UNIT 5 PROCESSING AND ANALYSIS OF DATA

Contents

5.0 OBJECTIVES

On completion of this Unit, you would be able to:

- edit your data;
- prepare the code book for your Interview schedule/questionnaire;
- prepare your Master Chart;
- plan your data analysis;
- classify your data;
- tabulate your data;
- analyse your data;
- know the application of computer for data processing and analysis.

5.1 INTRODUCTION

In the previous Unit we discussed about methods and tools of data collection. After data collection the researcher turns his focus of attention on its processing. In this Unit, we will discuss about one of the most important stages of the research process, i.e. data processing and analysis.

5.2 DATA PROCESSING

Data processing refers to certain operations such as editing, coding, computing of the scores, preparation of master charts, etc. A researcher has to make his plan for each and every stage of the research process. As such, a good researcher makes a perfect plan of processing and analysis of data. To some researchers data processing and analysis is not a very serious activity. They feel many times that data processing is a job of computer assistants. As a consequence, they have to be contended with the results given by computer

assistants which may not help them to achieve their objectives. To avoid such situations, it is essential that data processing must be planned in advance and instructed to assistants accordingly.

5.3 EDITING OF DATA

After collection of filled in questionnaires, editing of entries therein is not only necessary but also useful in making subsequent steps simpler. Many a times, a researcher or the assistants either miss entries in the questionnaires or enter responses, which are not legible. This sort of discrepancies can be resolved by editing the schedule meticulously. Another problem comes up at the time of tabulation of data when researcher asks for tabulation of responses from consecutive questions. In cases where data are not cleaned there has to be inconsistency in the tabulations. The researcher has to be very particular about consecutive questions where category 'not applicable' exists.

Check Your Progress I

Note: a) Use the space given below for your answer.
b) Check your answer with those given at the end of this unit.

1) Distinguish 'data processing' from 'data analysis'.

.....

.....

.....

.....

.....

.....

5.4 CODING OF DATA

Coding of data involves assigning of numbers to each response of the question. The purpose of giving numbers is to translate raw data into numerical data, which may be counted and tabulated. The task of researcher is to give numbers to response carefully. As we have already discussed various types of questions (such as open-end, close-end, matrix, factual, opinion, etc.) in the previous unit, the coding scheme will vary accordingly. For example a close-end question may be already coded and hence it has to be just included in the code book whereas coding of open-end questions involves operations such as classification of major responses and developing a response category of 'others' for responses which were not given frequently. The classification of responses is primarily based on similarities or differences among the responses. Usually, in the case of open-end questions, to classify responses researcher looks for major characteristics of the responses and puts it accordingly. In case of attitude scales, researcher has to keep in mind, the direction or weightage of responses. For example, a response 'strongly agree' is coded as 'five' the subsequent codes would be in order. Therefore, if there are responses like 'agree' 'undecided' 'disagree' and 'strongly disagree' they have to be coded as four, three, two, one. Alternatively, if strongly agree is coded as minus two, the subsequent responses would be coded as minus one, zero, +1 and +4. The matrix questions have to be coded taking in to consideration each cell as one variable. For example, if the column of matrix represents employment status, namely, 'permanent'

and 'temporary' and row represents employers or type of employer, namely, government and private, the first cell would represent a variable 'government-permanent'. The second cell would represent 'government-temporary' and so on. In order to demonstrate the points discussed above a section of the code-book is reproduced in Table 5.1.

Table 5.1: Code Book

Q.No	Var. No.	Information Sought	Responses	Code	Column No.	Remarks
		Respondents Number			1-3	
2	1	Age	Actual	-	4-5	
3	2	Designation	Worker Supervisor Manager	1 2 3	6	
4	3	Establishment	Public Private	1 2	7	
5	4	Level of Education	Graduate Intermediate High School Middle School Primary Illiterate Other	1 2 3 4 5 6 7	8	
6	5	Marital Status	Married Unmarried Widow Divorce	1 2 3 4	9	
36	35	Nature of work	Yes No	1 2	10	
	36	Duration of work	Yes No	1 2	11	
	37	Wages	Yes No	1 2	12	
	38	Promotion	Yes No	1 2	13	
43	42	Attitude of Employer Preference for male employees	Agree Undecided Disagree	1 2 3	14	

5.5 PREPARING A MASTER CHART

After a code book is prepared, the data can be transferred either to a master chart or directly to computer through a statistical package. Going through master chart to computer is much more advantageous than entering data directly to computers because one can check the wrong entries in the computer by comparing 'data listing' as a computer output and master chart. Entering data directly to computer is disadvantageous, as there is no way to check wrong entries, which will show inconsistencies in tabulated data at the later stages of tabulation. A sample of master chart prepared in accordance with the code book is presented in Table 5.2.

