UNIT 19 PREPARATION OF DATA MATRIX

Structure

19.0 OBJECTIVES

This unit aims at familiarising you with some basic quantitative methods that are of used in development planning. Officials seek to determine the efficacy of plans, policies, programmes and projects at every level. This requires obtaining information on the ground. While it is true that in official structures, instruction flows from the top to the bottom, information often flows from the bottom to the top. This unit is concerned with explaining the notions of data, variables and other related concepts, some quantitative methods, besides how data matrices are constructed and used as tools in development planning. After going through the unit, you should be able to:

- Define data and variables and explain data theory;
- Describe various data collection methods;
- Discuss and analyse descriptive and inferential data analysis methods; and
- Discuss the use of data matrices as tools in development planning.

19.1 INTRODUCTION

In previous units of this course, you learnt about the meaning of development, development administration and development planning. You have also learnt that planning is executed at various levels and that there are diverse development issues like agriculture and industry, health and education, and sustainability of development itself. You have had the occasion to know about the role of bureaucracy as well as NGOs in development planning.

You will appreciate that to formulate plans, whether at the national level or at different sectoral levels, as also to implement the plans and the constituent programmes and projects of the plans, information of various kinds would be needed. This information has to be collected, sorted and used. Some of it may even be discarded. This information can be facts regarding various sectors, areas, groups, etc., or it can be about characteristics of the people, profiles of their different aspects, or their problems and attitudes. All this information, which is used, goes to form *data*. So you can see the importance of data for development planning and administration.

This unit is concerned with data. Although the discussion would be general and somewhat abstract, an attempt has been made to provide examples, particularly in the developing nation's context. The unit discusses the meaning and nature of data, and the concept of measurement. It then goes on to deal with the notion of

dimensionality and finally provides a brief discussion of scaling methods, thus, setting the stage for unit 24 on scalogram analysis. The unit will also help in the study of the subsequent unit on trend analysis. While discussing the nature of policies, policy makers and planners often find it useful to obtain information about peoples' faith in government programmes. It becomes necessary to have a theoretical construct of an indicator of a particular attitude, so that this indicator can be usefully measured. We may say that an indicator is to be obtained which is valid and reliable. In the broadest sense, this empirical indicator may be referred to as a scale. In this discussion, we have taken the construction of a scale from a set of attitudinal items that together form an indicator of a particular concept, say trust in bureaucracy.

Administrators often need to evaluate the efficacy of the programmes. Policy makers studying social phenomena are confronted by observations. All observations are not useful for analysis. However, only those observations that are used for analysis comprise the data. Thus, the concept of data is a subjective one, determined in a particular context. Data constitute a subset of observations.

The main purpose of a *theory* of data is to suggest ways by which data can be extracted from observations. As we go along, we shall look at the basic concept related to data, that is, variables. How data are collected and analysed, and a 'spatial' notion of the relation between data and the observations will be discussed. These will lead us directly to the concept of data matrix, which is the crux of the unit.

This unit has to be read closely in conjunction with unit 24 on Scalogram analysis. That unit will deal with measurement concepts and scales, and there is going to be some amount of overlap in these two units.

19.2 DATA AND VARIABLES

We discussed in the previous section about the distinction between data and observations. Now, when we collect data, we do not know in advance what value it will take. For example, when we collect data, say on the incidence of a disease in a particular area, we cannot know the value in advance of this incidence. For us this value is likely to vary, a priori. Of course, on collection it becomes a number. Moreover, incidence of this disease will not be constant across places and over time. Thus, in this case, incidence of disease is a variable. A variable is something whose magnitude varies from case to case. In mathematics and other sciences, something whose value remains the same in all situations is, as the name implies, a constant. So in data collection, we are primarily concerned with variables.

Experimental and Measured Variables

A distinction needs to be made between experimental variables and measured variables. Those variables, which the researcher or the person during the study can control, are called manipulated or active variables. Variables that cannot be manipulated are attribute variable or subject-characteristic variables.

Another way to distinguish between variables, especially in experimental contexts, is to call some variables as stimulus and others response. A stimulus variable is under condition or capable of manipulation by the person carrying out the experiment. The stimulus evokes response in the organisation. Any kind of behaviour of the organisation that is induced by the stimulus is called a response variable. Thus, response variable depends on stimulus variables, and variables

that intrinsically characterise the organisation, on which the stimulus is applied, are called organic variables. Any property, characteristic or attribute of an organism (in social science contexts, an individual) is an organism variable. These are what have been referred to above as attribute variables.

