UNIT 10 TABULATION AND ORGANIZATION OF DATA

Structure

10.1 INTRODUCTION

In the last three units, we dealt with the nature of various tools used in the collection of data. These data are mostly expressed in quantified uncategorized metric terms. However, quantitative data may not be available in certain cases. In such a situation, the researcher has to consider the qualitative data presented as nominal, ordinal or even categorized or coded metric data. Indeed, he/she should be familiar not only with the two types of data – quantitative and qualitative, but also with the process of classifying data, and graphical representation of data.

The aim of this unit is to make you understand the nature of quantitative and qualitative data, the procedures for classifying and tabulating quantitative data and presenting them graphically.

Objectives

After studying this unit, you will be able to:

- define quantitative and qualitative data,
- describe how to process the quantitative data,
- prepare various types of graphs and tables for presenting data, and
- explain how to organize qualitative data.

10.2 TYPES OF DATA: QUANTITATIVE AND QUALITATIVE

You may recall studying earlier in Unit 9 that the data collected through the administration of various types of tools on the selected samples are of two types i.e. qualitative and quantitative. In quantitative data, numerical values are assigned to the characteristics or properties of objects or events, according to logically accepted

rules. It is a process wherein a number system like figures, ratings or scores is imposed on empirical data. For example, characteristics on metric scale can have numeric outcome such as birth weight $\geq$ 25 kg, apgar score 7 for a particular child or respiration rate 68 per minute. However, when the researcher takes into consideration the phenomenon as a whole and does not attempt to analyze it in measurable or quantifiable terms, the approach becomes 'qualitative'. Generally, in epidemiological research, both types of data, (i) quantitative and (ii) qualitative, are recognized. Use of quantitative variables such as body temperature, births weight and gestation is of course common, but at some stage, they tend to be interpreted as "qualities" such as high/normal/low and preterm/term etc. The clinical interpretability becomes easier by assigning such "qualities".

We look at the characteristics of both i.e. quantitative and qualitative data in the following sections. Section 10.3 deals with quantitative data their tabulation, frequency distribution and cumulative frequency distributions, the need to represent data graphically, the various types of graphs, and the methods of analyzing quantitative data, viz., measures of central tendency, variability, relative positions and relationship. In section 10.4, we shall briefly discuss qualitative data and their analysis with reference to content analysis, logical analysis and inductive analysis. The application of various parametric and non-parametric tests is discussed in more detail in Unit 11.

10.3 PROCESSING OF QUANTITATIVE DATA

Once data are collected, the researcher turns his/her focus of attention on its tabulation and processing. In this sub-section, we will discuss about one of the most important stages of the research process, i.e. data processing and tabulation. We begin with data processing.

10.3.1 Data Processing

A researcher has to make his/her 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 a time that data processing is a job of a computer assistant. As a consequence, they have to be contended with the results given by computer assistant 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. In this sub-section we will discuss about data processing.

Data processing refers to certain operations such as editing, coding, computing of the scores, prepration of master charts, etc. After collection of filled in questionnaires, and other data editing of entries therein are not only necessary but also useful in making subsequent steps simpler. Many a times, a researcher or the assistant 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. In the process of editing, the researcher has to be very careful about consecutive questions having 'not applicable' as a response. Hence data coding becomes important. The next sub-section focuses on this aspect.

10.3.2 Coding of Data

Coding of data involves assigning of symbols (numerical) to each response of the question. The purpose of giving numerical symbols is to translate raw data into numerical data, which may be counted and tabulated. The task of researcher is to give numbers/codes (1,2,3....) to response carefully. As we have already discussed

various types of questions (such as open-end, etc.,) in the previous block, 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 'other' 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, and 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 into 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 could 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 a code book used in a study is reproduced in Table 10.1.

Table 10.1: Code book

Q. No.	Var. No.	Information Sought	Responses	Code	Column No.	Remarks
		Respondents Number			1 - 3	
2	1	Age	Actual Worker	- 1	4 - 5	
3	2	Designation	Supervisor Manager	2 3	6	
4	3	Establishment	Public Private Graduate Intermediate	1 2 1 2	7	
5	4	Level of education	High School Middle School Primary Illiterate Other Married	3 4 5 6 7 1	8	
6	5	Marital status	Unmarried Widow Divorce	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 mothers towards breast feeding	Agree Undecided Disagree	1 2 3	14	

Once the data is coded, the next step involves preparing a master chart which is a crucial step as highlighted next.

10.3.3 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 10.2.

Table 10.2: Master chart

Correspondent Number				Variable Lables											
				Age	Designation	Establishment	Education	Marital Status	Nature of Work	Duration of Work	Wages	Promotions	Attitude of Employer	Weight (kg)	Height (cm)
				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	3		
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	5		
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		

Having gone through the concepts presented above certainly you would be better equipped to handle your data and process it in a way which will help you tabulate and organize your results in a scientific manner. Next, we shall review tabulation and organization of quantitative data.

10.4 TABULATION AND ORGANIZATION OF QUANTITATIVE DATA

Tables and graphs are commonly used to tabulate the data on a large number of subjects in a condensed and summarized form. Let us get to learn how to use tables in data tabulation. We may use the frequency table or cumulating frequency distribution or the contingency table as highlighted herewith.

10.4.1 Frequency Distribution

Data collected from a test and by using other gathering/measuring tools are raw and may have little meaning to the researcher until they are tabulated and organized in a systematic order. One of the ways of doing so is to prepare a frequency table or a frequency distribution which depicts the number of subjects distributed among the various groups or categories of characteristics. The method for tabulating the quantified data in a frequency distribution can be illustrated by considering the following scores of 40 students of say, Master's Programme in Dietetics and Food Service Management of the Indira Gnadhi National Open University in Course MFN-009 presented in Table 10.3.

Table 10.3: Tabulation of scores on a test in course MNF-009

57	70	80	82	87
60	72	80	82	88
64	73	80	82	87
67	70	78	80	93
67	76	77	84	95
62	76	78	85	97
61	75	80	85	98
63	70	78	85	90

It is difficult to see from the above list how the scores are distributed. Inspection of scores, however, shows that many scores occur more than once.

