UNIT 6 PROCESSING OF DATA

STRUCTURE

6.0 OBJECTIVES

After studying this unit, you should be able to:

- evaluate the steps involved in processing of data,
- check for obvious mistakes in data and improve the quality of data,
- describe various approaches to classify data,
- construct frequency distribution of discrete and continuous data, and
- develope appropriate data tabulation device.

6.1 INTRODUCTION

In Unit 3 we have discussed various methods of collection of data. Once the collection of data is over, the next step is to organize data so that meaningful conclusions may be drawn. The information content of the observations has to be reduced to a relatively few concepts and aggregates. The data collected from the field has to be processed as laid down in the research plan. This is possible only through systematic processing of data. Data processing involves editing, coding, classification and tabulation of the data collected so that they are amenable to analysis. This is an intermediary stage between the collection of data and their analysis and interpretation. In this unit, therefore, we will learn about different stages of processing of data in detail.

6.2 EDITING OF DATA

Editing is the first stage in data processing. Editing may be broadly defined to be a procedure, which uses available information and assumptions to substitute inconsistent values in a data set. In other words, editing is the process of examining the data collected through various methods to detect errors and omissions and correct them for further analysis. While editing, care has to be

taken to see that the data are as accurate and complete as possible, units of observations and number of decimal places are the same for the same variable. The following practical guidelines may be handy while editing the data:

1) The editor should have a copy of the instructions given to the interviewers.

2) The editor should not destroy or erase the original entry. Original entry should be crossed out in such a manner that they are still legible.

3) All answers, which are modified or filled in afresh by the editor, have to be indicated.

4) All completed schedules should have the signature of the editor and the date.

For checking the quality of data collected, it is advisable to take a small sample of the questionnaire and examine them thoroughly. This helps in understanding the following types of problems: (1) whether all the questions are answered, (2) whether the answers are properly recorded, (3) whether there is any bias, (4) whether there is any interviewer dishonesty, (5) whether there are inconsistencies. At times, it may be worthwhile to group the same set of questionnaires according to the investigators (whether any particular investigator has specific problems) or according to geographical regions (whether any particular region has specific problems) or according to the sex or background of the investigators, and corrective actions may be taken if any problem is observed.

Before tabulation of data it may be good to prepare an operation manual to decide the process for identifying inconsistencies and errors and also the methods to edit and correct them. The following broad rules may be helpful.

Incorrect answers: It is quite common to get incorrect answers to many of the questions. A person with a thorough knowledge will be able to notice them. For example, against the question “Which brand of biscuits do you purchase?” the answer may be “We purchase biscuits from ABC Stores”. Now, this questionnaire can be corrected if ABC Stores stocks only one type of biscuits, otherwise not. Answer to the question “How many days did you go for shopping in the last week?” would be a number between 0 and 7. A number beyond this range indicates a mistake, and such a mistake cannot be corrected. The general rule is that changes may be made if one is absolutely sure, otherwise this question should not be used. Usually a schedule has a number of questions and although answers to a few questions are incorrect, it is advisable to use the other correct information from the schedule rather than discarding the schedule entirely.

Inconsistent answers: When there are inconsistencies in the answers or when there are incomplete or missing answers, the questionnaire should not be used. Suppose that in a survey, per capita expenditure on various items are reported as follows: Food – Rs. 700, Clothing – Rs.300, Fuel and Light – Rs. 200, other items – Rs. 550 and Total – Rs. 1600. The answers are obviously inconsistent as the total of individual items of expenditure is exceeding the total expenditure.

Modified answers: Sometimes it may be necessary to modify or qualify the answers. They have to be indicated for reference and checking.

Numerical answers to be converted to same units: Against the question "What is the plinth area of your house?" answers could be either in square feet or in square metres. It will be convenient to convert all the answers to these questions in the same unit, square metre for example.

6.3 CODING OF DATA

Coding refers to the process by which data are categorized into groups and numerals or other symbols or both are assigned to each item depending on the class it falls in. Hence, coding involves: (i) deciding the categories to be used, and (ii) assigning individual codes to them. In general, coding reduces the huge amount of information collected into a form that is amenable to analysis.

A careful study of the answers is the starting point of coding. Next, a coding frame is to be developed by listing the answers and by assigning the codes to them. A coding manual is to be prepared with the details of variable names, codes and instructions. Normally, the coding manual should be prepared before collection of data, but for open-ended and partially coded questions. These two categories are to be taken care of after the data collection. The following are the broad general rules for coding:

1) Each respondent should be given a code number (an identification number).

2) Each qualitative question should have codes. Quantitative variables may or may not be coded depending on the purpose. Monthly income should not be coded if one of the objectives is to compute average monthly income. But if it is used as a classificatory variable it may be coded to indicate poor, middle or upper income group.

