landscapes, for example, may involve the analysis of a painting, a textual description, maps, or pictures. *Hermeneutics* is the task of interpreting meaning in such texts, extracting their embedded meanings, making a "reading" of the landscape. One set of approaches focuses on deconstruction of these texts and analysis of *discourses.* The importance of language in such representations is, of course, paramount. The world can only be understood through language that is seen as a method for transmitting meaning.

The Rise of Qualitative Research Methods in Geography

One consequence of the emergence of this extreme diversity to the approach of human geography is a renewed focus on developing suitable tools for this type of research. These so-called qualitative methods serve not as a competitor but more of a complement to the toolbox, which statistical methods offer to the researcher. The three most commonly used qualitative methods are interviews, techniques for analyzing textual materials (taken in the broadest sense), and observational techniques.

The use of data from interviews is familiar to most statisticians since the development of survey research was closely linked to developments in probability theory and sampling. However, most of the work in this field has focused on one form of interview—the personal interview, which uses a relatively structured format of questions. This method can be thought of as a relatively limiting one, and qualitative geographers tend to prefer more *semistructured* or *unstructured* interview techniques. When used properly, these methods can extract more detailed and personal information from interviewees. Like statisticians, those who employ qualitative methods encounter many methodological problems. How many people should be interviewed? How should the interview be organized? How can the transcripts from an interview be coded to elicit understanding? How can we search for commonalities in the transcripts? Would a different analyst come up with the same interpretations? These are not trivial questions.

In *focus groups,* 6 to 10 people are simultaneously interviewed by a moderator to explore a topic. Here, it is argued that the group situation promotes interaction among the respondents and sometimes leads to broader insights than might be obtained by individual interviews. Statisticians have employed focus groups to help design questionnaires. Marketing experts commonly use them to anticipate consumer reaction to new products. Today focus groups are being used in the context of many different types of research projects in human geography.

Textual materials, whether in the format of written text, paintings or drawings, pictures, or artifacts, can also be subjected to both simple and complex methods of analysis. At one end, simple *content analysis* can be used to extract important information from transcripts, often assisted by PC-based software. Simple word counts or coding techniques are used to analyze textual materials, compare and contrast different texts, or examine trends in a series of texts. Increasingly, researchers are interested in "deconstructing" texts to reveal multiple meanings, ideologies, and interpretations that may be hidden from simple content analysis.

Finally, qualitative methods of observing interaction in a geographical environment are increasingly common. Attempting to understand the structure and dynamics of certain geographic spaces at both the micro level (a room in a building) or in a larger context (a neighborhood or shopping mall) by observing how participants behave and interact can provide useful insights. Observers with weak or strong participation in the environment are possible. Compare, for example, the data likely to be available from a hidden camera recording pedestrian activity in a store, to the data obtained by a researcher living and observing activity in a small remote village. Clearly, one's *positioning* to the observed is important.

All of these techniques have their role in the study of geography. Some serve as useful complements to statistically based studies. For example, when statisticians make interpretations based on the results of surveys, it is often useful to use in-depth unstructured interviews to assess whether such interpretations are indeed valid. A focus group might be used to assess whether the interpretations being made are in agreement with what people actually think. It is easy to think of circumstances where one might wish to use quantitative statistical methods, purely qualitative techniques, or a mixture of the two.

The Role of Statistics in Contemporary Geography

What then is the role of statistics in contemporary geography? Why should we have a good understanding of the principles of statistical analysis? Certainly, statistics is an important component of the research methodology of virtually all systematic branches of geography. A substantial portion of the research in physical, urban, and economic geography employs increasingly sophisticated statistical analysis. Being able to properly evaluate the contributions of this research requires us to have a reasonable understanding of statistical methodologies.

For many geographers, the map is a fundamental building block of all research. Cartography is undergoing a period of rapid change in which computer-based methods are continuing to replace much conventional map compilation and production. Microcomputers linked to a set of powerful peripheral data storage and graphical devices are now essential tools for contemporary cartography. Maps are inherently mathematical and statistical objects, and as such they represent one area of geography where dramatic change will continue to take place for some time to come. This trend has forced many geographers to acquire better technical backgrounds in mathematics and computer science, and has opened the door to the increased use of statistical and quantitative methods in cartography. Geographic information systems (GIS)

are one manifestation of this phenomenon. Large sets of data are now stored, accessed, compiled, and subjected to various cartographic display techniques using video display terminals and hard-copy devices.

The analysis of the spatial pattern of a single map and the comparison of sets of interrelated maps are two cartographic problems for which statistical methodology has been an important source of ideas. Many of the fundamental problems of displaying data on maps have clear and unquestionable parallels to the problems of summarizing data through conventional descriptive statistics. These parallels are discussed briefly in Chapter 3, which focuses on descriptive statistics.

Finally, statistical methods find numerous applications in *applied geography*. Retail location problems, transportation forecasting, and environmental impact assessment are three examples of applied fields where statistical and quantitative techniques play a prominent role. Both private consulting firms and government planning agencies encounter problems in these areas on a day-to-day basis. It is impossible to underestimate the impact of the wide availability of microcomputers on the manner in which geographers can now collect, store and retrieve, analyze, and display the data fundamental to their research. The methodologies employed by mathematical statisticians themselves have been fundamentally changed with the arrival and diffusion of this technology. No course in statistics for geographers can afford to omit applied work with microcomputers in its curriculum.

In sum, statistical analysis is commonplace in contemporary geographical research and education, as it is in the other social, physical, and biological sciences. It is now being more thoughtfully and carefully applied than in the past and includes an ever widening array of specific techniques. Moreover, research using *both* quantitative and qualitative methods is increasingly common. Such an approach exploits the advantages of each class of tools, and minimizes their disadvantages when relying on either alone.