We observe that there are one 98, one 95, one 88, two 87s, three 85s, and so on. For our convenience, we can arrange the data in columns as shown in Table 10.4. In one column, we can arrange the marks in class-intervals and in the other, we can record the number of students who have scored these marks by tallies. Inspection of the scores in Table 10.4 shows that the highest score is 98 and the lowest is 57. The range is 41 (i.e. 98 – 57). Therefore, the distribution of scores can be conveniently arranged by dividing the range of 41 into eight or more class-intervals if the classes are taken to be of 5 points each. If we take the starting point at 56, the scores within the range 56 to 60, that is all scores with the values 57 and 60 will be grouped together to form the lowest class-interval. All scores from 61 to 65, that is, 61,62,63,64 and 65 will form the next class-interval. Similarly we shall group all scores within the ranges 66 to 70, 71 to 75 and so on. The highest class interval will be 96 - 100. Note, these are nine groups. The choice mostly depends on common sense evaluation of the utility of such groups in conveying the basic feature of the data. As a rule of thumbs it may be suggested that the number of such groups should generally be between four and eight.

Table 10.4: Frequency distribution of the scores of 40 students: Course MFN-009

Class Interval	Tallies	Frequency (f)
96 - 100	II	2
91 - 95	II	2
86 - 90	IIII	4
81 - 85	~~IIII~~ II	7
76 - 80	~~IIII~~ ~~IIII~~ I	11
71 - 75	III	3
66 - 70	~~IIII~~	5
61 - 65	IIII	4
56 - 60	II	2
Total number of scores N = 40		

In Table 10.4, the class-intervals have been arranged serially from the smallest at the bottom to the largest at top, each class-interval covering 5 scores. For each score, we have marked a 'tally' against the corresponding class-interval. The first score, 57, is represented by a tally placed against the class interval 56 - 60, the second score of 60 by a 'tally' marked against the class interval 56 - 60, and the third score 64 by a tally against the class interval 61 - 65. the remaining scores have been tabulated in the same way. When all the 40 scores are listed, the total number of tallies in each class-interval are counted and written in the next column f called 'frequency'. The total of 'f' gives the total number of scores (in the present case 40) and is denoted by N.

It may be noted that the interval 56 - 60 takes care of all scores from 56 upto 60. The score of 56 ordinarily means the interval 55.5 to 56.5 and that the score of 60 means 59.5 to 60.5. The mid-point of the bottommost class-interval is 58. Hence, the distribution represented in Table 10.4 may also be expressed as distribution given in Table 10.5.

Table 10.5: Frequency distribution of the scores of 40 students: Course MFN-009

Score Intervals	Exact Units of Class Intervals	Mid Point(x)	(f)
96 - 100	95.5 - 100.5	98	2
91 - 95	90.5 - 95.5	93	2
86 - 90	85.5 - 90.5	88	4
81 - 85	80.5 - 85.5	83	7
76 - 80	75.5 - 80.5	78	11
71 - 75	70.5 - 75.5	73	3
66 - 70	65.5 - 70.5	68	5
61 - 65	60.5 - 65.5	63	4
56 - 60	55.5 - 60.5	58	2
			N = 40

The distribution is *univariate* when the division of the subjects is presented by categories of one variable only as illustrated in Table 10.6 (i.e. scores on one test). It is *bivariate* when presented by categories of two variables simultaneously. An example of bivariate would be the frequency distribution of anaemic children by severity of disease and gender.

With this basic understanding about frequency distribution we next focus on cumulative frequency distribution.

10.4.2 Cumulative Frequency Distribution

In some cases, we may not be concerned with the frequencies within the class-intervals, but rather with the number or the percentage of values greater than or less than a specified value. The main purpose of computing a percentage is to be able to compare groups or class intervals in a frequency table. These values, called 'cumulative frequencies' or 'cumulative percentage frequencies' are obtained by adding successively the individual frequencies of class-intervals as shown in Table 10.6. We start with the frequency of the lowest score interval (56 - 60).

Table 10.6: Cumulative frequency distribution of the test scores of 40 students

Score Intervals	Exact Unit of Class-intervals	Frequency (f)	Cumulative Frequency (F)	Cumulative Percentage Frequency (%)
96 - 100	95.5 - 100.5	2	38 + 2 = 40	100.00 (40/40×100)
91 - 95	90.5 - 95.5	2	36 + 2 = 38	95.00 (38/40×100)
86 - 90	85.5 - 90.5	4	32 + 4 = 36	90.00 (36/40×100)
81 - 85	80.5 - 85.5	7	25 + 7 = 32	80.00 (32/40×100)
76 - 80	75.5 - 80.5	11	14 + 11 = 25	62.50 (25/40×100)
71 - 75	70.5 - 75.5	3	11 + 3 = 14	35.00 (14/40×100)
66 - 70	65.5 - 70.5	5	6 + 5 = 11	27.50 11/40×100)
61 - 65	60.5 - 65.5	4	2 + 4 = 6	15.00 (6/40×100)
56 - 60	55.5 - 60.5	2	2	5.00 (2/40×100)
N = 40				

Finally let us learn about the contingency table.

10.4.3 Contingency Tables

Contingency tables are used to record and analyze the relationship between two or more variables, most usually categorical variables. Suppose that we have two variables, sex (male or female) and iron status (anaemic or non-anaemic). We observe the values of both variables in a random sample of 100 people. Then a contingency table can be used to express the relationship between these two variables, as follows:

	Anaemic	Non-anaemic	Total
Male	9	43	52
Female	44	4	48
Total	**53**	**47**	**100**

From the example presented above, it must be evident that contingency table is a table of frequencies that shows the observed frequencies of data elements classified according to two variables, with the rows indicating one variable (iron status) and the columns indicating the other variable. The example above is for the simplest kind of contingency table, in which each variable has only two levels; this is called a 2 × 2 contingency table. A contingency table is called one-way, two-way, or r-way depending upon the number of variables on which the subjects are classified.

In a contingency table the subjects are classified into *mutually exclusive* and *exhaustive* categories. Categories are called mutually exclusive when only one of them is applicable to one subject and exhaustive when a subject cannot be classified beyond the specific categories. Let us understand this with the help of an example given in Table 10.7.

Table 10.7: Haemoglobin distribution among children according to age and sex

Age (months)	Male				Female			
	Anaemic		Normal		Anaemic		Normal	
	No.	%	No.	%	No.	%	No.	%
9 - 12	22	14.2	16	16.7	35	19.8	16	16.8
13 - 18	49	31.6	21	21.9	47	26.6	26	27.4
19 - 24	35	22.6	21	21.9	36	20.3	25	26.3
25 - 30	34	21.9	24	25.0	37	20.9	13	13.7
31 - 36	15	9.7	14	14.5	22	12.4	15	15.8
Total	**155**	**100.0**	**96**	**100.0**	**177**	**100.0**	**95**	**100.0**

Source: Kapur D, Agarwal KN, Sharma S. Detecting Iron Deficiency Anemia among Children (9 - 36 months of age) by implementing a Screening Programme in an Urban Slum. Indian Pediatr 2002; 39:671 - 676.