3) All responses including "don't know", "no opinion" "no response" etc., are to be coded.

Sometimes it is not possible to anticipate all the responses and some questions are not coded before collection of data. Responses of all the questions are to be studied carefully and codes are to be decided by examining the essence of the answers. In partially coded questions, usually there is an option "Any Other (specify)". Depending on the purpose, responses to this question may be examined and additional codes may be assigned.

Self Assessment Exercise A

1) How would you edit the research data?

.....

.....

.....

2) What do you mean by coding?

.....

.....

.....

6.4 CLASSIFICATION OF DATA

Once the data is collected and edited, the next step towards further processing the data is classification. In most research studies, voluminous data collected through various methods needs to be reduced into homogeneous groups for meaningful analysis. This necessitates classification of data, which in simple terms is the process of dividing data into different groups or classes according to their similarities and dissimilarities. The groups should be homogeneous within and heterogeneous between themselves. Classification condenses huge amount of data and helps in understanding the important underlying features. It enables us to make comparison, draw inferences, locate facts and also helps in bringing out relationships, so as to draw meaningful conclusions. In fact classification of data provides a basis for tabulation and analysis of data.

6.4.1 Types of Classification

Data may be classified according to one or more external characteristics or one or more internal characteristics or both. Let us study these kinds with the help of illustrations.

6.4.1.1 Classification According to External Characteristics

In this classification, data may be classified according to area or region (Geographical) and according to occurrences (Chronological).

Geographical: In this type of classification, data are organized in terms of geographical area or region. State-wise production of manufactured goods is an example of this type. Data collected from an all India market survey may be classified geographically. Usually the regions are arranged alphabetically or according to the size to indicate the importance.

Chronological: When data is arranged according to time of occurrence, it is called chronological classification. Profit of engineering industries over the last few years is an example. We may note that it is possible to have chronological classification within geographical classification and *vice versa*. For example, a large scale all India market survey spread over a number of years.

6.4.1.2 Classification According to Internal Characteristics

Data may be classified according to attributes (Qualitative characteristics which are not capable of being described numerically) and according to the magnitude of variables (Quantitative characteristics which are numerically described).

Classification according to attributes: In this classification, data are classified by descriptive characteristic like sex, caste, occupation, place of residence etc. This is done in two ways – simple classification and manifold classification. In **simple classification** (also called classification according to dichotomy), data is simply grouped according to presence or absence of a single characteristics – male or female, employee or unemployee, rural or urban etc. In **manifold classification** (also known as multiple classification), data is classified according to more than one characteristic. First, the data may be divided into two groups according to one attribute (employee and unemployee, say). Then using the remaining attributes, data is sub-grouped again (male and

female based on sex). This may go on based on other attributes, like married and unmarried, rural and urban so on… The following table is an example of manifold classification.

Population			
Employee		Unemployee	
Male	Female	Male	Female

Classification according to magnitude of the variable: This classification refers to the classification of data according to some characteristics that can be measured. In this classification, there are two aspects: one is variables (age, weight, income etc;) another is frequency (number of observations which can be put into a class). Quantitative variables may be, generally, divided into two groups - discrete and continuous. A **discrete variable** is one which can take only isolated (exact) values, it does not carry any fractional value. The examples are number of children in a household, number of departments in an organization, number of workers in a factory etc. The variables that take any numerical value within a specified range are called **continuous variables**. The examples of continuous variables are the height of a person, profit/loss of companies etc. One point may be noted. In practice, even the continuous variables are measured up to some degree of precision and they also essentially become discrete variables.

The following are two examples of discrete and continuous frequency distribution placed side by side.

a) Discrete frequency distribution		**b) Continuous frequency distribution**	
No. of children	**No. of families**	**Income (Rs.)**	**No. of families**
0	12	1,000-2,000	6
1	25	2,000-3,000	10
2	20	3,000-4,000	15
3	7	4,000-5,000	25
4	3	5,000-6,000	9
5	1	6,000-7,000	4
Total	68	Total	69

6.4.1.3 Preparation of Frequency Distribution

When raw data is arranged in conveniently organized groups, it is called a frequency distribution. The number of data points in a particular group is called frequency. When a discrete variable takes a small number of values (not more than 8 or 10, say), each of the observed value is counted to form the discrete frequency distribution. In order to facilitate counting, prepare a column of “tallies”. The following example illustrates it.

Illustration 1: A survey of 50 college students was conducted to know that how many times a week they go to the theatre to see movies. The following data were obtained:

3 2 2 1 4 1 0 1 1 2 4 1 3 3 2 1 3 4 3 2 0 1 3 4 3

1 4 3 2 2 1 3 1 2 3 2 3 4 4 2 4 3 4 2 3 3 2 0 4 3

To have a discrete frequency table, we may take the help of 'Tally' marks as indicated below.