Table 5.2: Master Chart

RESPONDENT NUMBER				VARIABLE LABELS									
				AGE	DESIGNATION	ESTABLISHMENT	LEVEL OF EDUCATION	MARITAL STATUS	NATURE OF WORK	DURATION OF WORK	WAGES	PROMOTIONS	ATTITUDE OF EMPLOYER
				QUESTION? VARIABLE/COLUMN NUMBER									
1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	1	4	1	2	1	2	3	2	3	4	3
0	0	2	2	1	2	3	4	5	6	7	8	9	5
0	0	3	3	3	3	3	4	2	6	7	2	6	
0	0	4	4	5	7	8	9	1	3	5	6	1	1
0	0	5	5	4	5	1	4	2	1	1	4	3	
0	0	6	6	3	1	2	3	4	5	6	3	1	5
0	0	7	7	5	4	5	6	9	7	8	5	2	4
0	0	8	8	1	4	2	5	3	6	3	7	8	9
0	0	9	9	2	2	4	2	6	7	8	2	1	5
0	0	0	0	9	5	6	8	7	9	2	4	4	3
0	0	1	1	2	2	8	4	9	3	4	7	3	8
0	0	2	2	2	5	7	9	5	1	4	2	4	3
0	0	3	3	2	9	4	5	6	7	2	6	9	6
0	0	4	4	2	8	7	9	5	2	4	6	2	3
0	0	5	5	3	5	4	8	7	9	2	4	2	3
0	0	6	6	2	4	8	7	9	5	8	4	5	6
0	0	7	7	8	7	4	9	4	3	4	6	3	4

Check Your Progress II

1) What is "Coding of Data".

.....

.....

.....

.....

.....

.....

.....

2) Highlight the importance of preparing of 'Master Chart".

.....

.....

.....

.....

.....

.....

5.6 TABULATION OF DATA

Tabulation is a process of presenting data in a compact form in such a way as to facilitate comparisons and show the involved relations. In other words, it is an arrangement of data in rows and columns. This also helps the researcher to perform statistical operation on the data to draw inferences. Tabulation can be generally in the form of uni-variate, bi-variate, tri or multi-variate tables. Accordingly, analysis proceeds in the form of uni-variate analysis, bi-variate analysis and tri or multi-variate analysis.

5.7 CLASSIFICATION OF DATA

The process of classification in accordance with the objectives and hypothesis of the study is arrived at with the help of frequency distributions. Re-classification is a process to rearrange responses with the help of statistical techniques. This helps researcher to justify the tabulation. We have seen earlier that the responses to a statement may be assigned scores or weightage. These scores or weightage are summated and re-classified may be as 'high,' 'medium' and 'low'. The basic principle in the process of classification or re-classification is that the categories thus obtained must be exhaustive and mutually exclusive. In other words, the categories have to be independent and not overlapping.

5.8 DATA ANALYSIS AND INTERPRETATION

The purpose of data analysis is to prepare data as a model where relationships between the variables can be studied. Analysis of data is made with reference to the objectives of the study and research questions if any. It is designed to test the hypothesis. It also involves re-classification of variables, tabulation, explanation and casual inferences.

The first step in data analysis is a critical examination of the processed data in the form of frequency distribution and cross tabulation. This analysis is made with a view to draw meaningful inferences and generalisation.

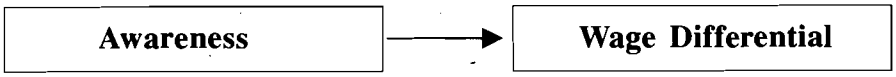
Setting up the Analytic Model

Before we begin the analysis of data we have to look back to the objectives of our research study and set up analytic models. These models are diagrammatic presentation of variables and their interrelationships.

Let us hypothesise that awareness about the Equal Remuneration Act affects wage differentials.