Continuous and Discrete Variables

Variables may be distinguished among each other based on whether they can take all possible values within a certain interval. Those that can are called continuous variable. For example, weight is a continuous variable. On the other hand, a discrete variable is one that can take only certain number of values. A particular type of discrete variables called nominal or categorical variables is such that there is no idea of ranking. Nominal measure means that there is subject of two or more objects from the basic set of objectives being measured. These kinds of variables are assigned for a subclass of a class on the basis of the objects possessing or not possessing the attributes that characterise the object. The word 'nominal' suggests that these variables are 'names' For example, if we categorise into rural and urban, or male and female, these are categorical variables.

Sometimes categorical variables are described as qualitative variables, in contrast to quantitative variables (which are continuous variables). But any nominal variable, if it can have two values, is in a sense, measured and quantitative. Thus, the term 'qualitative' variable is somewhat misleading. Sometimes codes like 'o', 'i', may be given to categorical variables like rural and urban. These are then called dummy variables, observable variables and latent variables.

Sometimes we come across variables that are observed, but underlying these observed variables is some other unobserved or latent variable. For instance, a student may perform very well on verbal and numerical tests, but these high scores (observed variables) on the tests may be due to the presence of an unobserved variable "intelligence". Latent variables are somewhat different from proxy variables, which are variables that are used as 'substitute' for variables, for which data are hard to collect or obtain. Another name for latent variable is intervening variable, but the term latent variable is more common. Very often, a latent variable is an underlying variable that is a variable related to several observed variables.

19.3 DATA THEORY

As we have mentioned earlier, data theory provides abstract models for comprehending the information conveyed by actual observations.

Every empirical observation involves underlying relations and comparison. However, there is a wide variety in the type of observations. Hence, it would be useful if there were one overarching framework, which could analyse empirical observations of various kinds, in an abstract manner.

One useful way to model observations is to take a geometrical metaphor. Suppose a given observation involves a relation or comparison between two entities, and then these entities can be represented as points within a space. The way the analysts choose to interpret the substantial comparison between entities, will dictate the representation of relations position of the two points. So, if we

say that a brinjal is purple, the point depicting brinjal would be close to each other. Or if I say that person K scored more in Mathematics than in Chemistry, the geometric position of chemistry would be more extreme than the position of mathematics in a geometric depiction.

The depiction of data by geometric mean should not be confused with the graphical representation of data like bar charts, pie charts, frequency polygons, etc. There what we are discussing is the basic abstract relation among stimuli, the subjects and the attributes of the subjects. We shall be discussing in this unit - two theories of data: that of Coombs and of Carroll, Arabie and Young. These theories of data led to suggestions regarding construction of data, matrices.

Research entails multiple observations pertaining to the phenomenon under study. Each observation is a separate and distinct relation and comparison between two entities. Hence, it is modelled as a distinct pair of points. Each observation provides a piece of information that summarises relations between entities. This can in turn be modelled as a comparative geometric relation between the members of the point pair. This collection of point pairs goes to constitute the data.

The notion of pair points is broad enough to encompass any kind of observation. Earlier, we had mentioned about stimuli, responses and attributes organisation (that is the subjects). All these different types of variables can be depicted as points along with a reading of measurement (such as a value obtained from a scale).

Coomb's Data Theory

Clyde Coombs put forward theory based on a geometric depiction of data. In his work, he suggested that the two elements in a pair constituting a single stimulus can either be drawn from different sets (such as farmer and agricultural product, a potent and symptoms or a stimulus and response) or they can be drawn from the same set (for example, student x and student y who take the same test, teaching method A and teaching method B). Secondly, he suggested that the two entities are a datum pair, can either have a dominance relationship or a proximity relationship. When one object has less or more of some characteristic as compared to another objective, there is a dominance relationship. When two objects watch or coincide with each other to a considerable extent, there is a proximity relationship. Sometimes the distinction between these types of relationship is clearly apparent from the nature of the empirical observation has with the analyst's interpretation of the observation. Of course, this is true of data in general.

These two types of dichotomies can be represented geometrically. If the elements of the pair of observations are from different sets, the space is called a joint space. If the objects are from the same set, then the space is called a 'stimulus space' or subject space. If the objects in the observation pair are characterised in the ordering of the points in the One-dimensional space, it is called 'an object place'. On the other hand, proximity is depicted as inter point distance. Thus, the two distinct types of dichotomies (observation from some or different sets, and whether the relation between the observation in the pair is one of dominance or proximity) can give rise to four types of data in Coomb's data theory:

1) **Single stimulus data:** These are data where the objects in the observation pair come from different sets and there is a dominance relation. An example is the respondents in a survey, which has items of questions.

Another example is the hotness of a liquid and marks of temperature along a thermometer. Actually, almost all-physical measurement falls within the single stimulus data category;

2) **Stimulus comparison data**: In this case, the pair of elements in an observation is drawn from the same set, and there is a dominance relation between the points in the pair. This usually is the case when similar objects are ranked on the basis of some common property. There is an ordering between two points. Thus, if crop type P has greater yield than crop type S, or one bull has a greater life than another.