Table 10.7 is an example of a three-way contingency table, which presents age distribution of children according to sex and iron status. In this table, metric scale for age in months is grouped while the sex (gender) is ordinal. The table shows frequencies of children in various age groups according to sex and iron status. The categories are mutually exclusive. For example, an anaemic girl of age 11 months belongs to only one category in the frequency table. In the same way, no girl (or for that matter no boy) can be classified beyond these categories shown in this table. If the column for normal iron status is omitted then the categories are not exhaustive and the table is not a contingency table.

Now contrast this with another example of frequency distribution given in Table 10.8.

Table 10.8: Frequency distribution of symptoms and/or signs (morbidity picture) in children

Sign/Symptoms	Number	Per Cent (%)
Upper Respiratory Tract Infection	243	44.2
Diarrhoea	120	21.0
Fever	101	18.5
Pallor	125	22.9
Pica	170	31.0
Underweight	487	89.4
Worms	63	11.6
Other Infections (Boils, Ear infection, Urinary tract infection)	36	6.6
Total	**545**	**100**

Source: Kapur D, Agarwal KN, Sharma S. Detecting Iron Deficiency Anemia among Children (9-36 months of age) by implementing a Screening Programme in an Urban Slum. Indian Pediatr 2002; 39:671 - 676.

Table 10.8 on clinical symptoms and/or signs in 545 children, 9 - 36 months of age, screened for iron deficiency anaemia is an example where categories are not mutually exclusive. You may have noticed in Table 10.8 that one patient can have two or more symptoms. This is called *multiple response*. Table 10.8 contains frequencies, but is not a contingency table.

We hope the examples presented above may have helped you understand the concept of contingency tables. With this we end our study of how to tabulate and organize

the quantitative data. Next, we shall focus on graphical presentation of quantitative data. But first we shall take a break and try to recapitualte what we have learnt so far, by answring the questions given in check your progress exercise 1.

Check Your Progress Exercise 1

1) Define quantitative data. Describe the various types of quantitative data along with examples.

.....

.....

.....

.....

2) Tabulate the following data in a frequency distribution using an interval of 5 units.

185	176	166	177	171
147	176	170	171	180
173	168	181	165	173
175	158	156	162	173
197	151	153	162	188
166	145	191	164	174
178	142	158	167	178
148	187	172	169	184
156	187	172	193	183
181	161	172	179	179

.....

.....

.....

.....

.....

.....

.....

.....

3) What is a contengency table? What is their use in epidemiological research.

.....

.....

.....

.....

.....

.....

10.5 GRAPHICAL PRESENTATION OF QUANTITATIVE DATA

In the previous section we reviewed the tabular presentation to summarize data. Like tabular presentation, graphical presentation often facilitates understanding of a set of data. For visual display, graphs, diagrams, charts and maps are commonly used. They are generally referred to as "Figures" in the literature. With the help of a well-drawn graph, the data can be read and interpreted very easily. Brief descriptions of the various types of graph which are useful in visualizing the important properties of a frequency distribution are enumerated in sub-section 10.5.1. Sub-section 10.5.2 focuses on graphs for nominal and ordinal data. Graphs for relation between two variables are presented in sub-section 10.5.3.

10.5.1 Representation of Frequency Distribution

The following three types of graph are commonly used for representing frequency distribution. There include:

i) Histogram or column diagram

ii) Frequency polygon

iii) Cumulative percentage curve or ogive

Let us review these types of graphs.

i) ***Histogram or column diagram***

A histogram or column diagram is a graph in which class-intervals (about which we learnt earlier in section 10.4) are represented along the horizontal axis and their corresponding frequencies are represented by areas in the form of rectangular vertical bars drawn on the intervals.

The following steps are followed in preparing a histogram.

Step 1: A horizontal line is drawn at the bottom of a graph paper. Units representing class-intervals are marked along this line.

Step 2: A vertical line is drawn at the left hand extreme of the horizontal axis. Along this vertical axis, units representing individual frequencies of the class-intervals are marked.

Step 3: Taking class units as bases, rectangles are drawn, such that the areas of rectangles are proportional to the frequencies of the corresponding classes.

Let us consider the data presented in Table 10.9 for drawing a histogram as an illustration of what we have said above.

Table 10.9: Frequency distribution of the scores of 40 students

Class Intervals	Exact Units of Class Intervals	Mid-Point	Freq-uency (f)	Cumulative Frequency (F)	Cumulative Percentage Frequency
35 - 39	34.5 - 39.5	37	4	40	100.00
30 - 34	29.5 - 34.5	32	8	36	90.00
25 - 29	24.5 - 29.5	27	11	28	70.00
20 - 24	19.5 - 24.5	22	8	17	42.50
15 - 19	14.5 - 19.5	17	6	9	22.50
10 - 14	9.5 - 14.5	12	3	3	7.50
N = 40					

The histogram drawn for the above data is shown in Figure 10.1.

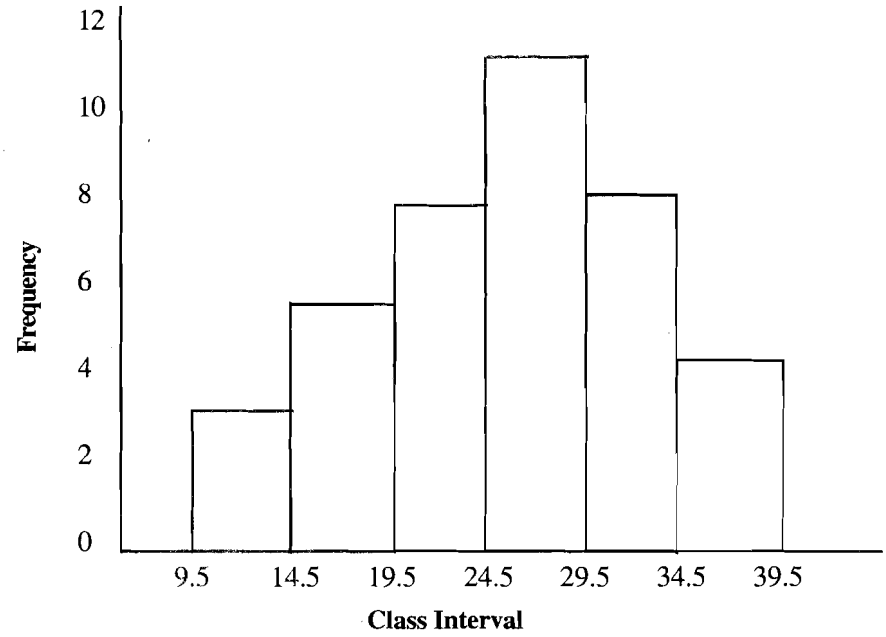

Figure 10.1: Histogram plotted from the data in Table 10.9

Thus histogram you would have noticed in Figure 10.1 is a set of contiguously drawn bars. The bars are drawn for each group/interval of values such that the area is proportional to the frequency in that group. Next, we move on to frequency polygon.