1.2. Data

Although Figure 1-2 seems to suggest that statistical analysis begins with a dataset, this is not strictly true. It is not unusual for a statistician to be consulted at the earliest stages of a research investigation. As the problem becomes clearly defined and questions of appropriate data emerge, the statistician can often give invaluable advice on sources of data, methods used to collect them, and characteristics of the data themselves. A properly executed research design will yield data that can be used to answer the questions of concern in the study. The nature of the data used should never be overlooked. As a preliminary step, let us consider a few issues relating to the sources of data, the kinds of variables amenable to statistical analysis, and several characteristics of the data such as measurement scales, precision, and accuracy.

Sources of Data

A useful typology of data sources is illustrated in Figure 1-4. At the most basic level, we distinguish between data that already exist in some form, which can be termed

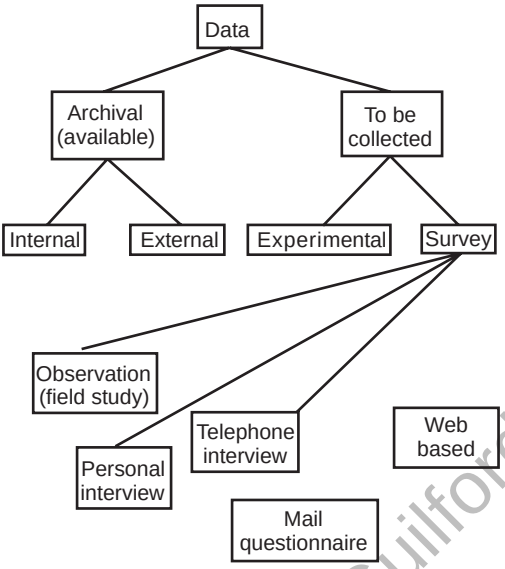

FIGURE 1-4. A typology of data sources.

archival, from data that we propose to collect ourselves in the course of our research. When these data are available in some form in various records kept by the institution or agency undertaking the study, the data are said to be from an *internal source.*

DEFINITION: INTERNAL DATA
Data available from existing records or files of an institution undertaking a study are data from an internal source.

For example, a meteorologist employed by a weather forecasting service normally has many key variables such as air pressure, temperature, and wind velocity from a large array of computer files that are augmented hourly, daily, or other predetermined frequency. Besides the ready availability of this data, the meteorologist has the added advantage of knowing a great deal about the instruments used to collect the data, the accuracy of the data, and possible errors. In-depth practical knowledge of many factors related to the methods of data collection, is often invaluable in statistical analysis. For example, we may find that certain readings are always higher than we might expect. When we examine the source of the data, we might find that all the data were collected from a single instrument that was incorrectly calibrated.

When an *external* data source must be used, many important characteristics of the data may not be known.

DEFINITION: EXTERNAL DATA
Data obtained from an organization external to the institution undertaking the study are data from an external source.

Caution should always be exercised in the use of external data. Consider a set of urban populations extracted from a statistical digest summarizing population growth over a 50-year period. Such a source may not record the exact areal definitions of the urban areas used as a basis for the figures. Moreover, these areal definitions may have changed considerably over the 50-year study period, owing to annexations, amalgamations, and the like. Only a primary source such as the national census would record all the relevant information. Unless such characteristics are carefully recorded in the external source, users of the data may have the false impression that no anomalies exist in the data. It is not unusual to derive results in a statistical analysis that cannot be explained without detailed knowledge of the data source. At times, statisticians are called upon to make comparisons among countries for which data collection procedures are different, data accuracy differs markedly, and even the variables themselves are defined differently. Imagine the difficulty of creating a snapshot of world urbanization collecting variables taken from national census results from countries in every continent. Organizations such as the United Nations spend considerable effort integrating data of this type so that trends and patterns across the globe can be discerned.

Another useful distinction is between *primary* and *secondary* external data.

DEFINITION: PRIMARY DATA
Primary data are obtained from the organization or institution that originally collected the information.

DEFINITION: SECONDARY DATA
Secondary data are data obtained from a source other than the primary data source.

If you must use external data, *always* use the primary source. The difficulty with secondary sources is that they may contain data altered by recording or editing errors, selective data omission, rounding, aggregation, questionable merging of datasets from different sources, or various *ad hoc* corrections. For example, never use an encyclopedia to get a list of the 10 largest cities in the United States; use the U.S. national census. It is surprising just how often research results are reported with erroneous conclusions—only because the authors were too lazy to utilize the primary data source or were unaware that it even existed.

Metadata

It is now increasingly common to augment any data source with a document or database that provides essential information about the data themselves. These so-called metadata or simply "data about data" provide information about the content, quality, type, dates of creation, and usage of the data. Metadata are useful for *any* information source including pictures or videos, web pages, artifacts in a museum, and of course statistical data. For a picture, for example, we might wish to know details about the exact location where it was taken, the date it was taken, who took the picture, detailed

physical characteristics of the recording camera such as the lens used, and any post-production modifications such as brightness and toning applied to it.

DEFINITION: METADATA

Metadata provide information about data, including the processes used to create, clean, check, and produce it. They can be presented in the form of a single or multiple set of documents and/or databases, and they can be made available in printed form or through the web.

The items that should be included in the metadata for any object cannot be precisely and unambiguously defined. While there is considerable agreement about what is to be included in the metadata, many providers augment these basic elements with other items specific to their own interests. In general, the goal of providing metadata *is to facilitate the use and understanding of data.* For statistical data, a metadata document includes several common components:

1. *Data definitions.* This component includes the content, scope, and purpose of the data. For a census, for example, this should include details of the questions asked to recipients, coding used to indicate invalid or missing responses, and so forth. These pieces of information can be obtained by examining the questionnaire and the instructions given to those who apply the questionnaire or the set of instructions given to the interviewers on how to code responses from the recipients.