Table 6.1: Frequency Distribution of Number of Movies Seen by 50 College Students in a Week

Number of Days	Tally Marks	Frequency
0	𝍸	5
1	𝍸 | | |	8
2	𝍸 𝍸 | |	12
3	𝍸 𝍸 𝍸	15
4	𝍸 𝍸	10
Total		50

From the above frequency table it is clear that more than half the students (27 out of 50) go to the theatre twice or thrice a week and very few do not go even once a week. These were not so obvious from the raw data.

It is possible to prepare frequency distribution for qualitative variables also. For example, one may construct a frequency distribution of brands of 100 cars or blood groups of 50 patients in a hospital

Construction of a Continuous Frequency Distribution: In continuous frequency distribution, the data is grouped into a small number of intervals instead of individual values of the variables. These groups are called classes. There are two different ways in which limits of classes may be arranged - exclusive and inclusive method. In the **exclusive method**, the class intervals are so arranged that the upper limit of one class is the lower limit of the next class, whereas in the **inclusive method**, the upper limit of a class is included in the class itself. The same frequency distribution is shown below using inclusive method. As an example, the frequency distribution is shown below using exclusive method and inclusive method.

Illustration 2

Table 6.2 : Frequency Distribution of Daily Wages of 65 Labourers.

Exclusive method		Inclusive method	
Daily wages of Labourers (Rs.)	**No. of Labourers**	**Daily wages of Labourers (Rs.)**	**No. of Labourers**
20-30	2	20-29.99	2
30-40	15	30-39.99	15
40-50	21	40-49.99	21
50-60	29	50-59.99	29
60-70	13	60-69.99	13
Total	80	Total	80

In the exclusive method, the upper class limit of the first class is the same as the lower limit of the second class. A labourer with a daily wage of exactly Rs. 30 will be included in the second class. Thus, a class interval 20–30 means "20 and above but below 30". This is the exclusive method and the upper limit is always excluded.

In case of inclusive method, the upper limits of the classes are not the same as the Lower limits of their next classes. Thus, class interval 20-29.99 means "20 and above, and 29.99 and below". It is to be noted that both the methods give the same class frequencies, although the construction of classes look different. For computation of positional values such as median, mode etc., it is necessary to convert the inclusive classes into exclusive form. This can be done with the help of the following formula:

$$\text{Correction Factor} = \frac{\text{Lower limit of the succeeding class} - \text{upper limit of the class}}{2}$$

The result so obtained is deducted from all lower limits and added to all upper limits. For instance, in the above example, table 6.2, the correction factor is (30-29.99)/2 = 0.005. Deduct this value from the lower limit and add to the upper limit of each class. You will obtain the exclusive form of classes as 19.995-29.995; 29.995-39.995 and so on.

Steps to construct frequency distribution: The following broad guidelines may be followed for construction of a frequency distribution.

1) The highest and the lowest values of the observations are to be identified and the lower limit of the first class and upper limit of the last class may be decided.
2) The number of classes to be decided. There is no hard and fast rule. It should not be too little (lower than 5, say) to avoid high information loss. It should not be too many (more than 12, say) so that it does not become unmanageable.
3) The lower and the upper limits should be convenient numerals like 0-5, 0-10, 100-200 etc.
4) The class intervals should also be numerically convenient, like 5, 10, 20 etc., and values like 3, 9, 17 etc., should be avoided.
5) As far as possible, the class width may be made uniform for ease in subsequent calculation.

It is often quite useful to present the frequency distribution in two different ways. One way is **relative or percentage relative frequency distribution**. Relative frequencies can be computed by dividing the frequency of each class with sum of frequency. If the relative frequencies are multiplied by 100 we will get percentages. Another way is **cumulative frequency distribution** which are cumulated to give the total of all previous frequencies including the present class, cumulating may be done either from the lowest class (from below) or from the highest class (from above). The following table illustrates this concept.

Illustration 3

Table 6.3: Construction of Relative Frequency Distribution for the Data on Daily Wages of 70 Labourers

Class Interval	Frequency	Relative Frequency	Relative Frequency (as Percentage)	Cumulative Frequency (less than)	Cumulative Frequency (more than)
(1)	(2)	(3)	(4)	(5)	(6)
15-20	2	0.0286	2.86	2	70
20-25	23	0.3286	32.86	25	68
25-30	19	0.2714	27.14	44	45
30-35	14	0.2000	20.00	58	26
35-40	5	0.0714	7.14	63	12
40-45	4	0.0571	5.71	67	7
45-50	3	0.0429	4.29	70	3
Total	70	1.0000	100.00		

One advantage of using relative frequency distribution is that it helps in comparing two frequency distributions (with same class intervals). It is also the basis of empirical probability. The topic of Probability is the subject matter of Unit 13 and Unit 14 of this course.