ii) ***Frequency Polygon***

Frequency polygon, you would notice, is a shape enclosed by straight lines. Frequency polygon is drawn by plotting the mid-point of each class-interval (i.e. bars in the histogram) at a height proportional to its respective frequency and then joining the points by straight lines including those with zero frequency at the two ends (Refer to Figure 10.2). The first two steps are identical to those used in the construction of a histogram. The next step to be followed is given as under:

Step 3: Directly above the mid-point of each class-interval along the horizontal axis plot the points at a height proportional to the respective frequencies. Join these points by straight lines. The frequency polygon for the distribution of Table 10.7 is shown in the Figure 10.2.

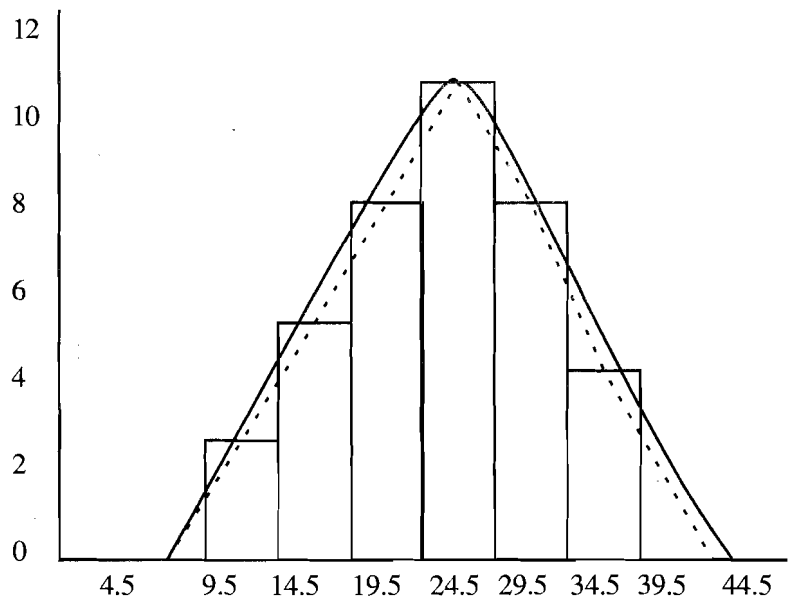

Figure 10.2: Frequency Polygon plotted from the data of Table 10.9

..... frequency polygon

— Frequency cruve

A frequency curve is a smooth curve as drawn in Figure 10.2. Next, let us review cumulative percentage curve. This can be imagined as the shape of the frequency polygon when the number of subjects as extremely large and the width of intervals extremely small.

iii) ***Cumulative Percentage Curve or Ogive***

When the frequencies are expressed as cumulative percentage of N on the vertical axis, the graphic representation is known as a cumulative percentage curve or ogive. After finding the cumulative percentage frequencies, the points are plotted on the exact upper limits of the class-intervals. A curve joining the points thus obtained is called the cumulative percentage curve or ogive.

The cumulative percentage curve or ogive of the distribution represented in Table 10.9 is illustrated in Figure 10.3.

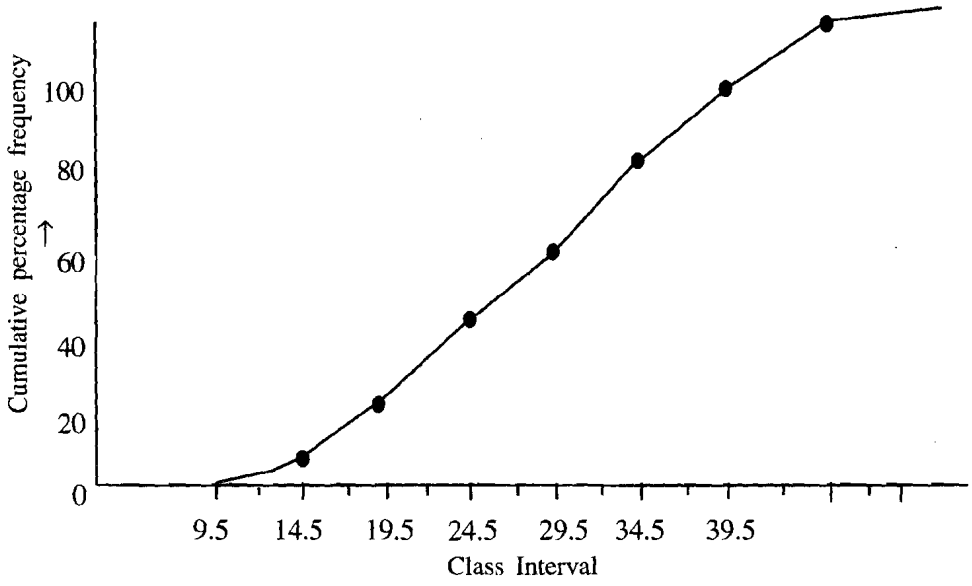

Figure 10.3: Cumulative percentage curve or ogive plotted from the data of Table 10.9

Now that we are familiar with the frequency distribution graphs, it is important to note that the basic information provided by histogram, polygon or curve is the nature of the distribution of the subjects over various values of the variable, *i.e.*, whether they are evenly distributed or that there is a concentration around some value(s), and whether the values are widely scattered or are compact. Thus, these figures are more of an exercise in data exploration than in analysis of data.

Besides the histogram, a variant of histogram is the stem-and-the leaf plot. Another way of representing the frequency distribution is through the Box-and-whiskers plot. Let us get to know about these graphs.

iv) ***Stem-and-Leaf Plot***

As already mentioned above the stem-and-leaf plot is a histogram-style tabulation of data. Consider the data set presented in Table 10.1. sort the data in the ascending order (i.e. starting from 57,60,61,63.....95,97,98).

