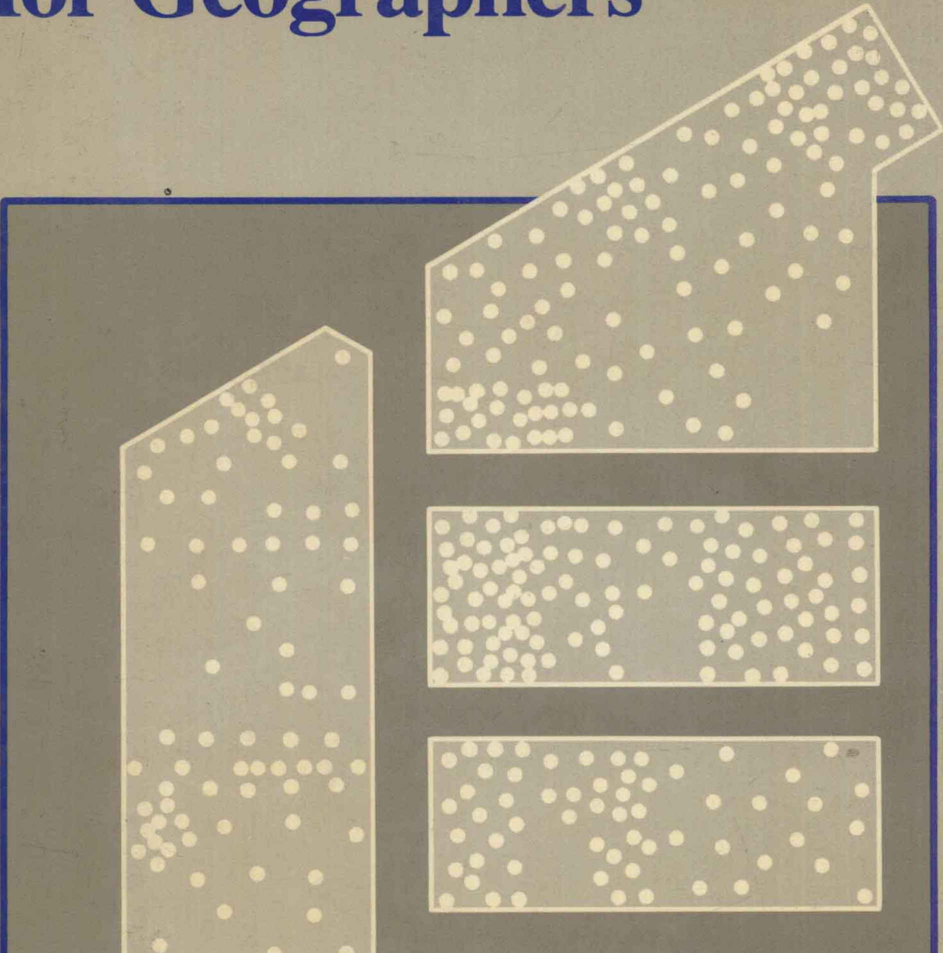
Statistical Methods for Geographers



W.A.V. CLARK P.L. HOSKING

Statistical Methods for Geographers

W. A. V. Clark
University of California, Los Angeles

P. L. Hosking
University of Auckland, New Zealand

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto • Singapore

For Irene and Carol

Library of Congress Cataloging in Publication Data:

Clark, W. A. V. (William A. V.)
Statistical methods for geographers.

Includes index.
1. Geography—Statistical methods. I. Hosking, P. L.
II. Title.

G70.3.C55 1985 910′.21 85-20309
ISBN 0-471-81807-0

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

Few geographers in the early 1950s would have envisaged the changes that were to take place in their discipline over the following twenty years—changes that were probably greater than in any period in the history of geography. Although these changes reflected, in part, new thinking about the philosophy of the subject, primarily they involved a methodological revitalization. In the move to a more "scientific" approach, geographers came to rely heavily on the methods of mathematics and statistics—to describe and test the concepts that previously they were satisfied to formulate subjectively. Although many branches of mathematics contributed ideas to this change, none did so more than the field of statistics, which in its widest sense can be considered as an analysis of information about real-world phenomena as an aid in their description, interpretation and prediction. This change in methodology was not without its problems. Examples of ill-conceived analyses, overexuberance in the use of some methods, and gross errors can be found in abundance in the geographic literature.