Several different documents can be included in the metadata for potential users. If the data has been collected from a survey, the original questionnaire is particularly useful since it will contain the exact wording used by interviewers. Since responses are highly sensitive to the wording choices of researchers, this is an essential component of any metadata. After it has been collected, many items are *coded* and assigned numerical or alphabetical codes representing the actual responses by the subject. For example, responses where the subject was unwilling to answer can be coded differently than those where the subject did not know the answer. In addition, responses might be simply *missing* or *invalid.*

Another useful set of information is sometimes available when the data are stored in a database. A *data dictionary* describes and defines each field or variable in the database, provides information on how it is stored (as text, integer, date, or floating point number with a given number of decimal places), and the name of the table in which it is placed. The file types, database schema, processing checks, transformations, and calculations undertaken on the raw data should be included.

If you wish to compare a number of different statistical studies on the same topic, you may find it essential to compare the background information on each data element used in the study. For example, suppose you want to compare vacancy rates in the apartment rental market in several different places. You may find that this task is particularly difficult when different studies have employed different definitions for both rental units and vacancies. While it is generally agreed that a vacancy rate measures the proportion of rental units unoccupied, there will undoubtedly be variations

on how this statistic was actually calculated. Were all rental units visited? Were postal records used to verify occupancy? Were landlords contacted to verify occupancy? As you can see, knowing how the data were collected is almost as valuable as the number itself!

2. *Method of sampling.* Many sources of data are based on samples from populations. How was the sample undertaken? Exactly what sampling procedures were used? How large was the sample? Were some items sampled more intensively than others? For example, when we estimate a vacancy rate, we inevitably combine data from different types of rental units, varying from large residential complexes of over 100 or even 500 units to small apartments rented (perhaps even illegally) by individual homeowners. Sometimes the results of the study may reflect the differential sampling used to uncover these units.

The size of the sample and the size of the population themselves are extremely important characteristics of the data source. A sample of 500 units from a population of a potential 100,000 units in some city is less useful than a sample of 500 taken from a city where the estimated number of rental units is only 10,000. The exact dates of the survey are also important, as vacancy rates vary considerably over the year. It is important to know the currency of the data as well. Situations change rapidly over time. Public opinion data are particularly problematic because they are sometimes subject to radical change in an extremely short period of time.

Sometimes the objects under study are stratified by type, and sampling within each stratum is undertaken independently and at differential sampling fractions. In order to combine the objects into a single result, they must be properly weighted to reflect this differential sampling. For example, in a vacancy rate study we might differentially sample types of units, spending more resources on units with lower rents than on those with higher rents. To combine the results in order to come up with a single measure for the vacancy rate, we apply weights to each type to reflect their relative abundance in the overall housing stock.

3. *Data quality.* When measures of data quality are available, they are also an important indicator of the usefulness of a data source. As we shall see in Section 1.3, we should examine our data for *accuracy*, *precision*, and *validity.* Suppose a study collected some data in the field and used a GPS to determine the location of the phenomenon. Depending on what type of GPS was used, its potential internal error, and the time period over which the coordinates of the location were determined, we may have data of different quality. The quality of the data collected may reflect the precision of the recording instruments, the training and experience of the interviewers, the ability of the survey instrument to yield the answers to questions of interest, and the care taken to verify and clean the data collected.

4. *Data dissemination and legal issues.* Information on how the data can be obtained and how they are distributed is also an important component of metadata. In an era when data are increasingly being distributed electronically, it is now common to specify the procedure for obtaining the data and the particular file formats used. Sometimes data analysis may be undertaken using a statistical package that imports the data provided in one format and alters it to make it compatible with the data commonly analyzed by the program. At times the import process can truncate the data or

change the number of decimal places. Errors can be introduced by file manipulations that truncate rather than round numbers if the number of decimal places is reduced. If the data are disseminated by the original organization that collected the data, this will often ensure that the data used in a study are the best available. This should be apparent in the metadata.

Not all data can be made publicly available, and a considerable number of data sources must deal with legal issues related to privacy and ownership. This is particularly true for data collected on individuals or households where it is possible to suppress the distribution of data that can lead to the identification of an individual household or small group of individuals. For example, figures on incomes earned by households are sensitive and are not normally made available except for large groups of households, for example, census tracts.

5. *Lists of studies based on the data.* It is no longer unusual for data collection agencies or providers to also include in their metadata a bibliography of studies and reports that have utilized the data. These may be internal reports, academic journal articles, research monographs, or other published documents. Being able to see how others have used the data and their conclusions can tell us a lot about the potential issues that may arise in our own study. Suppose an analyst using housing data to estimate vacancy rates feels that the study underestimated the vacancy rate since it placed to much emphasis on high-income properties and ignored low-rent properties that were often advertised only locally in particular neighborhood markets. It would be foolish of us to ignore this result if it might possibly affect the interpretations we developed in our study, which used the data to determine the length of time typical units were vacant.

6. *Geographic data.* Data are at the core of GIS, and metadata are now commonly provided for spatial data so that users can know the spatial extent, locational accuracy and precision, assumed shape of the earth, and projection used to develop a map integral to some dataset. It is obvious that when we are describing *areal* data, we need to know the exact boundaries of places and any changes to these areal definitions over time. For example, several GIS software packages contain a *metadata editor* so that the characteristics of any layer of spatial information can be completely detailed. Developing suitable official standards for geographic metadata is becoming increasingly imprtant.

7. *Training.* Some data collection agencies provide courses that introduce users to data sources, particularly large complex data collection exercises such as a national census. Training and help files are now provided online so that users can know a great deal about the data before beginning their analysis.

More and more, the need for the exchange of statistical data is creating a demand for the effective design, development, and implementation of parallel metainformation systems. As data become increasingly distributed using web-based dissemination tools, software tools that document metadata for statistical data will become increasingly important. As this trend continues, users will be able to undertake statistical analysis of data with a better understanding of the strengths and weakness of the data itself.