A stem-and-leaf plot of this data can be constructed by writing the first digits in the first column (under stem and leaf), then writing the second digits of all the numbers in that range to the right as shown in Figure 10.4.

Frequency	Stem	Leaf
1	5	7
7	6	0123477
12	7	000235667888
15	8	000002224555778
5	9	03578

Figure 10.4: Stem and leaf plot for scores of 40 students of the M.Sc. (DFSM) programme

The result is a histogram turned on its side, constructed from the digits of the data as you may have seen in Figure 10.4. The first column gives the frequency and the actual stem-and-leaf plot of the subjects over various values of the variable is in the second column. The first digit in each value under the column stem-and-leaf is considered the 'stem' and the second digit 'leaf'. The term 'stem-and-leaf' is used to describe the diagram since it resembles the right half of a leaf, with the stem at the left and the outline of the edge of the leaf on the right. This type of plot indicated whether the subjects are evenly distributed or that there is a concentration around some value(s), and whether the value(s) are widely scattered or are compact. A stem-and-leaf plot therefore shows the shape and distribution of data. It can be clearly seen in the Figure 10.4 above that the data clusters around the row with a stem of 7 and 8.

Thus it must be evident that the stem-and-leaf plot provides essentially the same information as the histogram, with the following added benefit that the plot displays not only the frequency for each interval, but also displays all of the individual values within that interval and the data is arranged compactly since the stem is not repeated in multiple data points. However, remember, stem-and-leaf plot figures are more of an exercise in data exploration than in analysis of data.

Next, let us learn about the box-and-whiskers plot, which indicates primarily how a data distributes about the median.

v) ***Box-and-Whiskers Plot***

A box-and-whiskers plot, sometimes referred to as box plot is also a histogram-like method of displaying data. This statistic is used to help understand the distribution of the data in terms of percentile. The median (middle value of the data) is called the 50th percentile, which means that 50% of the data are below the median and 50% are above the median. In the same way the 25th percentile is that number where 25% of the data are below that number and 75% are above. The 75th percentile is similar. Fifty per cent of the data lie between the 25th and 75th percentiles. Another statistical measure of position is the *quartile*. A *quartile* divides the distribution into quarters. The quartiles denoted by Q1, Q 2 and Q 3 in Figure 10.5 are the three numbers that occupy the 25th, 50th, and 75th percentiles, respectively.

Let us understand this with the help of an example. Suppose we have the following numbers (already ranked):

(1, 3, 4, 5, 5, 6, 7, 7, 7, 8, 8, 10, 10, 11 and 19)

Using this data now let us develop the box plot. Given herewith are the steps to draw a boxplot.

1. First identify the median and the 25th and 75th percentile value from the data.
2. Draw a *box* from the *25th to the 75th* percentile.
3. Split the box with a *line at the median.*
4. Draw a thin line (whisker) from the *75th percentile up to the maximum* value.
5. Draw another thin line from the *25th percentile down to the minimum* value.

Now tally your plot with the display of the whiskers and box plot representing the data given herewith.

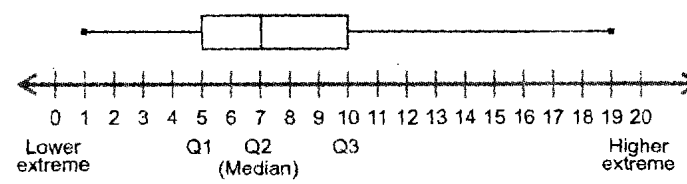

Figure 10.5: Developing the box and whiskers plot

So what did you infer from this plot? The box plot conveys the location of the middle half of the data (i.e. value 7 which is the median), the dispersion, and the skewness. The location of the median shows the central tendency. The rectangle (box) reveals the middle half of the data (i.e. 50% of the data lies here). The reach of the whiskers (vertical lines connecting the extreme points to the box) exposes the range, and the non-symmetry or symmetry of box and whiskers displays the skewness.

A box chart of the lung cancer study statistics presenting the relationship between smoking and mortality is presented in Figure 10.6 for your information.

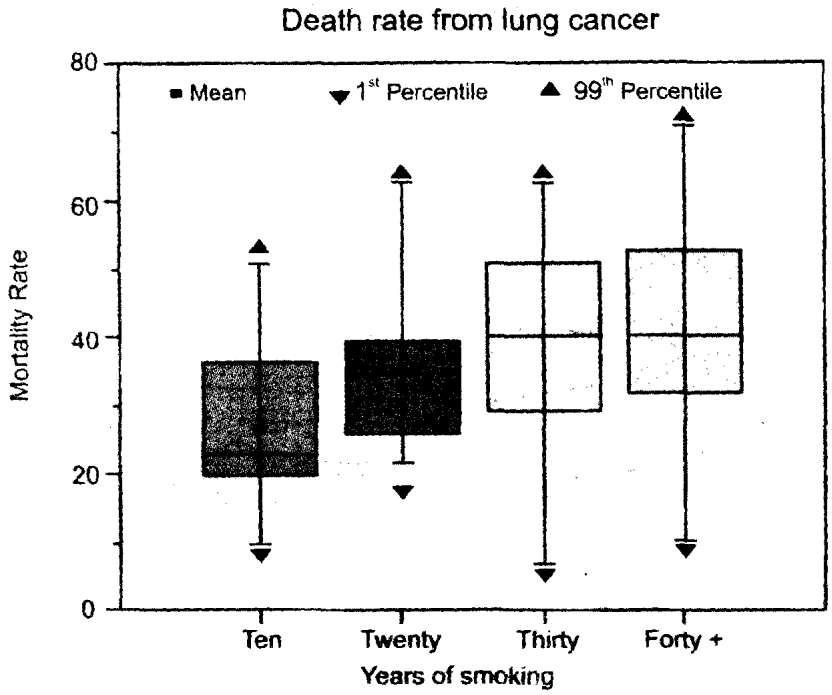

Figure 10.6: Box and whiskers chart of lung cancer study statistics

Box plots are helpful in quality analysis for interpreting the distribution of data since it can easily show whether the data is skewed and if there are unusual observations (outliers) in the dataset. Box plots are also very useful when large numbers of observations are involved and when two or more datasets are being compared.