   It might appear that single stimulus data and stimulus comparison data portray the same situations. But in the former, one type of object is compared to fundamentally another type of object, although both imply an ordering of a pair of points in a line or space. The information used for the similar geometric representation is different in the two cases;

3) **Similarities data:** Here observations are shown as pairs of entities drawn from the same set, with a proximity relation between them. It is not concerned about the ordering of the points within the space, but with the distance between a pair of points; and

4) **Preferential choice data:** Here the pair of objects in the observations is drawn from different sets, with a proximity relation between them. An example would be of people displaying a preference among different commodities. The people and the commodities are from different sets.

Carroll, Arabie and Young's Theory of Data

It is an alternative data theory, which is related to Coomb's theory of data. Coomb's data theory can be depicted as an abstract 2 by 2 matrix with each cell representing one particular type of data.

The basic idea of this alternative theory is that the number of ways and modes contained in the data matrix is the most important consideration. The dimensions of the data matrix are the ways of the matrix. The number of distinct, different objects represented by the ways of the matrix is given by the modes of the data matrix. Each way of matrix has its own number of levels. There is a minimum of two ways, because an observation always has a comparison between two objects. The type of objects will determine the number of modes. Suppose there is a survey with a set of questionnaire item. If there are p respondents and m items, then there are two ways (respondents and items). The first way has p levels and the second has m levels. In case, there is replication, the replication can be the third way.

Since the type of data matrix corresponds to the analyst's interpretation of the observations rather than the intrinsic nature of the entries involved, the number and type of distinct object may not correspond to the shape and size of the data matrix.

The role of data theory is to clarify the nature of information that is to be used for analysis. Most multivariate statistical methods need data matrix. Scaling methods on the other hand are more concerned with the basic nature of information in the data matrix.

19.4 DATA ANALYSIS AND DATA MATRIX

You have been acquainted with the nature of data and observations. We also discussed two important theories of data. In this section, we are going to look at some simple methods to gather data and to deal with the data thus obtained. Any planning or policy organisation or agency needs to analyse data to come up with improved policy measures. Hence, it is important for policy agencies to collect data, analyse it and construct data matrices. In fact, as you shall soon see, the important thing is the *analysis* of the data matrix. Making of the matrix is not so hard. When we deal with several variables together, which is what comes naturally in policy analysis, we end up analysing these several sets of data for the several variables together. This analysis is carried out using what are called multivariate analysis. For multivariate analysis, constructing data matrices is merely the first step. A multivariate analysis technique much used in data matrix analysis is called 'Principal Component Analysis'. This technique quantifies and arranges data presented in a matrix, to help find more general indicators that would clarify more complex information.

In this section, we are going to discuss some very simple ways to get data and to analyse these data. At the end of the unit there are some books mentioned, for example, the one by Welch and Comer, which you could go through for the details and fine points. Now, let us get on to a discussion about analysing data matrices.

Both experimental as well as non-experimental methods can collect data. In the former, the data collector often undertakes some action or experiment, which would give rise to responses from the members of a group. In experimental methods, there are often two groups, an experimental group, the members of which have the experiments conducted on them, and the control group whose members do not. In non-experimental methods, which are more prevalent in the social sciences, data are often collected from a sample of the total population, unless, of course, it is a census, when there is an enumeration of the entire population. In sampling, data are obtained through interviews or questionnaires. To draw up samples, it is important that each population have an equal probability of being in the sample. This is called random sampling. There are subcategories within random sampling, such as simple random sampling, stratified random sampling and so on. There are non-random sampling methods as well, such as purposive sampling. But it must be kept in mind that the results of inferences drawn based on random sampling will always be more robust than those inferences, which are drawn, based on non-random sampling methods.

After the data have been collected, these data are often presented using some pictorial representation, such as pie charts or frequency polygons. Descriptive methods are used to glean some information from the presented data. Measures of central tendency like the mean or median (“the middlemost”) or the mode are used, as also measures that show spread or dispersion of the data around the specified measure of central tendency, usually the mean. These measures of dispersion include the variance and its positive square root, the standard deviation. Other descriptive measures are those, which show the ‘flatness’ or ‘peaked ness’ of the frequency polygon. This measure is called kurtosis. Another set of descriptive measure has to do whether the frequency distribution curve is disproportionate to the left or right rather than evenly spread in both directions. In this case, the mode of the distribution would lie towards the low-valued or high-valued end of the observations. This measure is called skew ness.