Column (5) in the above table gives cumulative frequency of a particular class, which is obtained as discussed earlier. Cumulative frequency of the second class is obtained by adding of its class frequency (23) and the previous class frequency (2). Cumulative frequency of the next class is obtained by adding of its class frequency (19) to the cumulative frequency of the previous class (25). Cumulative frequencies may be interpreted as the number of observation below the upper class limit of a class. For example, a cumulative frequency of 44 in the third class (25-30) indicates that 44 labourers received a daily wage of less than Rs. 30. Cumulation from the highest class may also be done as shown in column (6). It has a similar interpretation.

At times relative frequencies are also cumulated to obtain cumulative relative frequency distribution.These cumulative frequencies are useful for researchers in two ways.

Firstly, Some simple graphs can be drawn to show all the frequency distributions. This is done in the next unit (Unit 7). Secondly, frequency distribution methods are also used for discrete data, if the number of observations is large and spread is more.

Bivariate frequency distribution: When data is collected on two variables, one may construct two frequency distributions separately. But it is possible to construct a two-way frequency distribution table. Here class intervals, based on value of one variable are placed in the row and class intervals based on value of the other variable is placed in the column. The following example illustrates this.

Illustration 4

Table 6.4: Bivariate Frequency Distribution of Sales and Profit of 200 Companies

Sl. No.	Sales (Rs. lakhs)	Profit (Rs. in thousands)						
		Upto 10	10-20	20-50	50-100	100-200	200 and more	Total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Upto 1	10	3					13
2	1-2	12	12	19				43
3	2-5	11	15	20	10	8		64
4	5-10	2	8	15	5	10		40
5	10-20		2	12	4	9	6	33
6	20 and more			2	1	2	2	7
7	Total	35	40	68	20	29	8	200

The above bivariate frequency table is prepared on the basis of sales and profit data of 200 companies. As discussed earlier, class limits for both Sales and Profit are decided first. Tally marks are placed in appropriate row and column (not shown here). Suppose a company's Sales and Profit figures are Rs. 2.5 lakhs and Rs. 49000 respectively. It is placed in class 3 of Sales (2 to 5 lakhs) and Column (5) showing class interval of profit 20 to 50 thousands. The last column (Column (9) gives the total over all class intervals of Profit. Hence it gives the frequency distribution of Sales. The frequency distribution in this column is known as Marginal Frequency distribution of Sales. Similarly, the figures in Serial No 7 (Row 7) are obtained by summing up over all the class intervals of Sales. This is the frequency distribution of profit or the Marginal Frequencies of Profit. The entire table is also known as Joint Frequency Distribution of Sales and Profit.

Self Assessment Exercise B

The following table gives values of production and values of raw materials used in 60 industrial units. Prepare (i) two individual frequency distributions for the variables. (ii) Prepare bivariate frequency table. For value of production you may take - 8000-9000, 9000-10000,…….., 130000 - 140000 as classes and for value of raw material you may take - 2500 - 3000, 3000-3500, ……,5000 - 5500 as class intervals.

Table : Values of Production and Values of Raw Material Used by 60 Industrial Units (Rs. in lakhs)

Sl. No.	Value of Production	Value of Raw Material	Sl. No.	Value of Production	Value of Raw Material
(1)	(2)	(3)	(1)	(2)	(3)
1	8952.69	2915.30	31	10394.05	3392.49
2	10147.82	3497.26	32	10751.31	3983.16
3	9938.00	3652.91	33	8685.23	3513.37
4	10278.61	3851.22	34	9393.46	3448.06
5	10225.37	3624.36	35	9352.66	3495.73
6	10324.95	3702.34	36	9405.84	3503.47
7	10921.96	3794.27	37	9692.46	3286.88
8	10885.23	4296.33	38	8783.27	3188.69
9	13324.16	5446.39	39	8963.49	3153.59
10	12154.16	3939.36	40	8956.34	3229.62
11	10835.82	3697.95	41	10920.31	3958.89
12	10864.50	3885.91	42	10094.23	3604.95
13	10698.37	3943.97	43	12038.21	4387.66
14	11136.22	4017.36	44	11199.94	4143.76
15	10644.51	3669.89	45	11522.36	3823.97
16	10070.44	3586.27	46	10862.59	3888.08
17	10857.33	3772.49	47	11797.08	3852.58
18	11561.56	4292.22	48	11762.56	3758.17
19	10544.07	4128.06	49	9687.11	3309.81
20	10163.53	3570.83	50	10905.26	3612.14
21	9580.77	3615.61	51	9806.59	4354.83
22	10493.76	3730.19	52	11614.45	2675.32
23	10454.36	3953.90	53	8260.93	3722.35
24	11026.60	3893.18	54	8498.49	3682.07
25	12808.22	4660.28	55	8796.21	3599.29
26	11681.71	4100.16	56	10372.99	3967.33
27	10631.05	3543.68	57	10714.01	4108.27
28	9441.43	3568.53	58	10212.08	3945.38
29	10311.69	3640.68	59	12817.07	3443.86
30	9152.95	3400.50	60	11865.77	3605.48