With this we end our study of how to represent frequency distribution through the use of various graphs. Next, in sub-section 10.6.2 we shall focus on graphs we can use for presenting the nominal and ordinal data.

10.5.2 Graphs for Nominal and Ordinal Data

In the sub-section above we have learnt about the histogram, polygram and frequency curve as means to display frequency distribution of a variable. Usually these display the frequencies on a metric scale. Various other forms of graphs we can also draw for other kind of data, preferably on nominal and ordinal scale though can be metric also. We shall get to know about these important graphs namely the pie diagram and the bar diagram, next.

i) *Pie Diagram*

Refer to Figure 10.7, which depicts a pie chart. The pie diagram is based on the data on the magnitude/severity of anaemia reported in a study among children 9 - 36 months of age.

As is evident from the figure, a pie is a circular diagram divided into segments, each segment representing frequency in a category. It is important that the categories must be mutually exclusive and exhaustive. To illustrate, the pie in Figure 10.7a, illustrate that 38.2% children from the total were suffering from moderate anaemia,7.8% from severe anaemia and so on.

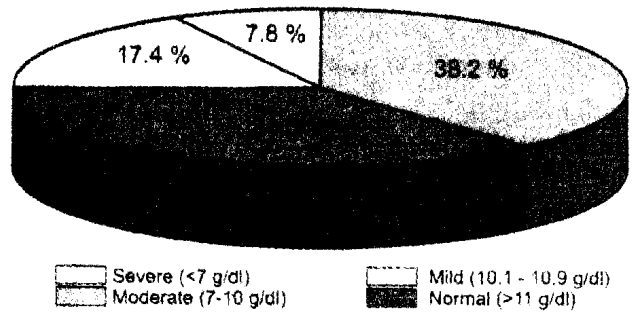

Figure 10.7a: Pie chart for magnitude/severity of anaemia among children 9 - 36 months of age (n = 523)

Source: (Kapur et al. Iron Status in Children Aged 9-36 months in an Urban Slum Integrated Child Development Services Project in Delhi, Indian Pediatr 2002;39:136-144).

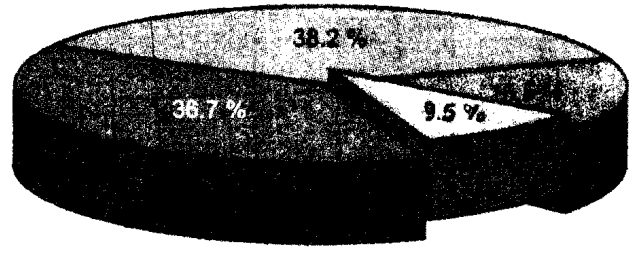

Figure 10.7b: Wedged pie chart

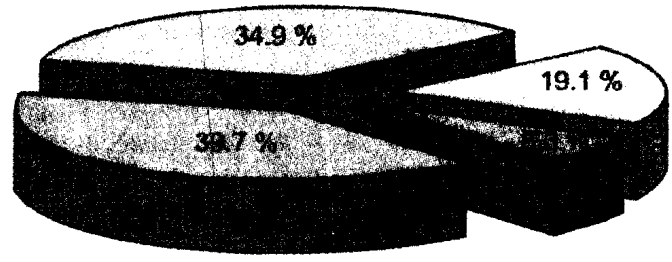

Figure 10.7c: Exploded pie chart

Generally, the largest components sector of pie diagram is placed beginning at 12 O'clock position on the circle. The other component sectors are placed in clockwise sequence in descending order to magnitude except for the component, "Miscellaneous", if any is shown last. Each component should be shaded or coloured to contrast with adjacent sector, whenever possible. When the attention is to be specially drawn to one particular category, the segment representing this category is wedged out (as illustrated in Figure 10.7b) and this feature is termed *wedging*. A chart with one or more sectors separated from the rest of the disk is called an *exploded pie chart* as shown in Figure 10.7c.

Pie charts have several advantages. They are both an informative way to depict proportional statistical information and allow for easy comparisons. Limitations of pie chart would include that they can only depict up to 8 slices without causing confusion. Anymore than that could mislead readers and not convey the message we are trying to get across. Also, categories that represent less than 5% of the whole cannot be easily distinguished in the graph.

Next, let us get to know about the bar graphs.

ii) *Bar Graphs*

Bar graphs are specially suitable for depicting data presented in nominal or ordinal categories. In fact, bar graphs are an excellent way to depict the data in the form of mean, rate or ratio from a cross-sectional study. What is a bas graph? The bar graph consists of parallel, usually vertical bars or rectangles with lengths proportional to the frequency with which specified quantities occur in a set of data. Refer to Figure 10.8 which illustrates a bar graph, which depicts data related to distribution of the haemoglobin concentration in children (9 - 36 months of age). Yes, you would have noticed that the bar graph resembles histogram. But, it differs from a histogram in the sense that the the bars are separated by spaces. Bar charts help in making the categories stand out from one another, therefore making it easier to compare each category.

While reviewing research articles you may have also come across *divided bar diagrams* as the one illustrated in Figure 10.9. In this graph the divided bar depicts the age of introduction of solid foods among children. Divided bar graphs are a variation that, similar to pie charts, show proportional relationships between data within each bar. In addition, divided bar graphs can show changes over time.

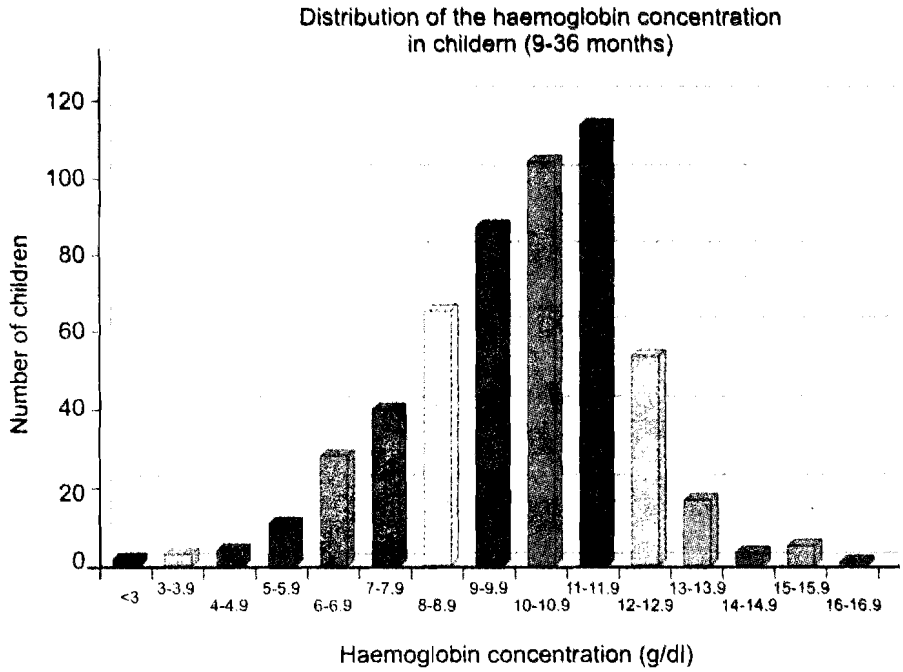

Figure 10.8: Bar graph

Source: Kapur et al. Indian Pediatr, 2002, 39:136:144 (Note: Bar graph prepared based on the data obtained from the above reference).