Data Collection

When the data required for a study cannot be obtained from an existing source, they are usually collected during the course of the study. It should be clear that any data collection procedure should be undertaken in parallel with an exercise in metadata creation. As our data collection takes place we continually augment our metadata file or document to reflect all characteristics that may be important to users. When, where, what, and how were the data collected? by whom? where? It is almost as difficult to provide accurate metadata as it is to provide the data themselves!

Data acquisition methods can be classified as either *experimental* or *nonexperimental*.

DEFINITION: EXPERIMENTAL METHOD OF DATA COLLECTION
An experimental method of data acquisition is one in which some of the factors under consideration are controlled in order to isolate their effects on the variable or variables of interest.

Only in physical geography is this method of data collection prominent. Fluvial geomorphologists, for example, may use a flume to control such variables as stream velocity, discharge, bed characteristics, and gradient. Among the social sciences, the largest proportion of experimental data is collected in psychology.

DEFINITION: NONEXPERIMENTAL METHOD OF DATA COLLECTION
A nonexperimental method of data collection or *statistical survey* is one in which no control is exercised over the factors that may affect the population characteristic of interest.

There are five common survey methods. *Observation* (or field study) requires the monitoring of an ongoing activity and the direct recording of data. This form of data collection avoids several of the more serious problems associated with other survey techniques, including incomplete data. While techniques based on observation are well developed in anthropology and psychology, their use within geographical research is more recent and limited.

In addition to observation, three other methods of data collection are often used to extract information from households, individuals, or other entities such as corporations or organizations: *personal interviews*, *telephone interviews*, and *web-based interviews*. In a personal interview, a trained interviewer asks a series of questions and records responses on a specially designed form. This procedure has obvious advantages and disadvantages. An alternative, and often cheaper, method of securing the data from a set of households is to send a *mail questionnaire.* This method is often termed *self-enumeration* since the individual completes the questionnaire without assistance from the researcher. The disadvantages of this method include nonresponse, partial response, and low return rates for completed questionnaires. Factors affecting the quality of data from mail surveys include appropriate wording, proper question

order, question types, layout, and design. For telephone and personal interviews there is the added impact of the rapport developed between the interviewer and the subject.

Over time, technological change has had an immense impact on these techniques. Computer-assisted telephone interviewing (CATI) is now the norm with random-digit dialing. Some interviews are now conducted using e-mail or web browser-based collection pages. Important issues related to these techniques include variations in coverage, privacy concerns, and accuracy. Groves et al. (2004) is an especially useful overview of the issues related to all types of survey techniques.

Characteristics of Datasets

Statistical analysis cannot proceed until the available data have been assembled into a usable form.

DEFINITION: DATASET
A dataset is a collection of statistical information or values of the variables of interest in a study.

Geographers collect or analyze two typical forms of *datasets.* An example of the first type, sometimes known as a *structural matrix,* is shown in Table 1-1. In this example, the observational units are climatic stations. Five variables are contained within the dataset. The information on every variable from one observational unit is often termed an observation; it is also common to speak of the data value for a single variable as an *observation* since it is the observed value. In this case, the rows of the dataset represent different locations, and the columns represent the different variables available for analysis. These places might represent areas or simply fixed locations such as cities and towns. Such a matrix allows us to examine the *spatial variation* or *spatial structure* of these individual variables.

Table 1-2 illustrates the second typical form of datasets, an *interaction matrix,* in which the variable of interest is expressed as the flow or interaction between various locations (A through G), which are both the row and column headings of the

TABLE 1-1
A Geographical Dataset

Climatic station	Days per year with precipitation	Annual rainfall, cm	Mean January temperature, °C	Mean July temperature, °C	Coastal or inland
A	114	71	6	16	C
B	42	48	12	16	C
C	54	32	–4	21	I
D	32	28	–8	20	I
E	41	129	16	18	C
F	26	18	1	22	I
G	3	8	24	29	I

TABLE 1-2
An Interaction Matrix

		To						
		A	B	C	D	E	F	G
From	A	58	49	60	91	92	34	14
	B	42	48	12	16	68	25	72
	C	54	32	72	73	63	82	81
	D	45	60	20	28	57	20	46
	E	41	12	17	48	33	99	14
	F	26	18	30	22	10	66	29
	G	13	18	24	19	29	77	29

matrix. Each entry in this matrix represents one observation. Looking across any single row allows us to see the outflows from a single location. Similarly, a single column contains all the inflow into one location. It is easy to see that this matrix can be analyzed in any number of ways in order to search for patterns of spatial interaction.

The variables in a dataset can be classified as either *quantitative* or *qualitative.* Quantitative values can be obtained either by counting or by measurement and can be ordered or ranked.

DEFINITION: QUANTITATIVE VARIABLE
A quantitative variable is one in which the values are expressed numerically.

Discrete variables are those variables that can be obtained by counting. For example, the number of children in a family, the number of cars owned, the number of trips made in a day are all counting variables. The possible values of counting variables are the ordinary integers and zero: 0, 1, 2, . . . , *n.* Quantities such as rainfall, air pressure, or temperature are measured and can take on *any* continuous value depending upon the accuracy of the measurement and recording instrument. *Continuous* variables are thus inherently different from discrete variables. Since continuous data must be measured, they are normally *rounded* to the limits of the measuring device. Heights, for example, are rounded to the nearest inch or centimeter, and temperatures to the nearest degree Celsius or Fahrenheit.

Qualitative variables are neither measured nor counted.

DEFINITION: QUALITATIVE VARIABLE
Qualitative variables are variables that can be placed into distinct nonoverlapping categories. The values are thus non-numerical.

Qualitative variables are sometimes termed *categorical* variables since the observational units can be placed into categories. Male/female, land-use type, occupation, and plant species are all examples of qualitative variables. These variables are defined by

the set of classes into which an observation can be placed. In Table 1-1, for example, climatic stations are classified as either coastal (C) or inland (I).