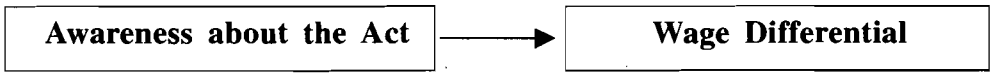
H1 ⟶ Awareness about the Act Affects Wage Differnetials

The two variables in the hypothesis are awareness about the Act and the wage differentials. The relation between the two variables can be diagrammatically presented as follows:



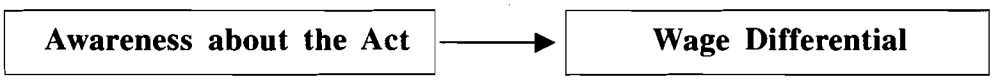
Further it is hypothesised that the bivariate relationships between the variables are affected by another variable regional development. Let us suppose that regional development has been categorised into three, namely, high, medium and low. This can be described as follows:

REGIONAL DEVELOPMENT: HIGH



REGIONAL DEVELOPMENT: MEDIUM

REGIONAL DEVELOPMENT: LOW



With the analytic model described above the researcher can proceed to analyse the data as discussed in the following sections:

Univariate Analysis

Univariate analysis refers to tables, which give data relating to one variable. Uni-variate tables which are more commonly known as frequency distribution tables show how frequently an item repeats. Examples of frequency tables are given below. The distribution may be symmetrical or asymmetrical. The characteristics of the sample while examining the percentages, further properties of a distribution can be found out by various measures of central tendencies. However, researcher is required to decide which is most suited for this analysis. To know how much is the variation, the researcher has to calculate measures of dispersion.

Usually, frequency distribution tables are prepared to examine each of the independent and dependent variables. Tables 5.3, 5.4 and 5.5 present two independent variables and one dependent variable.

Table 5.3: Showing Awareness of the Respondents

(The Independent Variable)

Level of Awareness	Distribution of Respondents	
	Freq.	Percentage
High	110	39.3
Medium	106	37.9
Low	64	21.8
Total	280	100.0

Table 5.4: Showing the Respondents by Regional Development

(The Independent Variable)

Regional Development	Distribution of Respondents	
	Freq.	Percentage
High	142	57.7
Medium	86	30.7
Low	52	14.6
Total	280	100.0

Table 5.5: Showing Wage Differentials of the Respondents

(The Dependent Variable)

Wage Differentials	Distribution of Respondents	
	Freq.	Percentage
High	78	27.9
Medium	134	47.9
Low	68	24.2
Total	280	100.0

A frequency distribution of a single variable is the frequency of observation in each category of a variable. For example, an examination of the pattern of response to variable 'awareness of the respondents' in Table 5.3 would provide a description of the number of respondents who have high, medium and low level of awareness. In case of nominal variables categories can be listed in any arbitrary order. Thus, the variable "Religion" may be described with the category 'Hindu' or the category 'Christian' listed first. However, the categories of ordinal, interval and ratio variables are arranged in order. Let us consider the frequency distribution (Tables 5.3, 5.4 and 5.5) which describes the awareness, wage differentials and regional development of

respondents. The tables have four rows, the first three being the categories of variables, which appear in the left-hand columns and the right hand columns show the number of observation in each category. The last rows are the totals of all frequencies appearing in tables. To analyse the data it is necessary to convert the frequencies into figures that can be interpreted meaningfully. Frequencies expressed in comparable numbers are called proportions or percentages. A proportion is obtained by dividing the frequency of a category by the total number of responses in the distribution. Proportions when multiplied by 100 become a percentage. For example, the relative weight of the category 'High' in Table 5.3 is expressed by the proportion 110/280=.393 or by the percentage 110/280 x 100=39.3 per cent. These figures indicate that only about 40 out of every 100 respondents in the group have 'high' level of awareness about the Act. Proportions and percentages permit the comparison of two or more frequency distributions, for instance, while distribution of respondents by regional development displayed in Table 5.4 clearly shows the predominance of respondents from 'high' development region whereas, distribution of respondents by wage differential in Table 5.5 indicates that the proportions of respondents with 'high' and 'low' wage differentials are almost equal.