The early part of this twenty-year period was marked by a major debate on the implications of the use of mathematical methods, evoking emotional pleas supporting one side or the other. But by the 1970s the arguments had ceased, to be replaced by other more fundamental arguments relating to the philosophy of the discipline; and mathematics—especially statistical methods—had become an accepted and essential tool in geographic research. With a twenty-year background in applying techniques developed by mathematicians (and statisticians in related disciplines), geographers at last could turn to expanding their methodologic armory by developing techniques specifically formulated to handle those problems arising from the very nature of their discipline—the distribution of phenomena over space. The period since the early 1970s has been marked by major developments in what we could simply call *spatial analysis,* and unlike the changes that occurred in the previous twenty years, much of the original development has been initiated through the efforts of geographers themselves.

One of the side-effects of this methodologic change has been an increasing emphasis on a 'modeling approach' in analyzing geographic problems. Many of these so-called models are simply reformulations of ideas presented in less rigorous or less systematic form several years earlier. As such, they may involve only a change in terminology or in presentation. However, the emphasis on models has forced investigators in geography to emphasize the formulation and conceptualization of the problem in which they are interested. Geographers have been forced to search for relevant factors and discard the irrelevant. Tests and verification of hypotheses have become an essential element of geographic investigation. The interest in verification has resulted in both an increased use of numerical methods and a demand for more meticulously constructed research designs. The importance of this preliminary step of preparing a sound research design before data collection begins cannot be overemphasized.

This book provides insights into research design as well as basic methods for testing hypotheses and analyzing functional relationships, very much like other statistical texts. However, there are some important differences. First, despite several introductory books by geographers, there is a need for a comprehensive book that focuses on geographic examples and spatial problems. When using a book written for, say, sociologists, many students have difficulty transferring the statistical concepts to the types of geographic problems they are meeting in their own work or that are being analyzed in their geography courses. In this book, there is a strong emphasis on the presentation of specific methods related to spatial data and perhaps more importantly, an emphasis on the problems of applying standard statistical methods to data distributed over space.

Second, this book emphasizes the use of statistical routines which are now readily available on most computers. The major emphasis is on SPSS and SAS, although BMD is introduced for specific problems. The book includes examples of raw data, statistical set-ups and the output of actual runs. For the advanced techniques in the text, this is a particularly useful way of grasping the nuances of their application to geographic problems.

The decision to use SPSS and SAS (occasionally BMD) statistical packages also requires a comment. With the proliferation of personal computers, one approach would be to gear the text to the uses of these machines with interactive statistical packages. We believe, however, that we are presently in a transition phase and that the large packages such as SPSS and SAS will be "down-loaded" to the newer and faster personal computers in the near future (SPSS is available for the new IBM PC-AT). Thus, a background in SPSS and SAS will provide a basis for their use on smaller machines and in an interactive mode.

The book is designed to be used at the upper division undergraduate level and first year graduate level in North American, British, Australian and New Zealand universities. The first nine chapters could form the basis for a one quarter or one term course introducing geographers to the whole field of statistical methods, to be followed by a further course on multivariate methods (Chapters 10–13); or the book could be used as a one-quarter "review and extension" course for those who have been introduced to statistical methods in a statistics course in mathematics or statistics. These courses would provide the basis for graduate courses on probability-based models and spatial statistics. The book emphasizes verbal interpretation of equations. Mathematical derivations and developments are carefully explained. No mathematical background is assumed, and mathematical concepts are introduced as they are needed.

The first three chapters introduce terms, and ways of displaying and describing distributions. Chapters 4 and 5 focus on probability and sampling methods and are followed by three chapters on statistical testing. Chapter 9 introduces simple linear regression. Chapters 10 to 13 are concerned with extensions of the simple linear model to multiple regression and its assumptions, stepwise logit regression, canonical and discriminant analysis. The focus throughout the more advanced topics is on one or more dependent variables as they relate to sets of independent variables and for this reason, we have not included the data exploratory techniques of component and factor analysis.