Numerical values are sometimes assigned to qualitative variables. For example, the yes responses to a particular question in a survey may be assigned the value 1 and the no responses a value of 2. The variable gender may be identified by males = 1 and females = 0 (or vice versa!). In both of these examples, each category has been assigned an *arbitrary* numerical value. As we shall see, it is improper to perform most mathematical operations on qualitative variables expressed in this manner. Consider, for example, the operation of addition. Although this operation is appropriate for a quantitative variable, it makes no sense to add the numerical values assigned to the variable gender.

Besides being described as qualitative or quantitative, variables can also be classified according to the *scale of measurement* on which they are defined. This scale defines the amount of information the variable contains and what numerical operations can be meaningfully undertaken and interpreted. The lowest scale of measurement is the nominal scale. Nominal scale variables are those qualitative variables that have no implicit ordering to their categories. Even though we sometimes assign numerical values to nominal variables, they have no meaning. Consider the variable in Table 1-1 that distinguishes coastal climatic stations (C) form inland ones (I). All that we can really do is distinguish between the two types of stations. In other words, we know that stations A, B, and E are coastal and therefore different from stations C, D, F, and G. Also, stations A,B, and E are all alike according to this variable. One way of summarizing a nominal variable is to count the number of observations in each category. This information can be summarized in a bar graph or a simple table of the following form:

Category	Count or frequency	Proportion	Percentage
C	3	0.429	43
I	4	0.571	57
	7	1.000	100

The use of percentages or relative proportions to summarize such data is quite common. For example, we would say that 0.429, or 43%, of the climatic stations are coastal and 0.571, or 57%, are inland.

Proportional summaries are often extremely useful in comparing responses to similar questions from two or more different surveys, or from different classes of respondents to the same question in a survey. For example, studies of migrants to cities of the Third World often include questions concerning the sources of information used by individual migrants in selecting their destination. A question of this type might be phrased in the following way: In reaching your decision to come here, you must have had some information about job possibilities, living conditions, income, and the like. Which of the following gave you the *most* information? The question would then list a variety of potential sources of information: relatives, friends, newspapers, radio, or

TABLE 1-3
Percentage Distribution of Responses Concerning Primary Information Source Concerning Migration Destinations Used by Migrants

Sources of information	Males	Females
Newspapers	13	7
Radio	3	2
Government labor office	2	3
Family members	40	28
Friends	27	41
School teacher	4	1
Career counselor	1	1
Other sources	10	17
Total	100	100

other. We could compare the proportional use of these information sources from different studies, or by gender or educational attainment of the respondent.

Table 1-3 summarizes the responses to this question and differentiates by gender of the respondent. Clearly, about two-thirds of the respondents cite family or friends as the dominant information source. However, men seem to rely more on family and less on friends in comparison to women. Also, men tend to use newspapers more frequently, and female migrants seem to use other sources more often. These results indicate the significant role played by kin and friends in the rural-urban migration process, but also point to potentially important differences in primary sources of information by gender. An interesting research question is whether or not these observed differences are truly important. If the respondents to the questionnaire were randomly selected, we might be able to use an inferential technique to determine whether this observed difference is due to random sampling or is indicative of an important difference between males and females. Of course, we would also wish to compare the results of a number of similar studies from different cities in different countries before confirming the importance of this hypothesis as a *general* rule. The hypothesis may be limited to the current study.

If the categories of the *qualitative* variable can be put into order, then the scale of measurement of the variable is said to be *ordinal.* An example of an ordinal variable is the strength of opinion measured in responses to a question with the following categories:

Strongly agree	Agree	Neutral	Disagree	Strongly disagree
2	1	0	–1	–2
1	2	3	4	5
5	4	3	2	1

Three different numerical assignments are given, and each is consistent with an ordinal scale for this variable. In all cases, the stronger the agreement with the question, the higher the value of the numerical assignment. In fact, any assignment can be used as long as the values assigned to the categories maintain the ordering implicit in the wordings attached to the categories. The assignment of values –200, –10, 270, 271, 9382 meets this criterion. Note that it doesn't matter which end of the scale is assigned the lowest values, nor does it matter if they are given negative values. All of the categories could be given negative values, all positive values, or a mixture of the two. Of course, the scales defined in the table above have the added advantage of simplicity.

The numerical difference between the values assigned to different categories has no meaning for an ordinal variable. We can neither subtract nor add the values of ordinal variables. Notice, however, that the numerical assignments used to define ordinal variables often use a constant difference or unit between successive categories. In the first scale defined above, each category differs from the next by a value of 1. This does not mean that a respondent who checks the box for strongly agree is 2 units higher than a respondent who checks the box for neutral. We can only say that the first respondent agrees more with the question than does the second respondent. Only statements about *order* can be made using the values assigned to the categories of an ordinal variable.

Quantitative, or numerical variables, whether discrete or continuous, can be classified into two scales of measurement. *Interval* variables differ from ordinal variables in that the interval-scale of measurement uses the concept of *unit distance.* The difference between any two numbers on this scale can always be expressed as some number of units. Both Fahrenheit and Celsius temperature scales are examples of interval-scale variables. Although it makes sense to compare differences of interval scale variables, it is not permissible to take ratios of the values. For example, we can say that 90°F is 45°F hotter than 45°F, but we cannot say that it is twice as hot. To see why, let us simply convert these temperatures to the Celsius scale: 90°F = 32°C and 45°F = 7°C. Note that 32°C is *not* twice as hot as 7°C.

At the highest level of measurement are ratio-scale variables. Any variable having the properties of an interval scale variable as well as a natural origin of zero is measured at the ratio scale of measurement. Distance measured in kilometers or miles, rainfall measured in centimeters or inches, and many other variables commonly studied by geographers are measured on the ratio scale. It is possible to compute ratios of such variables as well as to perform many other mathematical operations such as logarithms, powers, or roots. Because we can take ratios of distances, we can therefore say that a place 200 miles from us is twice as far as one 100 miles from us. While there is a logical and theoretical distinction between ratio and interval variables, this distinction rarely comes into play in practice.