Bivariate Analysis

A researcher might be interested in knowing the relationships between the variables. To know the relationship between these variables, the data pertaining to the variables are cross tabulated. Hence, a bi-variate table is also known as cross table. A bi-variate table presents data of two variables in column and row simultaneously. An example of a bi-variate table is given below:

Table 5.6: Showing Levels of Awareness Towards the Act and Wage Differentials of the Respondents

Awareness about the Act	Wage Differentials			Total
	High	Medium	Low	
High	94 (66.2)	9 (10.5)	7 (13.5)	110 (39.3)
Medium	37 (26.1)	58 (67.4)	11 (21.2)	106 (37,9)
Low	11 (7.7)	19 (22.1)	34 (65.3)	64 (22.8)
Total	142 (57.70)	86 (30.7)	52 (14.6)	280 (100.0)

The table presents data with regard to two variables namely awareness about the Equal Remuneration Act and the level of wage differential. First row presents data with regard to respondents who were aware of the Act. The second row presents data about who were not. Similarly, the first column gives data pertaining to workers who have low wage differentials. The second column presents data of workers whose differentials were medium and the last column represents the respondents who felt high wage differentials. For example, the first cell (in the left-hand corner) represents 94 respondents who were fully aware of the Act and perceived high wage differentials.

The association between two variables can be explained either by comparing the percentages of respondents column wise or row wise. The relationship between

the variables can also be examined by various statistical techniques depending upon the level of measurement of the data. Apparently, the two variables are associated, therefore, more people who were aware of the Act have perceived low wage differentials than who were not aware of the Act. Alternatively, comparatively smaller percentage of people who were aware of the Act has perceived high wage differentials than people who were not aware of the Act.

In bivariate analysis the researcher also explains the nature and the degree of association. That is whether the relationship is positive or negative and it also indicates the degree of relationships in terms of high, moderate or low.

Trivariate Analysis

Sometimes researcher might be interested in knowing whether there is a third variable which is effecting the relationships between two variables. In such cases the researcher has to examine the bi-variate relationship by controlling the effects of third variable. One way of controlling the effects of a third variable is to prepare partial tables and examine the bi-variate relationship. Let us take an example. In the above table, if researcher wants to examine whether there is effect of regional development on the bivariate relationship he may prepare three partial tables giving data relating to awareness of the Act and wage differential for high, medium and low regional development.

Table 5.7: Regional Development = High (N = 142)

Awareness about the Act	Wage Differentials			Total
	High	Medium	Low	
High	7 (21.9)	10 (21.8)	3 (9.3)	20
Medium	13 (40.6)	26 (33.3)	9 (28.1)	48
Low	12 (37.5)	42 (53.8)	20 (62.5)	74
Total	32	78	32	142

Table 5.8: Regional Development = Medium (N = 86)

Awareness about the Act	Wage Differentials			Total
	High	Medium	Low	
High	8 (36.4)	11 (25.6)	4 (19.0)	23
Medium	9 (40.9)	17 (39.5)	8 (38.1)	34
Low	5 (34.9)	15 (34.9)	9 (42.9)	29
Total	22	43	21	86

Table 5.9: Regional Development = Low (N = 52)

Awareness about the Act	Wage Differentials			Total
	High	Medium	Low	
High	14 (58.3)	7 (53.8)	7 (46.7)	28
Medium	8 (33.3)	4 (30.8)	6 (40.0)	18
Low	2 (8.3)	2 (15.4)	2 (13.3)	6
Total	24	13	15	52

On examination of these three partial tables, if the researcher finds out that bi-variate relationships do not hold good he may infer that it is the third variable, the regional development which is affecting the bi-variate relationship. In the partial tables for higher regional development, the proportion of people perceiving high wage differential are those who are having high level of awareness about the Act. The similar trend can be noticed in the remaining two partial tables, which means regional development does not effect the bi-variate relationships between wage differential and awareness about the Act.

5.9 USE OF COMPUTER IN DATA PROCESSING AND TABULATION

Research involves large amounts of data, which can be handled manually or by computers. Computers provide the best alternative for more than one reason. Besides its capacity to process large amounts of data, it also analyses data with the help of a number of statistical procedures. Computers carry out processing and analysis of data flawlessly and with a very high speed. The statistical analysis that took months earlier takes now a few seconds or few minutes. Today, availability of statistical software and access to computers has increased substantially over the last few years all over the world.

While there are many specialised software application packages for different types of data analysis, Statistical Package for Social Sciences (SPSS) is one such package that is often used by researchers for data processing and analysis. It is preferred choice for social work research analysis due to its easy to use interface and comprehensive range of data manipulation and analytical tools.

