UNIT 8 : PROCESSING OF DATA

UNIT STRUCTURE

8.1 Learning Objectives
8.2 Introduction
8.3 Data and Data Types
8.4 Data Measurement Techniques
8.5 Editing of data
8.6 Classification of data
8.7 Tabulation of data
8.8 Let us Sum up
8.9 Answers to Check Your Progress
8.10 Further Reading
8.11 Possible Questions

8.1 LEARNING OBJECTIVES

After reading this chapter you will learn :

- definition of data and various types of data
- various measurement issues related to data and data types
- all about data editing
- how to classify the data
- how to tabulate the data in uni-variate and multi-variate distributions

8.2 INTRODUCTION

The basic aim of any research is to make generalisations and build theory upon them. Note that all scientific methods of research require careful observations at some level or other. Speaking plainly, observation is the act of acquiring knowledge and facts through our five sense organs. That is why observation is also known as sensory perceptions. The act of gathering sensory perception is known as empirical research.

Clearly, therefore, in empirical research we try to observe the attributes and whatever is obtained as a result of observation is called data. We will take up the issue of data in some detail a little later. The data help us in validating our research hypothesis as to whether we can accept or reject

The most familiar numbers are the natural numbers or counting numbers: 1, 2, 3 and so on. Traditionally, the sequence of natural numbers started with 1 (0 was not even considered a number by the Ancient Greeks.) However, in the 19th century mathematicians started including 0 within natural number system. In the decimal system (i.e. 10 base system), the symbols for natural numbers are written using ten digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. In this base ten system, the rightmost digit of a natural number has a place value of one, and every other digit has a place value ten times that of the place value of the digit to its right.

The negative of a positive number is defined as a number that produces zero when it is added to the corresponding positive number. Negative numbers are usually written with a negative sign (a minus sign). As an example, the negative of 7 is written ?7, so that 7 + (?7) = 0. When the set of negative numbers is combined with the set of natural numbers (which includes zero), the result is defined as the set of integer numbers, also called integers.

On the other hand, a rational number is one that can be expressed as a fraction or ratio with an integer numerator and a non-zero natural number denominator. Fractions are written as two numbers, the numerator and the denominator, with a dividing bar between them e.g. .

The real numbers include all of the measuring numbers. Real numbers are usually written using decimal numerals, in which a decimal point is placed to the right of the digit with place value one. Every rational number is also a real number. It is not the case, however, that every real number is rational. If a real number cannot be written as a fraction of two integers, it is called irrational number. For example, the real number ? (pi), or 22/7 neither ends nor forever repeats. It, therefore, cannot be written as a fraction, and is an example of an irrational number.

Moving to a greater level of abstraction, the real numbers can be extended to the complex numbers. Discussion on complex number is, however, beyond the scope of this lesson.

After the various types of numbers, let us briefly discuss the types of data in general. As has been mentioned earlier data can be quantitative or qualitative. Depending on unit of collection, data can also be classified as cross sectional, time series or panel. The data is called cross sectional

8.4 DATA MEASUREMENT TECHNIQUES

Let us begin by understanding what is meant by a measurement scale. For a moment, consider that you are standing in a room somewhere and you want to locate your position. Now, your location can be measured as a distance from any of the four walls of the room. It is possible that depending on the wall you choose, your distance can vary. The point here is that for measuring a distance one needs to refer to a particular point from where the measurement is actually to be taken. Mathematically, this point of reference is known as the origin of the measurement.

Next important issue involved here is the unit of measurement. You know that distance can be measured in different units - metre, centimetre, millimetre etc. Note that these units of measurement are nothing but intervals that divide the total scale into number of constituent parts. A careful examination will reveal that these units are, in fact, indicative of number of intervals created within the scale - 'centi' is indicative of hundred, 'mili' is indicative of thousand and so on. Important point here is to note that these intervals can either be equal or unequal. Very common example of an equal interval scale is decimal scale and that of an unequal scale is logarithmic scale. In the case of the first, each interval is equal to the preceding and succeeding intervals. In the case of the second, each interval succeeds by a larger interval. While dealing with quantitative data this equality or inequality of interval becomes very crucial because some arithmetic operations like addition and subtraction are meaningful only when intervals are equal.

It is, therefore, evident that a measurement scale essentially relates to two important notions - origin and interval. Depending on the nature of origin and interval we can have different type of scales, and following that different tools of numerical analysis. Let us spend some time in briefly discussing different type of scales.

Depending on the type of origin and interval used in measurement scales there can be four types of scales. First and the weakest of all is the nominal or categorical scale. In social sciences, it is quite common to denote or nominate different categories or options with numbers. For instance, categories like 'male' and 'female' are often represented by say 1 and 2 respectively. Similarly options like 'yes' and 'no' can be represented by 1

that we propose to analyse for not all methods and tools are permissible for all the types of quantities.

8.5 EDITING OF DATA

After collection of data, before the researcher analyses them it is essential to edit the data so that reliable and meaningful results are obtained. Note however, that editing is not manipulating. Editing is done for specific purposes without changing the information contained in data. Editing is carried out to in the first place to check whether there is any invalid code or value in the data. Invalid code or value is one which is not defined for the study. For example if male and female are coded as 1 and 2 value of 3 is not defined. While editing the researchers has to make sure that all data are in the values of 1 and 2. Second, while editing consistency amongst data are to be checked. For instance, while asking about reproductive health related questions it is expected that respondent should be married female. Therefore, such responses are to be checked thoroughly whether they have been provided by married females or not. Third, editing also involves classification of data. We will discuss classification below in detail.