Of far greater interest is the specification of the level of measurement of a variable measured indirectly. For example, the scale of measurement of a variable constructed from the responses to a question or a set of questions in an interview may not be easily identified. This variable may be ordinal, interval, or even ratio. That is, respondents may treat the categories of scale with the labels *strongly agree, agree,*

TABLE 1-4
Levels of Measurement: A Summary

Level	Permissible operations[a]	Examples
Nominal	A = B or A ≠ B, counting	Presence or absence of a road linking two cities or towns Land-use types Gender
Ordinal	A < B or A > B or A = B	Preferences for different neighborhoods in rank order Ratings of shopping center attractiveness
Interval	Subtraction (A – B)	Temperatures °F
Ratio	Addition (A + B) Multiplication (A × B)	
	Take ratios A/B and compare	Distances (imperial or metric)
	Square roots	Density (persons per unit area)
	Powers	Stream discharge
	Logarithms	Shopping center square footage
	Exponentiation	Wheat or corn yield

[a] Permissible mathematical operations at each level of measurement include all operations valid at lower levels of measurement.

neutral, *disagree*, *strongly disagree* as if they were part of an interval, not ordinal, scale. A significant amount of research in psychology has examined methods for deriving scales with interval properties from test questions that are, strictly speaking, only ordinal in character. Attitude scales and some measures of intelligence are two examples where interval properties are desirable. An investigator must sometimes decide what sorts of operations on the variables collected are meaningful, or whether the operations exceed the information contained in the variable. A summary of the scales of measurement, along with a list of permissible mathematical operations and examples, is given in Table 1-4.

1.3. Measurement Evaluation

The utility of any statistical analysis ultimately rests on the quality of the data used, regardless of how sophisticated the analysis itself is. We may be less likely to overstate the significance of our results if we first closely examine the nature of our data. For example, suppose our data suggest that men travel 10 minutes longer than females on the journey to work. On the face of it, this appears to be important. However, when we examine the source of our data, we find that it comes from a self-administered questionnaire in which respondents *reported* their travel times in a simple question expressed as "How long do you spend on your journey to work?" How did respondents answer this question? Did they report their journey time on the day of the survey, or was it based on some sort of average time? Suppose we independently found that over one-half of respondents report travel times that differ from the true travel

time by 20 minutes or more. What does this say about our difference of 10 minutes by gender?

It is therefore *essential* to rigorously evaluate the quality of our data with respect to several key principles—before we undertake any statistical analysis. These principles include measurement *validity, accuracy, and precision.*

Validity

Measurement validity is perhaps the most abstract of these principles. Loosely speaking, it is the degree to which a variable measures what it is supposed to measure. In many physical science applications, validity is usually not an issue. For example, the variable temperature is commonly used as a measure of thermal energy, or heat. The relation between heat and temperature is well known; thus there is not likely to be much debate about whether or not temperature measures an object's energy content. Furthermore, the principles of liquid thermometers are understood, thus it is clear that measurements taken by liquid thermometers do in fact reflect temperature.

DEFINITION: MEASUREMENT VALIDITY
Measurement validity refers to the degree of correspondence between the concept being addressed and the variable being used to measure that concept.

There are, however, instances in both physical and social sciences where the situation is less clear, and measurement validity is questionable. In the first case, it may be that the concept being studied is imperfectly defined. To pick just a few examples, we could mention intelligence, social status, drought, and ecosystem diversity. Everybody has a rough idea of what these concepts are about, but precise definitions are elusive. To take another example, perhaps we want to know if air pollution control policies have improved air quality. Many air quality measures are available, including average pollutant level, maximum level, and number of days above some threshold. If the definition of "air quality" is vague, there will likely be questions about the validity of the variable chosen to measure it. It may be that several different *dimensions* of the term *air quality* need to be addressed, including its maximum, persistence, and regularity. Issues of measurement validity are rarely given the emphasis that they deserve. This often leads to the publication of research results that are misleading or erroneous, or simply wrong.

Accuracy

Another important consideration is *accuracy.*

DEFINITION: MEASUREMENT ACCURACY
Measurement accuracy refers to the absence of error, or the degree of agreement between a measured and true value.

We would say, for example, that a thermometer is accurate if the measured value is close to the true temperature. Note that this does not imply that temperature is a good measure of heat; thus *accuracy* need not imply *validity.* Note also that the definition implies the existence of a "true" value that is at least potentially observable when the measurement is made. Without this assumption, the concept of accuracy has little meaning.

It is convenient to describe the concept of accuracy within an analysis of the errors possible in the measurement process. Let us divide the total error of measurement into two components: *systematic* and *random* error.

Total error = Systematic error + Random error

Ideally then, we would like our total error to be small, which is accomplished by minimizing each of the two components. Let us examine each of the components separately. The first component, *systematic* error, arises if the instrument consistently gives high or low values. Of course, we would like our systematic error to be zero. If it is, we say that our measurements are *unbiased.*

The second component is error that is not attributable to poor calibration of the measurement instrument leading to systematic error, but appears to be random or unpredictable in nature. The actual physical process used to estimate temperature is subject to a range of effects, each of which is individually small but together appear random in nature. Better instruments may be able to reduce this error by controlling for these effects. In this way, our total error is also reduced.

Precision

In general, the *precision* of a measurement refers to the level of exactness or to the range of values possible in the measurement process. A thermometer that can provide estimates within a tenth of a degree is more precise than one that can only provide estimates within a whole degree. In terms of our error equation, precision then is related to the random component of error. It is easy to see that our total error is also reduced by making this as small as possible.

The difference between these two measurement errors is best illustrated within the example of repeated temperature measurements. Let us suppose that four different thermometers A, B, C, and D are used to measure an air temperature known to be 20°C. Five readings are taken with each thermometer over the course of an hour. The recordings from each of the four instruments are illustrated in the temperature scales of Figure 1-5. We can see that thermometer A has no systematic error and leads to readings that may be above or below the true temperature within a range of random error of about one-half a degree. Thermometer B has a systematic error of about +1° since the average reading seems to be around 21°, but the measurements also vary within about one-half a degree. We would say that both of these thermometers are equally *precise,* but thermometer A is more accurate.