Basic Steps in Data Processing and Analysis

There are four basic steps involved in data processing and analysis using SPSS. They are:

1) Entering of data into SPSS,
2) Selection of a procedure from the Menus,
3) Selection of variables for analysis, and
4) Examination of the outputs.

You can enter your data directly into SPSS Data Editor. Before data analysis, it is advised that you should have a detailed plan of analysis so that you are

clear as to what analysis is to be performed. Select the procedure to work on the data. All the variables are listed each time a dialog box is opened. Select variables on which you wish to apply a statistical procedure. After completing the selection, execute the SPSS command. Most of the commands are directly executed by clicking 'O.K'. on the dialog box. The processor in the computer will execute the procedures and display the results on the monitor as 'output file'.

Check Your Progress III

1) What is an analytic model?

.....

.....

.....

.....

.....

2) What is 'bivariate' and 'trivariate' analysis?

.....

.....

.....

.....

.....

5.11 KEY WORDS

Univariate Analysis : refers to tables which give data relating to one variable.

Bivariate Analysis : refers to relationship between two variables.

Trivariate Analysis : controlling the effects of third variable which is effecting the relationship between two variables.

5.12 SUGGESTED READINGS

Bailey, Kenneth, D. (1978), *Methods of Social Research*, The Free Press, London.

Baker, L. Therese (1988), *Doing Social Research*, McGraw Hill, New York.

Doby, J.T. et al. (1953), *An Introduction to Social Research*, Harrisburg: Stackpole Co., London.

Kerlinger, Fred, R. (1964), *Foundations of Behavioural Research*, Surjeet Publications, Delhi.

Lal Das, D.K. (2000), *Practice of Social Research : A Social Work Perspective*, Rawat Publications, Jaipur.

Monete, Duane, R. et. al. (1986), *Applied Social Research: Tool For the Human Services*, Holt, Chicago.

Nachmias, D and Nachmias, C. (1981), *Research Methods in the Social Sciences*, St. Martins press, New York.

Sellitz, G. et. al. (1973), *Research Methods in Social Relations,* Holt, Rinehart and Winston (3rd edition), New York.

Sjoberg, G. and Nett. B. (1968), *A Methodology of Social Research,* Harper & Row, New York.

Wilkinson, T.S. and Bhandarkar, P.L. (1977), *Methodology and Techniques of Social Research*, Himalaya, Bombay.

5.13 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) While data processing refers to editing and coding of responses, preparation of master charts and computing of the scores, where as, data analysis involves tabulation, cross tabulation and explanation of data to study the relationships between the variables. Analysis of data is made with reference to the objectives of the study and research questions if any. It is also designed to test the hypothesis and causal inferences.

Check Your Progress II

1) Coding of data involves assigning of numbers to each response of the question. The purpose of giving numbers is to translate verbal data into numerical data, which may be counted and tabulated. For example, a response 'strongly agree' is coded as '1' the subsequent codes would be in order. Therefore, if there are responses like 'agree' 'undecided' 'disagree' and 'strongly disagree' they have to be coded as '2', '3' ,'4' and '5'.

2) After a code book is prepared, the data can be transferred from interview schedules/questionnaires to a master chart with the help of the code book. Entering data from master chart to computer is much more advantageous than entering data directly to computers because one can check the wrong entries in the computer by comparing 'data listing' as a computer output and master chart. Entering data directly to computer is disadvantageous, as there is no way to check wrong entries, which will show inconsistencies in tabulated data at the later stages of tabulation.

Check Your Progress III

1) An analytic model is a plan for data analysis. In other words, an analytic model is diagrammatic presentation of variables and their interrelationships. The purpose of preparing an analytic model is to visualise relationships between the variables as per the objectives of the study. It is also designed to test the hypothesis.

2) A bivariate table presents data of two variables in column and row simultaneously. When a researcher is interested in knowing the relationships between two variables the data pertaining to the variables are cross tabulated and a bivariate table is prepared.

   Sometimes researcher might be interested in knowing whether there is a third variable which is effecting the relationships between two variables. In such cases the researcher has to examine the bivariate relationship by controlling the effects of third variable. This is known as trivariate analysis. One way of controlling the effects of a third variable is to prepare partial tables and examine the bivariate relationship.