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

6.5 TABULATION OF DATA

Presentation of collected data in the tabular form is one of the techniques of data presentation. The two other techniques are diagrammatic and graphic presentation, which will be discussed in Unit 7 of this course. Arranging the data in an orderly manner in rows and columns is called tabulation of data. Sometimes data collected by survey or even from publications of official bodies are so numerous that it is difficult to understand the important features of the data. Therefore it becomes necessary to summarize data through tabulation to an easily intelligible form. It may be noted that there may be loss of some minor information in certain cases, but the essential underlying features come out more clearly. Quite frequently, data presented in tabular form is much easier to read and understand than the data presented in the text.

In classification, as discussed in the previous section, the data is divided on the basis of similarity and resemblance, whereas tabulation is the process of recording the classified facts in rows and columns. Therefore, after classifying the data into various classes, they should be shown in the tabular form.

6.5.1 Types of Tables

Tables may be classified, depending upon the use and objectives of the data to be presented, into simple tables and complex tables. Let us discuss them along with illustrations.

Simple Table: In this case data are presented only for one variable or characteristics. Therefore, this type of table is also known as one way table. The table showing the data relating to the sales of a company in different years will be an example of a single table.

Look at the following tables for an example of this type of table.

Illustration 5

Table 6.5 : Population of India During 1961–2001 (*In thousands*)

Census Year	Population
1961	439235
1971	548160
1981	683329
1991	846303
2001	1027015

Source: Census of India, various documents.

Any frequency distribution of a single variable is a simple table

Table 6.6 : Frequency Distribution of Daily Wages of 65 Labourers

Daily Wages of Laboures (Rs.)	No. of Labourers
20-30	2
30-40	5
40-50	21
50-60	19
60-70	11
70-80	5
80-90	2
Total	65

A simple table may be prepared for descriptive or qualitative data also. The following example illustrates it

Table 6.7 : Education Levels of 40 Labourers

Education Level	No. of Persons
Illiterate	22
Literate but below primary	10
Primary	5
High School	2
College and above	1
All	40

Complex Table: A complex table may contain data pertaining to more than one characteristic. The population data given below is an example.

Illustration 6

Table 6.8 : Rural and Urban Population of India During 1961–2001
(In thousands)

Census Year	Population		
	Rural	Urban	Total
1961	360298	78937	439235
1971	439046	109114	548160
1981	523867	159463	683329
1991	628691	217611	846303
2001	741660	285355	1027015

Note: The total may not add up exactly due to rounding off error.
Source: Census of India, various documents.

In the above example, rural and urban population may be subdivided into males and females as indicated below.

Table 6.9 : Rural and Urban Population of India During 1961–2001 (sex-wise)
(In thousands)

Census Year	Population					
	Rural		Urban		Total	
	Male	Female	Male	Female	Male	Female
(1)	(2)	(3)	(4)	(5)	(6)	(7)

In each of the above categories, the persons could be grouped into child and adult, worker and non-worker, or according to different age groups and so on. A particular type of complex table that is of great use in research is a cross-table, where the table is prepared based on the values of two or more variables. The bivariate frequency table used earlier (illustration 4) is reproduced here for illustration.

Illustration 7

Table 6.10 : Sales and Profit of 200 Companies

Sl. No.	Sales (Rupees in lakhs)	Profit (Rupees in thousands)						
		Upto 10	10-20	20-50	50-100	100-200	200 and more	Total
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Up to 1	10	3					13
2	1-2	12	12	19				43
3	2-5	11	15	20	10	8		64
4	5-10	2	8	15	5	10		40
5	10-20		2	12	4	9	6	33
6	20 and more			2	1	2	2	7
7	Total	35	40	68	20	29	8	200

From bivariate table, one may get some idea about the interrelationship between two variables. Suppose, that all the frequencies are concentrated in the diagonal cells, then there is likely to be a strong relationship. That is positive relationship if it starts from top-left corner to bottom-right corner or if it is from bottom-left corner to top-right corner then, we could say there is negative relationship. If the frequencies are more or less equally distributed over all the cells, then probably there is no strong relationship.