Consider now thermometers C and D. Like thermometer A, thermometer C has no bias, but it is much less accurate than A. The problem with C is its low precision,

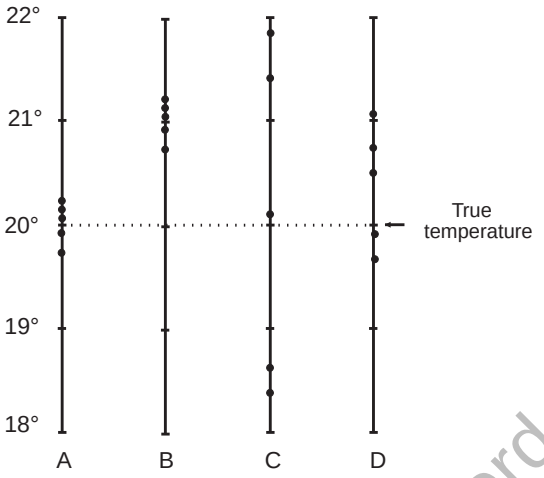

FIGURE 1-5. Differentiating between random and systematic error. • indicates recorded temperature.

since readings from the thermometer vary between roughly 18° and 20°. This is almost four times less precise than A. Thermometer D is biased but has a small random error, leading to an accuracy higher than the unbiased thermometer C. In the search for high overall accuracy, one must consider both sources of error. In particular, there are times when one will prefer a biased instrument over an unbiased alternative simply because the overall errors are smaller. In still other cases the concern is almost exclusively with bias, so that random error components are very much less important.

When the results of an empirical study are analyzed, a good first step is always to closely examine the operational definitions chosen for the variables. Are they suitable? Are they unambiguously defined? Are they the best dataset that could have been used? Are they sufficiently precise? Are all variables unbiased? Could they be responsible for any misleading inferences? The need for caution in the use of data applies to all data sources, including those that are based on experiments as well as those derived from surveys or observation.

1.4. Data and Information

Some historians and other commentators describe the age we live in as the *information age.* This age is characterized by the widespread use and adoption of information communications and technologies. Some argue that one of the fundamental keys to economic success at the firm or institution level—and indeed going all the way to the national level—is the ability to create information and successfully analyze it. *Information technology* is a general term for the technologies involved in collecting, processing, organizing, interpreting, and presenting data. Where does data and statistical analysis fit into this paradigm?

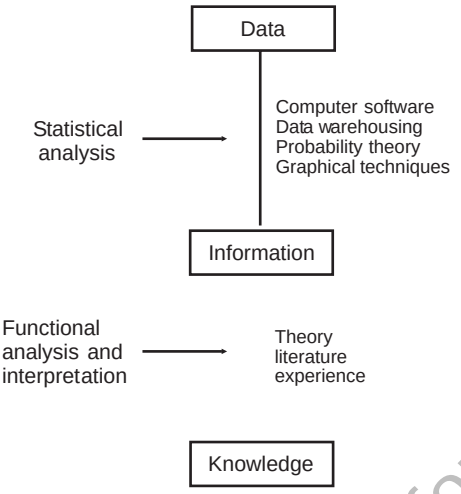

FIGURE 1-6. Statistics and information.

Figure 1-6 displays the concept of the *hierarchy of information* as described by many information system analysts. At the lowest level of the hierarchy is *data.* The meaning of data in this view is quite similar to what we have described in Section 1.3: specific observations of measured numbers. At the next level of the hierarchy, we have *information.* Information is developed from the data by processing and summarizing the data to yield facts and ideas.

DEFINITION: INFORMATION
Data that have been processed into a meaningful form, one that has value to the user, have been transformed into information.

In turn, these ideas are turned into *knowledge.* Knowledge is developed by converting this information using specific theories or understandings based on the techniques and analysis drawn from the specific subject areas. It is selectively organized information that provides understanding, recommends courses of action, and provides the basis for many decisions.

Statistical analysis can thus be viewed as the link *between* data and information. Using concepts of probability theory and also computer software, we examine the available data and search for patterns or facts that will aid us in expanding knowledge within the experience, literature, or theory available to us. Viewed in this way, we can see parallels between, for example, production workers who monitor the quality and productivity of their machines in order to operate and control them, and researchers interested in development studies who seek to understand the role of capital accumulation in economic growth. In each case, data are collected, processed, and summarized in order to create information, and then they are drawn together and interpreted to increase the knowledge of the subject area.

It is now not even uncommon to view information, and therefore data as *power.* Those who can best collect, store, manipulate, analyze, and understand their data will be more powerful. In a global economy, it is argued, such power will be translated to economic success, and firms, or even national states, that embrace information technology and analysis will be at a distinct advantage over those who do not. Data is now being accumulated at an unprecedented rate. Attempting to automate the statistical analysis of this data is leading to the development of new techniques of statistical analysis such as *data mining*, *knowledge discovery,* and *online analytical processing.*

The explosion of data in response to the growth of *automatic* data collection tools means that we may drown in data unless we can develop related automatic tools to turn these data into useful information. The use of conventional statistical analysis software is becoming too time consuming and labor intensive to keep up with the growth of data. Soon, the discovery of patterns, regularities, and simple rules from these databases will only be feasible using a more robust set of tools that are able to uncover previously unknown, nontrivial, and useful information from these extremely large data repositories. To this point, most of the advances in this area have focused on business applications, but these advances will soon spread to other research areas including geography.

1.5. Summary

Statistical analysis includes methods used to collect, organize, present, and analyze data. Descriptive statistics refers to techniques used to describe data, either numerically or graphically. Inferential statistics includes methods used to make statements about a population characteristic on the basis of information from a sample. Statistical inference includes both hypothesis testing and estimation methods.