We would like to acknowledge the help of many colleagues and students, both present and past, who have contributed directly and indirectly to this book. The material presented here has been worked out over a number of years in presentations to classes both in the United States and New Zealand. We would like to thank specifically Michael Hollis for running the programs and analyzing the examples used in the book, and Maggi Sokolik for typing and proofreading the many drafts that this book has gone through. We are also grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd. for permission to reprint Table V from their book, *Statistical Tables for Biological, Agricultural and Medical Research* (6th Edition, 1974).

W. A. V. Clark
P. L. Hosking

Contents

CHAPTER 1
An Introduction to Statistical Methods 1

1.1 Populations and Samples 2
1.2 Variables and Their Measurement 6
1.3 The Nature of Geographic Data 16
1.4 The Analysis of Geographic Data 23
1.5 The Compass of Quantitative Geography 30

CHAPTER 2
The Display of Distributions 35

2.1 Tabulation 36
2.2 Graphic Display 45
2.3 Mapping Spatial Distributions 51
2.4 Computer Methods 62

CHAPTER 3
Statistical Summaries of Distributions 67

3.1 Curve Description 69
3.2 Measures of Central Tendency 74
3.3 Variability, Skewness, and Kurtosis 82
3.4 Measures of Relative Position 88
3.5 Sediment Size Analysis 91
3.6 Statistical Moments of a Distribution 93
3.7 Measures for Spatial Distributions 98

CHAPTER 4
Probability and Probability Functions 109

4.1 Probability Definitions 110
4.2 Mathematical Properties of Probabilities 114
4.3 Random Variables and Probability Functions 121
4.4 Univariate Probability Functions 136

CHAPTER 5
Sampling Designs and Sampling Methods 149

5.1 Sampling and Inferential Statistics 150
5.2 Methods of Sampling 155
5.3 Sampling Geographic Distributions 164
5.4 Sample Distributions and Sampling Distributions 167

CHAPTER 6
Statistical Inference: Fitting Probability Functions 177

6.1 Selection of the Probability Function 179
6.2 Parameter Estimation 184
6.3 The Normal Probability Function 189
6.4 The Lognormal Probability Function 200
6.5 Other Probability Functions 208

CHAPTER 7
Statistical Inference: Interval Estimation and Hypothesis Testing 219

7.1 Principles of Interval Estimation 220
7.2 Interval Estimation for Population Parameters 227
7.3 Principles of Hypothesis Testing 233
7.4 Hypothesis Testing for Population Parameters 239
7.5 Tests for Goodness of Fit 242
7.6 Statistical Inference with Computers 247

CHAPTER 8
An Introduction to Bivariate Relationships 249

8.1 Testing Differences of Summary Measures 250
8.2 Contingency Tests and Associated Structures 261
8.3 Analysis of Variance 266

CHAPTER 9
The Simple Linear Regression Model 279

9.1 Bivariate Functional Relationships 281
9.2 Steps in the Analysis of a Functional Relationship 291
9.3 Linear Regression Analysis 293
9.4 Linear Correlation Analysis 310
9.5 Assessment of the Simple Linear Regression Model 316

CHAPTER 10

The General Linear Model—Multiple Regression 333

10.1 The Multiple Regression Model 334
10.2 The Coefficients of the Model 338
10.3 Matrices 349
10.4 Matrix Solutions for the Coefficients of the General Linear Model 354
10.5 Tests for the Regression Model 358

CHAPTER 11

Issues in the Application of the General Linear Model 365

11.1 Assumptions of the General Linear Model 366
11.2 The Analysis of Residuals 369
11.3 Normality, Linearity, and Homoscedasticity 373
11.4 Autocorrelation 378
11.5 Multicollinearity 400
11.6 Aggregation and Ecological Correlation 404

CHAPTER 12

Extensions of Multivariate Linear Regression Methods 411

12.1 Variable Transformations 413
12.2 Stepwise Regression and the "Best" Model 419
12.3 Analysis with Dummy Variables 426
12.4 Trend Surface Analysis 433

