Chapter 4

Statistical Methods

In common usage, the term *statistics* usually applies to facts and figures. Technically, it refers to a *branch of applied mathematics* concerned with the science, or perhaps the art, of the collection, presentation, description, analysis, inference, significance, testing and prediction of numerical information. Although the variety of such techniques is almost infinite, they may be broadly devided into seven categories—*univariate*, *bivariate*, *multivariate*, *time-series*, *directional*, *network* and *spatial* (Fig. 4.1).

The *univariate* analysis concerns a single variable and allows the distribution of points along the line to be described and statistically tested (data description: frequency distribution, range, average, spread and shape). In *bivariate* analysis, two variables are analysed together to describe and analyse the shape of the scatter for the purpose of investigating the relationship between the data points and/or the relationship between the variables (bivariate correlation coefficient, regression: linear, polynomial, power, logarithmic and exponential). In *time-series* analysis, the sequence of data in time is explored for orders.

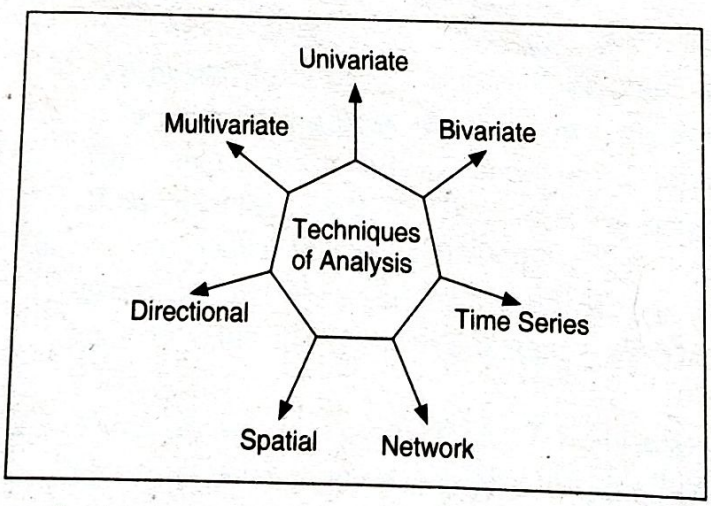

Fig. 4.1 Techniques of Geographical Data Analysis

Directional analysis concerns data expressed in terms of angles or azimuth or bearings from north that can be ordinated on a circle. Such data are of two types—directional and oriented, data that can be analysed with measures of central tendency and randomness. *Network* analysis involves the evaluation of parameters like, structure connectivity, optimisation and shortest path analysis etc.

In *spatial* analysis more than two variables are analysed together, two (or three) of which are spatial coordinates: grid references or latitude/longitudes, with or without altitude or depth. The remaining variable is a measurement of geographical interest and is regarded as varying continuously over the space. The data may be imagined as points in three dimensions and are analysed with the objective of constructing a smooth surface to describe the spatial variation.

The *multivariate* analysis includes the general methods applicable to any number of variables analysed simultaneously and is usually applied to more than three variables. The objective is to reduce the dimensionality so that the shape of the data scatter can be viewed better, relations between and among the variables explored and accordingly properly explained. The methods are multivariate regression, multiple correlation, principal

component analysis, factor analysis, discriminant analysis, cluster analysis, etc.

Statistical data analysis in geography is unique as it concerns the *spatial* data or *geographically referenced* data that are characteristics of multivariate situations. It has two distinct components—*locations* and *attributes at locations*. In the geographer's data matrix (GDM), a *column* represents the variations of attributes of the natural or socio-economic characteristics across some geographical spaces (Fig. 4.2). A *row*, on the other hand, denotes a specific location in geographic space. Therefore, each *cell* formed by the rows and columns of the GDM contains a specific item of geographic fact that can be found at a particular location. In a GDM, the columns can be compared to study the nature of spatial variation of the geographic characteristics. By comparing the rows of a GDM, we study the differences among different places, an aspect known in geography as areal differentiation.