Multivariate tables may also be constructed but interpretation becomes difficult once we go beyond two variables.

So far we have discussed and learnt about the types of tables and their usefulness in presentation of data. Now, let us proceed to learn about the different parts of a table, which enable us to have a clear understanding of the rules and practices followed in the construction of a table.

6.5.2 Parts of A Statistical Table

A table should have the following four essential parts - title, caption or box head (column), stub (row heading) and main data. At times it may also contain an end note and source note below the table. The table should have a title, which is usually placed above the statistical table. The title should be clearly worded to give some idea of the table's contents. Usually a report has many tables. Hence the tables should be numbered to facilitate reference.

Caption refers to the totle of the columns. It is also termed as "box head". There may be sub-captions under the main caption. Stub refers to the titles given to the rows.

Caption and stub should also be unambiguous. To the extent possible abbreviations should not be used in either caption or stub. But if they are used, the expansion must be given in the end note below. Notes pertaining to stub entries or box headings may be numerals. But, to avoid confusion, it is better to use some symbols (like *, **, @ etc) or alphabets for notes referring to the entries in the main body. If the table is based on outside information, it should be mentioned in the source note below. This note should be complete with author, title, year of publication etc to enable the reader to go to the original source for crosschecking or for obtaining additional information. Columns and rows may be numbered for easy reference.

Some of these features are illustrated below with reference to the table on Rural and Urban Population during 1961-2001, which was presented in earlier illustration-6, Table 6.8.

1. Title of the Table	Table: Rural and Urban Population of India during 1961–2001 (in thousands)

2. Caption or Box Head	Population		
	Rural	Urban	Total

	Census Year
3. Stub (Row Heading)	1961
	1971
	1981
	1991
	2001

4. Body (Main Data)	360298	78937	439235
	439046	109114	548160
	523867	159463	683329
	628691	217611	846303
	741660	285355	1027015

5. End Note	Note: The total may not add up exactly due to rounding off of error.

6. Source Note	Source: Census of India, various documents.

Column Number	(1)	(2)	(3)	(4)	(5)

	1
	2
Row Number	3
	4
	5

The boxes above are self-explanatory.

Arrangement of items in stub and box-head

There is no hard and fast rule about the arrangement of column and row headings in a table. It depends on the nature of data and type of analysis. A number of different methods are used - alphabetical, geographical, chronological/ historical, magnitude-based and customary or conventional.

Alphabetical: This method is suitable for general tables as it is easy to locate an item if it is arranged alphabetically. For example, population census data of India may be arranged in the alphabetical order of states/union territories.

Geographical: It can be used when the reader is familiar with the usual geographical classification.

Chronological: A table containing data over a period of time may be presented in the chronological order. Population data (1961 to 2001) presented earlier (Tables 6.5 and 6.8) are in chronological order. One may either start from the most recent year or the earliest year. However, there is a convention to start with the month of January whenever year and month data are presented.

Based on Magnitude: At times, items in a table are arranged according to the value of the characteristic. Usually the largest item is placed first and other items follow in decreasing order. But this may be reversed also. Suppose that state-wise population data is arranged in order of decreasing magnitude. This will highlight the most populous state and the least populous state.

Customary or Conventional: Traditionally some order is followed in certain cases. While presenting population census data, usually 'rural' comes before 'urban' and 'male' first and 'female' next. At times, conventional geographical order is also followed.

One point may be noted. The above arrangements are not exclusive. In a big table, it is always possible and sometimes convenient to arrange the items following two or three methods together. For example, it is possible to construct a table in chronological order and within it in geographical order. Sometimes information of the same table may be rearranged to produce another table to highlight certain aspects. This will be clear from the following specimen tables.

Table A

Sl.No.	Census Year	Rural		Urban	
		Male	Female	Male	Female
(1)	(2)	(3)	(4)	(5)	(6)
1					
2					

Table B

Sl.No.	Census Year	Male		Female	
		Rural	Urban	Rural	Urban
(1)	(2)	(3)	(4)	(5)	(6)
1					
2					

Tables A and B contain the same information. Table A compares male-female differences for rural and urban areas whereas Table B highlights rural-urban contrasts for both the sexes.

Tables are prepared for making data easy to understand for the reader. It should not be very large as the focus may be lost. A large table may be logically broken into two or more small tables.

6.5.3 Requisites of a Good Statistical Table

After having an understanding of the parts of a statistical table, now let us discuss the features of an ideal statistical table. Besides the rules relating to part of the table, certain guidelines are very helpful in its preparation. They are as follows:

1) A good table must present the data in as clear and simple a manner as possible.

2) The title should be brief and self-explanatory. It should represent the description of the contents of the table.

3) Rows and Columns may be numbered to facilitate easy reference.