CHAPTER 13

Alternative Forms of Multivariate Analysis 447

13.1 Logit Models 448
13.2 Canonical Correlation 466
13.3 Discriminant Analysis 479

Statistical Tables

A Random Numbers 501
B Ordinates of the Standard Normal Density Function 503
C Cumulative Standard Normal Distribution 504
D Cumulative *t* Distribution 505
E Cumulative Chi-Square Distribution 506
F Critical Values of *D* in the Kolmogorov-Smirnov One-Sample Test 507

G Cumulative *F* Distribution 508
H Critical Values of the Durbin-Watson Statistic 511

Index 513

CHAPTER 1
An Introduction to Statistical Methods

1.1 Populations and Samples:
Elements and populations; Characteristics of populations; Sampling from populations; Characteristics of samples.

1.2 Variables and Their Measurement:
Characteristics of variables; Measurement of observed variables; Measurement of derived variables; Variable measurement and errors; Levels of precision and the recording of scores.

1.3 The Nature of Geographic Data:
The spatial component of geographic data; Spatial distributions, spatial analysis, and statistical methods; Sources of data.

1.4 The Analysis of Geographic Data:
Setting up the design, coding, and recording scores; Introduction to the use of SPSS for statistical analysis; The mechanics of statistical analysis.

1.5 The Compass of Quantitative Geography

In a book about the use of statistical methods in geography, it is appropriate to begin with a few brief definitions. The word *statistics* can take on a variety of meanings. In everyday usage, it simply refers to a set of data; but in the physical and social sciences, *statistics* also refers to a body of knowledge—a part of the field of mathematics or, more particularly, applied mathematics. In this sense, statistics or statistical methods has been defined as being concerned with the analysis of information about real-world phenomena as an aid in their description, interpretation, and prediction. More specifically, statistics involves the analysis of *distributions* of *scores* for *variables* derived from the *measurement* of *elements* from a *population* or *sample*. All these italicized terms require formal definitions—and to a considerable extent, that is what this introductory chapter is about—but for most of them, we probably already have some intuitive feeling about what is implied in their use.

The material that makes up the subject of statistics can be subdivided in a number of ways. First, we note a very broad separation into two fields: *descriptive statistics,* concerned with describing phenomena (leading ultimately to methods of interpretation and prediction); and *inferential* or *inductive statistics,* a set of methods to enable conclusions to be made about populations from a subset of that data—a sample. Another way of looking at statistics is in terms of the complexity of the problem—whether we are interested in just the description of a single variable (*univariate statistical methods*) or whether we are interested in the interaction (relationship) of two variables (*bivariate statistical methods*) or more than two variables (*multivariate statistical methods*).

1.1 POPULATIONS AND SAMPLES

Before attempting to outline even the most elementary statistical methods, we must examine carefully some basic concepts required to set up a piece of research involving statistical analyses. We emphasize again and again throughout this book the need to employ a careful, systematically derived

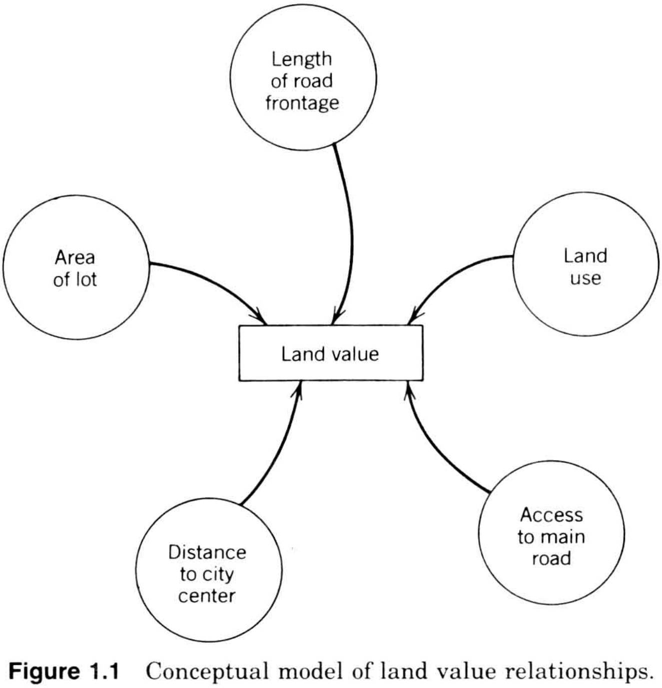


Figure 1.1 Conceptual model of land value relationships.