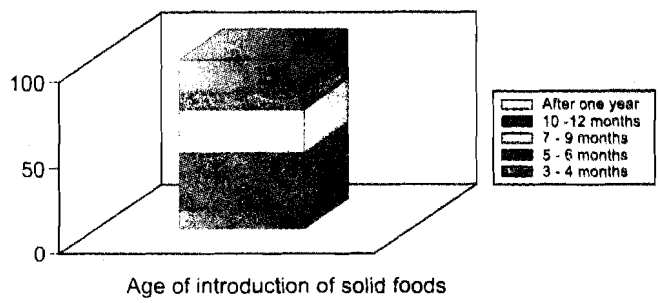

Figure 10.9: Divided bar graph

Source: Kapur et al. Unpublished data.

In our discussion so far we have looked at graphical representation of data for frequency distribution, as well as, for nominal and ordinal data. Very often you would also be requird to present data depicting the relationship between two or more variables under study. The next sub-section will focus on this aspect.

10.5.3 Graphs for Relation between Two Variables

Scatter diagrams and line diagrams are the two graphic representations which are used to study possible relationships between two variables. Let us review these two graphs.

i) *Scatter graph*

As the name suggests, scatter graphs gives information regarding the relationship between two variables when the points are scattered around the graph. A strong relationship between the two variables is observed when most of the points fall along an imaginary straight line with either a positive or negative slope. No relationship between the two variables is observed when the points are randomly scattered about the graph. Thus scatter graph is a useful tool for identifying a potential relationship between two variables.

How is the scatter graph composed? It is composed of a horizontal axis containing the measured values of one variable and a vertical axis representing the measurements of the other variable. Note it must always be a metric variable in all scatter diagrams. Look at Figure 10.10 which illustrates a scatter diagram depicting frequency distribution of haemoglobin in children (9 - 36 months of age) as compared to standard. As is evident from the scatter diagram, the population under study had an excess of children with low haemoglobin values (with majority of the children below the 3rd centile) as compared to standards.

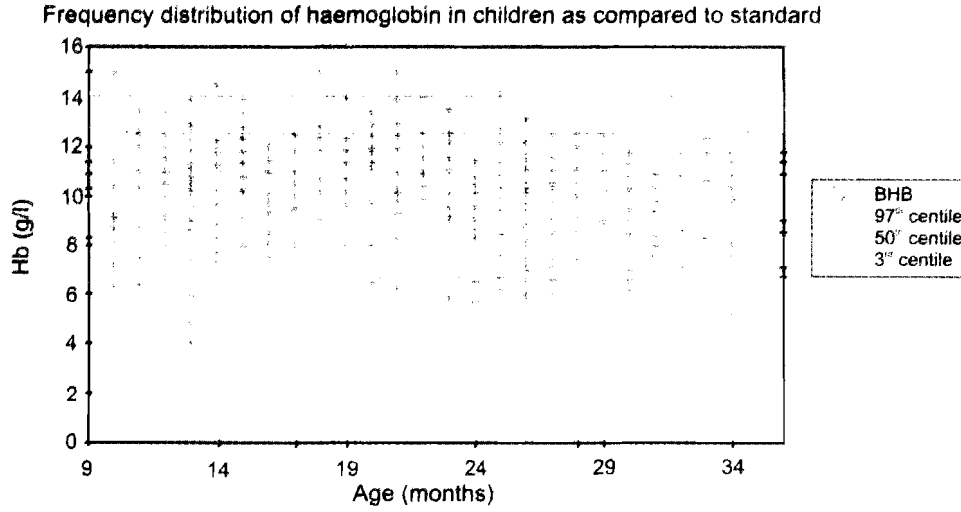

Figure 10.10: Scatter diagram for frequency distribution of haemoglobin among children (9 - 36 months of age) as compared to standards

Source: Kapur et al. Indian Pediatric, 2002, 39:36-144.

(Note: Scatter plot prepared based on the data obtained from the above reference).

Very often the scatter diagram is used to display what happens to one variable when another variable is changed. The diagram is used to test a theory that the two variables are related.

Next, let us focus on the line diagram.

ii) *Line diagram*

All of you must have seen a line chart at some point in your life. Certainly you may remember working with the child growth charts about which we have studied in the Advance Nutrition Course (MFN-004). A growth chart, you may recall, is constructed

by plotting different percentile points of weight-for-age of healthy children in a population (refer to Figure 14.1 in Unit 14 of the Course MFN-004). So you may have noticed that a growth chart, in fact, is a line graph, conventionally called a chart, which is mainly used for longitudinal monitoring of the child over a period of time. Here the trend of the childs growth is monitored and it is important to note that the trend should follow the same pattern as the reference curve. Any drop or faltering is a sign of lack of thriving.

A line graph now then actually plots points and then connects them with a line. It is used to show trend of one variable over the other. Refer to Figure 10.11, which shows the trend of mean haemoglobin among parasitic infested and non-infested children. As you must have noticed two lines are drawn one for infested and the other for non-infested. The plotted points are joined by a line.