4) Table should not be too narrow or too wide. The space of columns and rows should be carefully planned, so as to avoid unnecessary gaps.

5) Columns and rows which are directly comparable with one another should be placed side by side.

6) Units of measurement should be clearly shown.

7) All the column figures should be properly aligned. Decimal points and plus or minus signs also should be in perfect alignment.

8) Abbreviations should be avoided in a table. If it is inevitable to use, their meanings must be clearly explained in footnote.

9) If necessary, the derived data (percentages, indices, ratios, etc.) may also be incorporated in the tables.

10) The sources of the data should be clearly stated so that the reliability of the data could be verified, if needed.

Self Assessment Exercise C

The following report is obtained from 50 unskilled workers in a factory in Faridabad. Prepare three simple tables based on caste, education and place of origin and a complex table by considering all the factors.

S. No	Caste	Education	Place of Origin
1	OC	LITERATE BUT BELOW PRIMARY	URBAN
2	OC	LITERATE BUT BELOW PRIMARY	URBAN
3	OC	PRIMARY	RURAL
4	BC	ILLITERATE	RURAL
5	SC	LITERATE BUT BELOW PRIMARY	URBAN
6	SC	PRIMARY	RURAL
7	ST	ILLITERATE	RURAL
8	OC	HIGH SCHOOL	RURAL
9	SC	HIGH SCHOOL	RURAL
10	ST	ILLITERATE	RURAL
11	OC	LITERATE BUT BELOW PRIMARY	RURAL
12	OC	PRIMARY	RURAL
13	OC	HIGH SCHOOL	RURAL
14	BC	ILLITERATE	RURAL
15	SC	PRIMARY	RURAL
16	SC	PRIMARY	RURAL
17	OC	HIGH SCHOOL	RURAL
18	OC	HIGH SCHOOL	RURAL
19	SC	LITERATE BUT BELOW PRIMARY	RURAL
20	ST	PRIMARY	URBAN
21	SC	PRIMARY	RURAL
22	SC	HIGH SCHOOL	RURAL
23	ST	PRIMARY	RURAL
24	OC	LITERATE BUT BELOW PRIMARY	RURAL
25	OC	PRIMARY	RURAL
26	OC	HIGH SCHOOL	URBAN
27	OC	PRIMARY	URBAN

28	OC	PRIMARY	URBAN
29	BC	ILLITERATE	RURAL
30	SC	ILLITERATE	RURAL
31	SC	ILLITERATE	URBAN
32	ST	ILLITERATE	URBAN
33	OC	PRIMARY	URBAN
34	OC	PRIMARY	RURAL
35	BC	ILLITERATE	URBAN
36	OC	HIGH SCHOOL	URBAN
37	OC	HIGH SCHOOL	URBAN
38	OC	ILLITERATE	RURAL
39	BC	ILLITERATE	URBAN
40	OC	PRIMARY	RURAL
41	BC	LITERATE BUT BELOW PRIMARY	RURAL
42	BC	ILLITERATE	URBAN
43	OC	ILLITERATE	RURAL
44	BC	PRIMARY	RURAL
45	BC	LITERATE BUT BELOW PRIMARY	RURAL
46	SC	PRIMARY	RURAL
47	SC	ILLITERATE	URBAN
48	ST	PRIMARY	RURAL
49	OC	PRIMARY	RURAL
50	OC	LITERATE BUT BELOW PRIMARY	RURAL

6.6 LET US SUM UP

Once data collection is over, the next important steps are editing and coding. Editing helps in maintaining consistency in quality of data. Editing is the first stage in data processing. It is the process of examining the data collected to detect errors and omissions and correct them for further analysis. Coding makes further computation easier and necessary for efficient analysis of data. Coding is the process of assigning some symbols to the answers. A coding frame is developed by listing the answers and by assigning the codes to them. The next stage is classification. Classification is the process of arranging data in groups or classes on the basis of some chracteristics. It helps in making comparisons and drawing meaningful conclusions. The classified data may be summarized by means of tabulations and frequency distributions. Cross tabulation is particularly useful as it provides some clue about relationship and its direction between two variables. Frequency distribution and its extensions provide simple means to summarize data and for comparison of two sets of data.

6.7 KEY WORDS

Attribute : Characteristics that are not capable of being measured.

Caption or Box-head : Column headings of a table.

Class Interval : The difference between the upper and lower limits of a class.

Class Limits : The lowest and the highest values that can be included in the class.

Coding : A method to categorize data into groups and assign numerical values or symbols to represent them.

Continuous Variable : A variable that can take values to any degree of precision.

Cumulative Frequency Distribution : A distribution which shows cumulative frequencies instead of actual frequencies.

Discrete Variable : A variable that can take only certain values (but not fractional values).