research design. To assist in this process, we use *operational definitions*—carefully constructed descriptions or sets of rules outlining explicitly how each step or process in an analysis is to be performed. Throughout this introductory chapter, we use a single example—an example that although simple, illustrates one type of research problem faced by geographers. In a small city of, say, 15,000 people, we are interested in how land value varies throughout the central area of the city. We want to describe and analyze this variation and to attempt to explain the pattern of variation. We refer to this example as the *Urban Land Value Survey.* As a first step in the analysis, a simple conceptual model of possible relationships between land value and other potential influencing factors is prepared (Figure 1.1). We postulate that land value is a function of land use, distance to city center, and so on, where the changing values for one variable—land value in this case (called the *dependent variable*)—is dependent on (is a function of) the changing values for the other variables (called the *independent variables*). Statistical methods can be used to describe and test various parts of simple models of this sort.

Elements and Populations

Applying statistical methods to describe real-world phenomena is based on the process of *measurement*—the conversion of some characteristic of a phenomenon into symbols (numbers). As we will see, this measurement process can be interpreted in a much wider sense in statistics than in its normal usage. The measurement process is directed at an individual entity—the smallest sampling unit about which measures are to be obtained and to which we will give the general name *element.* Alternative names for the element found in many texts are *case, observation,* or *observational unit.* An element may be a readily observed entity: for example, a person, a plant, a town, or a rainfall station; but in some situations the identity of the element that is being analyzed may not be as apparent. In geographic work, it is common to find that the element under consideration consists of a collection of a number of *subelements,* each of which has to be subjected to some form of actual measurement. Some examples might be the study of sediment size over a beach where the elements are individual sediment samples consisting of individual grains (subelements) or a study of the population characteristics of a city based on census information where the element would be the census subdivision (census tract) consisting of a specified number of individual people (subelements). In these situations, measuring the characteristics of the element would be derived from measuring the individual subelements. The distinction between these two types of elements (those of indivisible entities and those with subelements), although not always readily apparent, is of critical importance in preparing the research design. We defer a detailed discussion of the grouping of subelements and its implications until much later in the text.

Statistics, then, is concerned with the analysis of elements. The nature of the analysis is determined by the characteristics of the elements to be measured (the *variables*—examined in Section 1.2) and by the "area of interest" of the analysis. This area of interest may be defined as the *universe* or *population*—the totality of all individual elements being examined.

Characteristics of Populations

A population consists of a set of *elements*. We need an operational definition of the element under study. For example, an operational definition for an element in the Urban Land Value Survey might be:

> **the recorded *lot* as specified on the map of the City of Woodford (Valuation map 82/01 Sheet 7) dated May 24 1982.**

The population itself requires an operational definition, in conjunction with and directly related to the defined element. For example:

> **all rεcorded lots within the area demarcated as Central Business District Core by the Woodford City Planning Department and recorded on Planning Map C/4 dated January 1 1982.**

A population is also defined by its size—the number of included elements. The population size, usually designated by uppercase N, may be infinite, finite and known, or finite and unknown. In the Urban Land Value Survey it is finite and, using the operational definitions of an element and the population outlined above, $N = 86$.

The above examples using the Urban Land Value Survey are simple, but they illustrate the sort of problem frequently faced by geographers. We have not yet mentioned the type of information that is to be measured from the population of elements, but this obviously would have been determined before decisions were made as to the operational definitions outlined. The full implications of these operational definitions will become more apparent when the whole survey design is discussed in Section 1.4. Some examples of populations and their elements are illustrated in Figure 1.2.