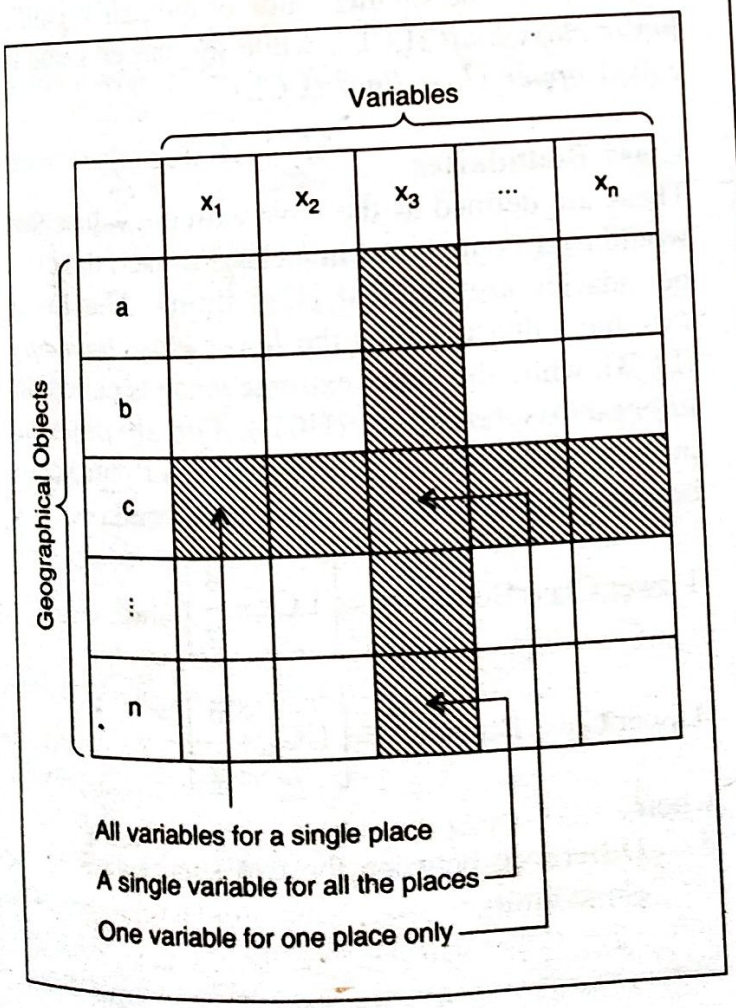

Fig. 4.2 Structure of a GDM

Some Basic Concepts

Data

The term *data* means a body of information in numerical form. A set of data arranged in a tabular form is normally referred to as a data matrix.

Raw Data

It refers to the *unclassified* data from which a frequency distribution may be prepared for further analysis (Table 4.1). The arrangement of raw data is unsystematic and irrespective of any order.

The Array

It is defined by an arrangement of data in which all items are placed *sequentially* in order of magnitude (Table 4.2). The array helps one to see

Table 4.1 Raw Dataset: Population Density of 125 Districts of a Country, 2001 (person/sq km)

840	671	524	671	687
450	355	320	374	526
422	878	486	436	354
671	392	498	610	496
486	492	317	374	486
436	671	486	452	608
826	311	325	767	354
448	614	538	362	782
110	381	328	494	463
662	752	576	392	448
138	333	330	455	392
536	307	780	751	494
162	381	428	288	486
934	576	612	582	642
195	303	292	436	448
458	470	464	881	597
740	264	655	273	461
472	764	256	346	496
486	260	576	587	486
754	456	242	231	880
436	202	765	448	448
562	588	221	581	346
594	217	486	470	594
416	337	470	229	337
982	555	412	464	732

Table 4.2 Array: Population Density of 125 Districts of a Country, 2001 (person/sq km)

110	337	448	494	642
138	337	448	494	655
162	346	448	496	662
195	346	450	496	671
202	354	452	498	671
217	354	455	524	671
221	355	456	526	671
229	362	458	536	687
231	374	461	538	732
242	374	463	555	740
256	381	464	562	751
260	381	464	576	752
264	392	470	576	754
273	392	470	576	764
288	392	470	581	765
292	412	472	582	767
303	416	486	587	780
307	422	486	588	782
311	428	486	594	826
317	436	486	594	840
320	436	486	597	878
325	436	486	608	880
328	436	486	610	881
330	448	486	612	934
333	448	492	614	982

at a glance the range of values, the nature of their concentration and the degree of continuity of the data. It also gives a rough idea of the distribution.