Editing : Methods to substitute inconsistent values in a data set.

Exclusive Class : A class at which the upper limit is excluded from that class and included as lower limit in next class.

Frequency Distribution : Distribution of frequencies (number of observations) over different classes of a variable.

Joint Frequency Distribution : Distribution of frequencies (number of observations) over different classes of two or more variables.

Inclusive Class : A class in which both its lower and upper limits are included in that class.

Stub : Row headings of a table.

6.8 ANSWERS TO SELF ASSESSMENT EXERCISES

B. Frequency Distribution for Value of Production (Rupees in lakh)

Sl. No.	Class Interval	Frequency
(1)	(2)	(3)
1	80–90	8
2	90–100	10
3	100–110	27
4	110–120	10
5	120–130	4
6	130–140	1
7	Total	60

Frequency Distribution for Value of Raw Material (Rupees in lakh)

Sl. No.	Class Interval	Frequency
(1)	(2)	(3)
1	25–30	2
2	30–35	11
3	35–40	36
4	40–45	9
5	45–50	1
6	50–55	1
7	Total	60

C. Table : Caste-wise distribution of 50 Unskilled Workers

Sl. No.	Caste	No of Workers
(1)	(2)	(3)
1	SC	12
2	ST	6
3	BC	9
4	OC	23
5	All Castes	50

Table : Distribution of 50 Unskilled Workers According to Educational Level

Sl. No.	Educational Level	No. of Workers
(1)	(2)	(3)
1	Illiterate	14
2	Literate but below Primary	9
3	Primary	18
4	High School	9
5	All	50

Table : Distribution of 50 Unskilled Workers According to Place of Origin

Sl. No.	Place of Origin	No. of Workers
(1)	(2)	(3)
1	Rural	34
2	Urban	16
5	All	50

Complex Table

Table : Distribution of 50 Unskilled Workers

Education Level	Place of Origin										Total
	Rural					Urban					
	SC	ST	BC	OC	Total	SC	ST	BC	OC	Total	
Illiterate	1	2	3	2	8	2	1	3	0	6	14
Below Primary	1	0	2	3	6	1	0	0	2	3	9
Primary	5	2	1	6	14	0	1	0	3	4	18
High School	2	0	0	4	6	0	0	0	3	3	9
Total	9	4	6	15	34	3	2	3	8	16	50

6.9 TERMINAL QUESTIONS/EXERCISES

1) What do you mean by Editing of data? Explain the guidelines to be kept in mind while editing the statistical data.
2) Explain the meaning of coding? How would you code your research data?
3) "Classification of data provides a basis for tabulation of data. Comment.
4) Discuss the various methods of classification.

5) Form a frequency distribution for the following data by inclusive method and exclusive method.

13	18	17	20	22	15	27	14	7	10	10	16
9	6	15	11	19	21	25	23	28	9	25	9
27	11	30	13	14	2	34	18	28	25	28	12
14	15	18	18	16	20	21	24	21	16	22	4

6) Draw a "less than" and "more than" cumulative frequency distribution for the following data.

Income (Rs.)	500-600	600-700	700-800	800-900	900-1000
No. of families	25	40	65	35	15

7) What is tabulation? Draw the format of a statistical table and indicate its various parts.

8) Describe the requisites of a good statistical table.

9) Prepare a blank table showing the age, sex and literacy of the population in a city, according to five age groups from 0 to 100 years.

10) The following figures relate to the number of crimes (nearest-hundred) in four metropolitan cities in India. In 1961, Bombay recorded the highest number of crimes i.e. 19,400 followed by Calcutta with 14,200, Delhi 10,000 and Madras 5,700. In the year 1971, there was an increase of 5,700 in Bombay over its 1961 figure. The corresponding increase was 6,400 in Delhi and 1,500 in Madras. However, the number of these crimes fell to 10,900 in the case of Calcutta for the corresponding period. In 1981, Bombay recorded a total of 36,300 crimes. In that year, the number of crimes was 7,000 less in Delhi as compared to Bombay. In Calcutta the number of crimes increased by 3,100 in 1981 as compared to 1971. In the case of Madras the increase in crimes was by 8,500 in 1981 as compared to 1971. Present this data in tabular form.

Note: These questions/exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university for assessment. These are for your practice only.

6.10 FURTHER READING

A number of good text books are available for the topics dealt with in this unit. The following books may be used for more in depth study.

1) Croxton, F E, D J Cowden and S Klein, 1979. *Applied General Statistics*, Prentice Hall of India, New Delhi.

2) Saravanavel, P, 1987. *Research Methodology*, Kitab Mahal, Allahabad.

3) Spiegel, M R, 1992. *Statistics*, Schaum's Outline Series, Mc Graw Hill, Singapore.