Other than descriptive measures, one important use of the sample data is to say something of interest about the entire population. In other words, we infer something about the whole population based on information about the sample. This is called inferential statistics. This has two aspects. One aspect is to estimate the value of some population parameter based on some sample measure. This sample measure is called a statistic. For example, we can estimate the population mean on the basis of the sample mean. In this case, the sample mean would be a statistics. The other aspect of inferential statistics is what is called hypothesis testing. For this we form a hypothesis about some aspect of the population. This is called a null or a maintained hypothesis. At the same time, we also mention what is called an alternative, which is in a sense contrary to the null hypothesis. In inferential statistics, methods from probability theory, namely, the probability distribution of the sample statistics, called the sampling distribution, are used to determine whether we can reject the null hypothesis or not. We use several different types of tests of hypotheses to determine the validity of rejection or acceptance of hypotheses.

When the question is to determine relationships among two or more variables together, then techniques of regression and correlation are used. Regression is used more to determine causal relation among variables, that is, to validate if a variable Y depends on, or is caused by, variables X1, X2, X3 and so on. Correlation is used to determine if the values of different variables move together.

When we consider more than two variables together we are using multivariate statistics. There are several types of multivariate statistical methods like principal components method, factor analysis, analysis of variance and analysis of covariance. All of these require the use of data matrices.

An important operation in metric scaling is the decomposition of a data matrix into its basic structure. This process by which a matrix is decomposed into its basic structure is called singular value decomposition.

All metric scaling methods have three phases:

1. The original data are transformed by some normalisation on transformation procedure;
2. A singular value decomposition of the transformed data summarise the basic structure in the data with a set of new vectors, a set of column sectors and a set of singular values; and
3. The sets of row and column reactors may be rescaled or otherwise weighed to provide a final set of row and column sectors.

The second step is the most crucial in metric scaling procedure.

There are various types of metric scaling procedures, that is, principal component analyse, multidimensional preference scaling and correspondence analysis of contingency tables.

The main use of the theory of data is to help in the construction of scales about which you will learn in unit 23. Here, we may simply say that ranking objects, with or without mentioning the difference between the various ranks or their weights, is called a scale. One particular type of scaling is metric scaling. The basic idea behind metric scaling is the representation of relations between 2 sets of variables. The aim is to summarise and reveal mutual relationships,

interrelationships among variables of different kind, without necessarily focusing on dependent or independent variables. One begins with data sets, and then inputs the raw data visually to form some notion of the possibilities it contains. Then we perform some exploratory data analysis on the data set.

19.5 DATA MATRIX IN DEVELOPMENT PLANNING

We have given some idea of the theoretical conceptualisation of data, data theory, data analysis, and some rudimentary notions of metric scaling. However, the main idea behind data theory is to construct data matrices so as to perform some multivariate analysis. This would go some way towards helping to formulate better policies.

Data matrices have been used in decentralised, multi-level planning, particularly in rural development, about themes and topics that have a spatial and geographical orientation, along with trend analysis and scalogram, about which you will learn in other units.

Imagine policy planning at the district or block level. There would be a welter of data about a lot of variables. There would be, for a village, for instance, data on socio-economic and demographic profile of the people, data on cattle, livestock, data on agriculture, etc. Now, a lot of these variables would affect each other. To draw meaningful relationships and interdependencies among various variables, data matrices are constructed and analysed. The analysis of such data would help to improve policy-making. District administration can use data matrices to draw up some analysis of data and transmit the analysis at higher levels for macro planning purposes.

19.6 ACTIVITY

1. Distinguish between observations and data. Suppose you were working in government administration in the area of rural development, mention five variables of interest to you, whose data you would collect. What method would you use?
2. Briefly compare Coomb's data theory with that of Carroll, Arabie, and Young.
3. Describe some uses of data tables and data matrices. How is it related to metric scaling?

19.7 CONCLUSION

This unit aimed at familiarising you with some basic concepts about data, its nature, its analysis and use of some multivariate methods to analyse the data. The unit began by making a distinction between observation and data. It then went on to discuss two very important theories and representation schemes of data, namely Coomb's theory and the Carroll-Arabie-Young theory of data. It stressed the importance of conceptualising about the data in a geometric manner and alternatively in form of a matrix.

Next, the unit proceeded to mention, in a brief manner, the various methods of data collection, diverse methods of describing that data, as well as drawing inferences about population parameters on the basis of sample data. It further briefly elucidated the relation between data matrices and metric scaling. The unit

talked a little of multivariate methods, for example, principal components analysis, and also described how simple data matrices are constructed and used as development planning techniques.

19.8 REFERENCES AND FURTHER READINGS

Blalock, H.M., *Conceptualisation and Measurement in the Social Sciences*, Sage Publications, California, 1982.

McKay, David, Norman, Schofield, and Paul, Whiteley, eds., *Data Analysis and the Social Sciences*, Frances Pinter, London, 1983.

Welch, Susan and John, Comer, *Quantitative Methods for Public Administration* (second edition), The Dorsey Press, Chicago, 1988.

Weller, Susan and Romney, A Kimball, *Metric Scaling: Correspondence Analysis*, Sage Publications, London, 1990.