8.6 CLASSIFICATION OF DATA

Sometimes, we may need to classify the frequency for making sense. Consider the following ages of 40 people:

12	23	21	28	31	46	49	51	69	72
84	16	77	43	37	33	26	19	54	42
11	06	17	33	45	30	40	55	73	38
14	22	26	33	36	76	80	67	62	59

Note the mere frequency will not summarise the above data effectively, for values are not repeated much. One therefore, can devise a classification scheme so that the data can be summarised properly. Let us make the age-classes as "up to 18", "18 to 30", "30 to 45", "45 to 60" and "more than 60". For classification tally marks are used for each class. The classified frequency then would be more convenient for summarising the distribution as shown below:

In short, a variable comprises of a set of responses and quantitative variables, in particular, comprises of numbers measured in terms of a particular scale. Let us suppose that one is interested in the study of the attribute age and he records ages of 10 persons as given below:

Age : 12, 23, 32, 18, 25, 29, 36, 41, 37, 26

It is evident from above that the variable age is given by a set of responses or values. The set of values (or responses) of a given variable, thus obtained, is called the distribution of the variable. In case, we are interested in only one variable, the distribution is called univariate distribution. When two variables are considered together, the two distributions together are known as bivariate distribution. When number of variable is more than two, it is called multivariate distribution.

The values of a variable can appear only once in a distribution or several times repeatedly. The number of times a value appears in a distribution is called its frequency. In other words, frequency of a value indicates how frequently the value appears in the distribution. When values are arranged along with their respective number of appearance or frequency the distribution is called frequency distribution. For example, consider the following values of a distribution

2, 2, 3, 3, 3, 1, 4, 2, 3, 4, 2, 1, 1, 3, 2, 3, 3, 2, 1, 1, 1, 2, 3, 2

The frequency distribution of the distribution is given as below

Value (X) :	1	2	3	4
Frequency (f) :	6	8	8	2

Speaking mathematically, frequency distribution f represents the number of times the value x is repeated in a particular distribution or series of a variable.

After collecting relevant data, researchers need to present them in an effective way to put across the basic arguments. Frequency tables are the first step towards this. A frequency table describes the data in a condensed way and shows how the values are repeated in a series. For instance, suppose one survey 15 people about their gender and data obtained are given as below

Age classes	**Sex**		**Total**
	Male	**Female**	
up to 18	4	3	7
18 to 30	2	4	6
30 to 45	10	1	11
45 to 60	5	2	7
more than 60	4	5	9
Total	**25**	**15**	**40**

It may be noted that in cross-table we have three different totals: total over rows, total over columns and total over the table. Accordingly we will have three different percentages for each cell value. One has to be careful about interpreting the results. Let us consider the value of male in the age group 30 to 45 viz. 10. Now one can calculate column percentage for the value as = 40 percent. Similarly row percentage for the value can be calculated as =91 percent. Also, total percentage for the value can be obtained as =25 percent. It is important to understand that all three percentages mean different things. The first case implies that out of total male 40 percent fall in the age group of 30- 45. The second means out of total sample in the age group 30-45 some 91 percent are male. While the third case indicates out of the total sample 25 percent are male and belong to age group 30-45.

It is thus, clear that for single variable distribution frequency table and for two variable distribution cross tables can be used. However, in most of the cases we have to deal with more than two variables. How do we accommodate a third variable in a table?

A third variable can be accommodated in a cross table by using two concepts - called nesting and layer. Let us accommodate another variable say place of residence (viz. rural and/or urban) in the age-sex cross of above to understand these two concepts. When the variable is accommodated within column it is called nesting as shown below:



CHECK YOUR PROGRESS

Q.4 : What is interval scale?

.....

.....

Q.5 : What are discrete variables ?

.....

.....

Q.6 : Which is the weakest scale?

.....

.....

Q.7 : Where are continuous variables generally applied?

.....

.....



8.8 LET US SUM UP

- In this chapter we have discussed the definition of data and also various types of data. We have discussed primary and secondary data, quantitative and qualitative data, cross-section and time series data. We have also discussed various measurement issues related to data and data types. We have introduced the idea of scale of measurement and elaborated nominal, ordinal, interval and ratio scales and their utilities. We have also briefly discussed about data editing and its requirement. Then we have gone onto describe classification of data. Next we have discussed about single variable frequency tables and two-variable cross tables. Finally we have introduced the idea of nesting and layer to accommodate more than two variables in a table.



8.10 FURTHER READING

- "Fundamental of Statistics"; S C Gupta, S Chand and Company, New Delhi
- "Fundamental of Statistics"; A M Gun, M K Gupta, B Dasgupta, Vol. I, World Press
- "Basic Statistics"; A L Nagar, Oxford University Press
- "Basic Statistics"; B L Agarwal, New Age International
- "Statistics without Tears", D Rowntree, Penguin Books



8.11 POSSIBLE QUESTIONS

Q.1 : What is data? What are the different types of data?

Q.2 : What do your understand by a measurement scale? What are the basic elements of a scale? Define different measurement scales.

Q.3 : Distinguish between primary and secondary data.

Q.4 : What is data editing? Why is data editing necessary?

Q.5 : What is classification of data? How a discrete class can be made continuous?

Q.6 : What is a cross table? How can one accommodate a third variable in a cross-table?

****