Sampling from Populations

In situations where the size of the population is infinite, obviously not all elements can be measured, and some means of selecting a set of elements—a *sample*—from the population must be devised. Even with finite populations, it may not be possible to examine all the elements of a population for reasons of time, cost, or inaccessibility. In fact, it may not even be desirable to examine a complete population, because statistical methods are so powerful that valid conclusions can be "inferred" about a population simply from information obtained from a sample.

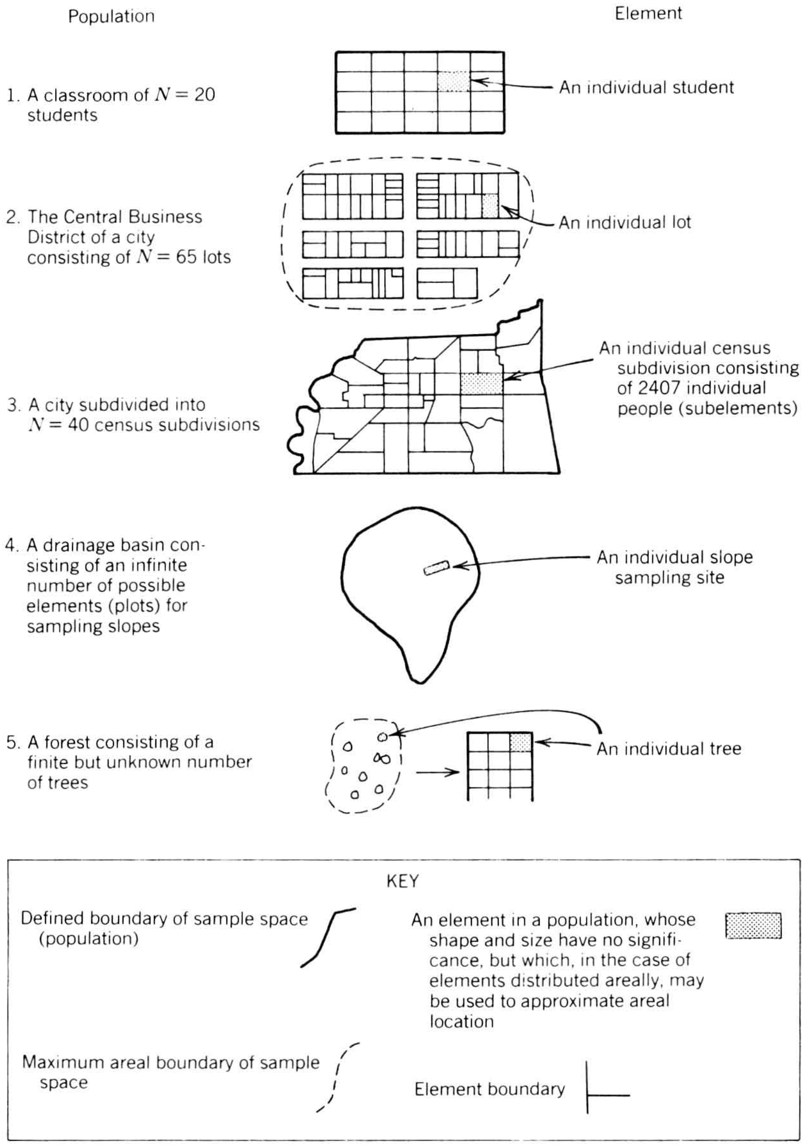

Population
Element
1. A classroom of $N = 20$ students
An individual student
2. The Central Business District of a city consisting of $N = 65$ lots
An individual lot
3. A city subdivided into $N = 40$ census subdivisions
An individual census subdivision consisting of 2407 individual people (subelements)
4. A drainage basin consisting of an infinite number of possible elements (plots) for sampling slopes
An individual slope sampling site
5. A forest consisting of a finite but unknown number of trees
An individual tree
KEY
Defined boundary of sample space (population)
An element in a population, whose shape and size have no significance, but which, in the case of elements distributed areally, may be used to approximate areal location
Maximum areal boundary of sample space
Element boundary

Figure 1.2 Diagrammatic representation of some populations and elements.