Variables

The items about which information have been collected are usually referred to as *individuals*. In the data matrix, each column gives the value of a specific property or characteristic that varies from one individual to the other. These are called *variables*. The variables that can take fractional values are called *continuous variables*, e.g., height, length, etc. On the other hand, the variables that can take only whole numbers and not fractions are called *discrete* variables. Commonly these concern count data, e.g., population, frequency, etc.

Class

In organising and summarising the statistical data, a frequency distribution is normally prepared. In this, the whole range of data is divided into some groups defined by intervals and the number of observations falling in each group is stated. The groups are called *class intervals* or simply *classes* and the observations are called *frequencies*. The classes are defined by limits or boundaries on two ends. When one end of the class is not specified, it is called an *open-end class*. Normally in a distribution, all classes are of the same size but in some cases, classes of *unequal* sizes may also be taken, particularly for a *highly dispersed* data.

Class Limits

In a grouped frequency distribution, the classes are defined by pairs of numbers such that the upper end of one class does not coincide with the lower end of the following class. These numbers are called class limits which signify the limits of a class for the purpose of tallying with the original distribution. The smaller value of the pair is called *lower class limit* (LCL), while the larger value is called *upper class limit* (UCL).

Class Boundaries

These are defined as the most extreme values that would ever be included in a class. In fact, the class boundaries are the real class limits. The lower extreme value is called the *lower class boundary* (LCB), while the upper extreme value is called the *upper class boundary* (UCB). For all practical purposes of graphical representation of data, class boundaries are used. The working formula is:

$$\text{Lower Class Boundary} = \left[\text{LCL} - \frac{d}{2}\right] \text{ and,}$$

$$\text{Upper Class Boundary} = \left[\text{UCL} + \frac{d}{2}\right]$$

where,

d = Difference between the two successive class limits

Class Width

Width or size of a class is the difference between the lower and upper class boundaries. It is expressed as,

$$\text{Class Width (w)} = (\text{UCB} - \text{LCB})$$

Class Mark

Class mark or mid-value of a class refers to the value that lies exactly at the middle of a class. It is used as the representative value of a class for the purpose of calculation of descriptive measures. The common formula is:

$$\text{Class Mark (x)} = \frac{1}{2}(\text{LCL} + \text{UCL})$$

or,

$$= \frac{1}{2}(\text{LCB} + \text{UCB})$$

Class Frequency

The number of observations contained in a class is known as its *class frequency*. The sum of all class frequencies in a distribution is called the *total frequency*. Classes with zero frequencies are called *empty classes*. Therefore,

$$\text{Total Frequency, } N = \sum_{i=1}^{n} f_i$$

where,

n = Number of class and

f_i = Frequency of the i^{th} class

Relative Frequency

Relative frequency is simply the class frequency (f_i) expressed as a proportion of the total frequency (N) of a distribution. The sum of all relative frequencies in a distribution is equal to unity. In expression form,

$$\text{Relative Frequency, } Rf_i = \left(\frac{f_i}{N}\right) \text{ and } \sum_{i=1}^{n} Rf_i = 1$$

where, n = Number of class.

Frequency Density

Frequency density of a class is defined as its frequency per unit width. It indicates the concentration of frequency in a class. The common expression is:

$$\text{Frequency Density } (f_d) = \frac{\text{frequency (f)}}{\text{class width (w)}}$$

Description

For ordinary as well as for spatial dataset, *description* means the specific and concise measures of their statistical characteristics. Thus, it implies summarising the nature of a large set of data.

Inference

In most cases, analysis is done with data obtained from samples rather than with all the possible data about a particular situation. Hence, it is assumed that the sample is representative of the whole set of data (or the *population*) from which it has been drawn. The *inferential measures* can help, within certain strictly defined limits, to make statements about the characteristics of a population based only on the sample data.

Significance

An observed difference or relationship between two sets of sample data can be tested for significance: *whether there is a difference within the population from which the samples were drawn or whether the observed differences in the samples appear merely due to chance in the sampling procedure*. Statistical measures can be pursued to find the probability that under certain specified conditions, a *relationship is significant* or that the inferences made on the basis of the samples are valid.

Prediction

This means statements about something in the past or in the future from the analysis of a set of period data. Statistical measures are there to do this within certain limits and with certain probabilities. In situations of deterministic processes, predictions and postdictions can be done with absolute precision. But in multivariate situations, the degree of certainty is less and only some sort of probabilistic prediction is possible.