Within geography, most applications of statistical methodology are rather recent, having become a significant part of the research literature only after the "quantitative revolution" of the 1950s and 1960s. Statistical methodology is most commonly utilized by geographers advocating a scientific approach to the discipline, an approach that is now common in many of the systematic branches of the field. Historical and cultural geographers find fewer uses for the methodology. Recent advocates of humanistic and structuralist approaches tend to be particularly critical of the basis of the scientific method and often reject statistical methodology. Even in these fields, however, there are several instances where statistical methodology can be fruitfully employed in applied research. Recently, the toolbox of human geographers has been enriched by the addition of qualitative techniques. These techniques can often be used hand-in-hand with statistical techniques in many research problems.

One of the first tasks of the statistician is to evaluate the data being used or being proposed in a research inquiry. If suitable data are not already available, or the limitations of existing data sources preclude their confident use, the researcher must collect new data. In some instances, an experimental approach is possible. It is mostly in various systematic specialties within physical geography that this is a feasible alternative, and statistical surveys are more commonly used to collect suitable data.

Whenever statistical surveys are undertaken, it is necessary to proceed with caution, recognizing the limitations of the data collected in this manner. The greater the control exercised in the design of data collection procedures, the better the data ultimately available to the researcher. It follows that the ability of geographers to make sound judgments in their research often rests on the very first steps taken in their research design. Generating precise, accurate, and valid sets of variables can only assist the development of theory and explanation in geography.

REFERENCES

H. Carey, *Principles of Social Science* (Philadelphia: Lippincott, 1858).

R. Hartshorne, *The Nature of Geography* (Lancaster, PA: Association of American Geographers, 1939).

E. Ravenstein, "The Laws of Migration," *Journal of the Royal Statistical Society* 48 (1885), 167–235.

J. C. Weaver, "Crop Combinations in the Middle West," *Geographical Review* 44 (1954), 175–200.

FURTHER READING

Nearly every introductory statistics textbook for social scientists includes a presentation of much of the material discussed in this chapter. See, for example, McGrew and Monroe (2000) for a slightly different approach from that taken here. Bluman (2004) is a slightly more mathematical presentation than this text but contains many simple examples drawn from many fields. For additional information on information science and statistics, see Hand et al. (2000) and Han and Kamber (2001). The book on survey methodology by Groves et al. (2004) provides an excellent overview of the current status of survey methodology. Finally, students wishing to compare statistical methodology to the rapidly growing area of qualitative methods should consult either Hay (2005) or Limb and Dwyer (2001).

Allan G. Bluman, *Elementary Statistics: A Step by Step Approach, 6th ed.* (Boston: McGraw-Hill, 2007).

Robert M. Groves, Floyd J. Fowler, J., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau, *Survey Methodology* (Hoboken, NJ: Wiley-Interscience, 2004).

Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (San Francisco: Morgan Kauffman, 2001).

David J. Hand, Heikki Mannila, and Padhraic Smyth, *Principles of Data Mining* (Boston: MIT Press, 2000)

Iain Hay, *Qualitative Research Methods in Human Geography, 2nd ed.* (London: Oxford University Press, 2005).

Melanie Limb and Claire Dwyer, *Qualitative Methodologies for Geographers: Issues and Debates* (London: Arnold, 2001).

J. Chapman McGrew and Charles B. Monroe, *An Introduction to Statistical Problem Solving in Geography, 2nd ed.* (New York: McGraw-Hill, 2000).

PROBLEMS

1. Explain the meaning of the following terms:
   - Descriptive statistics
   - Inferential statistics
   - Statistical population
   - Population characteristic
   - Variable
   - Population census
   - Sample
   - Sampling error
   - Nonsampling error
   - Representative sample
   - Random sample
   - Statistical estimation
   - Hypothesis test
   - Measurement precision
   - Metadata
   - Inductive approach
   - Deductive approach
   - Internal data
   - External data
   - Primary data
   - Secondary data
   - Experiments
   - Surveys
   - Quantitative and qualitative variables
   - Scale of measurement
   - Ordinal, interval, and ratio scales
   - Measurement validity
   - Measurement accuracy
   - Qualitative methods
   - Data as information

2. Under what conditions might it be advantageous to undertake a population census rather than a sample?

3. What impacts do you think the Internet might have on the distribution and analysis of statistical data?

4. What is the level of measurement of the following variables?
   a. SAT score
   b. Number of tests or quizzes in your statistics course
   c. Acres of land devoted to corn
   d. Number of break-ins in 2004 by neighborhood
   e. Social insurance or Social Security number
   f. Impression of a certain place selected by recipients from a scale of 1 to 5
   g. Name of birthplace
   h. Year of birth

5. Explain why human geographers have undertaken so few experimental studies in their research.

6. Use the Internet to search the government data collection agency in a specific country (e.g., *www.statcan.ca* [Canada], *www.census.gov* [United States], or *www.abs.gov.au* [Australia]). Examine the depth and availability of data from this source. Find another website where the some of the same data can be obtained. Can you locate the metadata for these data, or at least some elements of them?

7. Use the Internet to locate five different sources of data that can be used in statistical analyses. For each source:
   a. Identify the source.

b. Evaluate the data according to the typology of data sources discussed in section 1.3 of this chapter.
c. Edentify any questions you might have before you think the data should be used in any inferential study.

8. The United Nations typically undertakes studies that require the evaluation of data collected in many different countries in many different formats. Locate a study of this type and describe the statistical issues that need to be addressed in this form of analysis.

9. Some research projects are developed using the case study approach. Using this approach, we choose a single observation for detailed study, but we still want to generalize the results of our survey to wider populations. For example, we might choose to study one inner city neighborhood, rather than the inner city as a whole, but we are interested in making conclusions relevant across the city or even to other cities. When do you think case studies are potentially more valuable than surveys? What can be done to generalize the results of the study?