If we define a sample as a subset of elements from the population, then the interpretation of the information obtained from the sample is dependent on the method of selection of the subset of elements. Purposeful sampling may be defined as the procedure whereby elements are selected from a population because they have certain characteristics in which we are interested. For example, in a study of urban labor force characteristics, 5 cities from a population of 50 cities are selected because they are in a region of interest, because they are the largest, because they are growing the fastest, etc. The extent to which we can regard this sample as being "representative" of the population depends entirely on our definition of the population. As a general statement, however, we can say that the method just outlined would produce a sample from which information obtained could refer only to the sampled elements. More usually, our aim would be to select a sample from which information could be used to make conclusions about the entire population from which the sample was selected. Such an inference requires that the sample was based on probability sampling methods (to be discussed in Chapter 4), or, more specifically, a simple random sampling method, where each element of the population has an equal chance of being selected into the sample.

Characteristics of Samples

Samples consist of elements drawn from a population. Thus, we need operational definitions of an element and of the population itself.

The sample, or more accurately, the sampling method, requires an operational definition. For example, in the Urban Land Value Survey, if we are interested in obtaining a representative sample of 20 elements, an operational definition might be:

> **the 86 lots in the Central Business District Core were numbered 1 to 86 on the plan, and two-digit numbers were drawn from a random numbers table without replacement; the first 20 valid (between 01 and 86) two-digit numbers drawn were used to locate the 20 sampled lots.**

We note that a random numbers table is a special table of numbers having the property that every item has the same probability of occurrence; "without replacement" means that once an item has been chosen it cannot be chosen again.

A sample has a specified size, usually designated by lowercase n—the number of selected elements. If the population size is known or approximately known, then the relative sample size, usually expressed as a percentage, should also be specified. For the Urban Land Value Survey, a sample of size $n = 20$ would give a relative sample size of 20/86 or a 23% sample.

1.2 VARIABLES AND THEIR MEASUREMENT

A variable may be defined as any characteristic (or attribute) of an element of a population that can be measured in some form. Although an element of a population will obviously have innumerable characteristics, we are usually interested in only one or a few of these. These relevant characteristics—the variables—are the ones that would have been used to assist us in our definition of the population (Section 1.1).

Characteristics of Variables

Each variable should be carefully constructed and the measurement process fully described in an operational definition. For example, from the Urban Land Value Survey, a variable describing the area of the lot might have the following as its operational definition:

> **from the City of Woodford Valuation records held in the City Assessor's Land Tax Office, the lot area (in square meters) was extracted. The City Assessor gave the date of measurement as May 1982.**

In statistical work, we frequently need to refer to variables, and some form of abbreviated description is often required. It is possible to distinguish four levels of description, which in increasing order of abbreviation are as follows:

A full *operational definition,* as used in the foregoing example, which should provide a complete description of the variable and its measurement.

A *descriptive title,* which might be used in the text of an analysis, for example: "Area of Lot."

A short *one-word acronym,* which gives an indication of what is being measured and would be used in computer analyses where it would appear in summary tables and results. For example, the Statistical Package for the Social Sciences (SPSS) computer programs (Section 1.4) provide a maximum of eight alphabetic or numeric characters for variable description at this level, for example: AREA.

A *single symbol* representing a variable is an advantage in computational work. The symbols used include the whole range of alphabetic characters (including Greek letters), but the most common are X and Y. In a particular calculation, each symbol used would be defined using the descriptive title. For example, X where X = Area of Lot.

Using the operational definition of the variable, each element is assigned a *score*. This is the process of *measurement.* For each variable in a population or sample, there will be a set of scores constituting the input data for an analysis, one score for each element. Such a set of scores can be termed a *distribution of scores.* Where some symbol, such as X, can be used to describe the entire set of scores for the variable, subscripts can be used to refer to a particular score for that variable. Thus, X_4 would refer to the score for the variable X obtained from the fourth element. Alphabetic subscripts (especially lowercase i, j, and k) are frequently used to refer to any, or every, score for a variable. Thus, X_i refers to any score for the variable X, and the X_i would mean the whole distribution of scores for the variable X. Distributions can take on a number of forms, but there are two forms in particular.