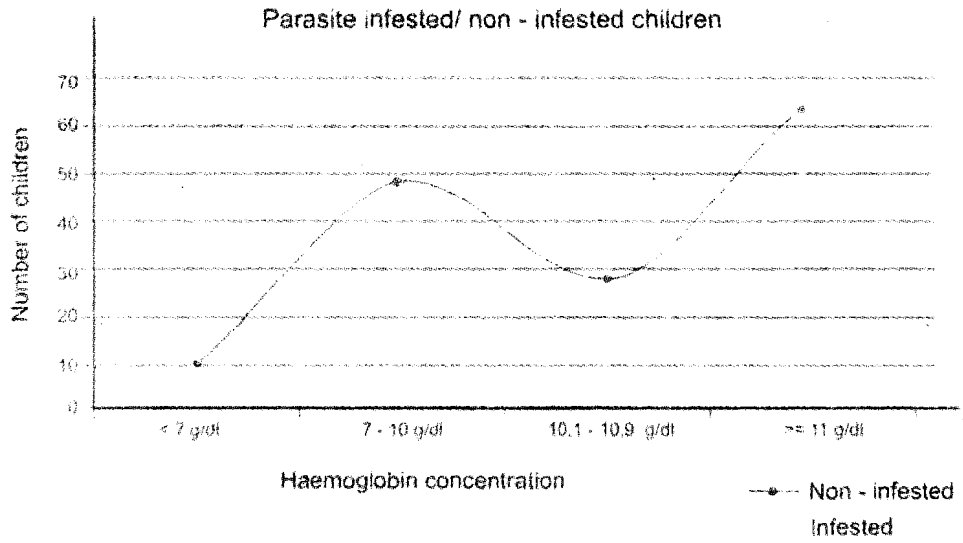

Figure 10.11: Line diagram showing the trend of mean haemoglobin among parasitic infested and non-infested children

Besides the graphical representations presented in this section, there may be some other special diagrams such as the *dendrogram*, *pedigree charts*, *epidemic curves* which may be used to present data in nutritional and health related research. For details related to these you may refer to any statistic text book, here we shall not enumerate further. However we would like to end this section by highlighting few cautions which must be exercised in visual display of data.

Cautions in Visual Display of Data

A diagram should not be too complex.

Restrict the number of relationships that can be shown in one diagram to not more than two relationships, or not more than three variables in one diagram.

We may make diagrams choosing the appropriate scale to show steeper or flatter relationships among variables.

Use graphic diagrams mostly for exploring the data than to draw conclusions. Why? Since the sample size is important for conclusions, and since small or big sample size does not affect the diagram, the conclusion based on the diagram can be very fallacious.

Check Your Progress Exercise 2

1) What is a histogram? How does it differ from a bar chart.

.....

.....

2) List the various types of graph you can use for representing the frequency distribution of a data.

.....

.....

.....

3) Enumerate the graphs you would prepare for nominal or ordinal data.

.....

.....

.....

4) Relationship between the variables can be depicted effectively through what kind of graphs? Explain their advantages briefly.

.....

.....

.....

10.6 QUALITATIVE DATA

It has already been explained that quantitative measurement makes use of tools that provide a standardized framework in order to limit data collection to certain predetermined responses or categories. The variables that describe a phenomenon are fit in the standardized categories to which numerical values are attached. But in some situations it is difficult to analyze a phenomenon into various components or variables which can be measured in quantified terms. In such cases the researcher takes into consideration the phenomenon as a whole and assumes that there is some quality in the phenomenon in its entirety. When the researcher attempts to retain the totality of a phenomenon while verifying propositions regarding it, he/she adopts a qualitative approach. While using this approach the researcher seeks to capture what people have to say in their own words. Qualitative approach describes the experiences of people in depth and permits the researcher to record and understand people in their own perceptions.

Qualitative data, we learnt, consists of 'detailed descriptions' of situations, events, people, interactions, and observed behaviours. These data are also available in the form of 'direct quotations' from people about their experiences, attitudes, beliefs, and thoughts. The verbal data gathered through questionnaires, observations and interviews are mostly qualitative in nature. The 'excerpts' or 'entire passages' from documents, correspondence records and case histories are also examples of data of a qualitative nature. It may be noted that detailed descriptions, direct quotations, and case documentation of a qualitative nature are raw data from empirical situations.

10.6.1 Organization of Qualitative Data

The responses to open-ended questions on a questionnaire are pretty extensive; they are neither systematic nor standardized. However, they permit the researcher to understand situations as seen and felt by the respondent. The data gathered through participant observation or an open-ended/unstructured interview are also descriptive in nature. These descriptions may be in the form of field notes specifying some basic information pertaining to the place where the observation has taken place, as well as, descriptions about the people who participated in the activities and their extrinsic

behaviour in the course of the activities. However, it is not possible to interpret minds while observing their extrinsic behaviour. Through an open-ended/unstructured interview, we can know more about those events which had occurred earlier or could not be observed during participant observation. It provides a framework within which the researcher should be able to gather information from people conveniently and accurately. The information may pertain to a programme, the reaction of participants about the programme and the type of change the participants perceive in themselves after their involvement in the programme. The data are mostly in the form of responses to structured and unstructured questions put to respondents by the researcher during conversation. The responses are generally direct quotations from respondents in their own words and provide details about the situations, events, people, experiences, behaviours, values, customs, etc.

The qualitative data gathered using open-ended questionnaires, participant observations and in-depth interviews are voluminous. They need to be organized and classified into specific patterns categories and descriptive units to avoid any chaos. However, before any such classification is done, it is advisable to make some copies of the data. Once copy may be stored in a safety deposit box so that in case of a loss, this copy can be stored by the researcher. The third copy may be used for further treatment of the data throughout. The third copy may be used to fill the missing gaps, identified during scrutiny by the researcher. Additional notes can also be recorded in this copy. Since the organization of qualitative data involves a lot of cutting and pasting, a fourth copy may be used for that purpose.

Actual classification or organization of the data can begin only after the copies are made. There are no formal or universal rules for organizing the data in various units, patterns, or categories. It requires a creative approach and a lot of perseverance to give a meaningful look to the data. The contents of field notes about interviews or observations may be read carefully by the researcher and he/she may note down his/her comments on the margins or attach small pieces of paper with his/her written comments/notes using staples or tags. The arrangement of data in topics, using abbreviations, is the next step. The abbreviated topics are written down either on the margins of the relevant data or on slips of paper which may be attached to the relevant pages. The process of classifying or labeling various kinds of data help in the preparation of a *data index*. Sometimes there are large data. In such situation, computers are helpful in developing systematic and comprehensive classification schemes using code numbers of different categories and sub-categories as already discussed earlier in section 10.4. The computerized classification system permits the use of organized data by several groups of people over a long period of time. It permits easy cross-classification and cross-comparison of descriptive narrations for complex analysis.

Check Your Progress Exercise 3

1) Define qualitative data. Give some examples of these data.

.....

.....

.....

.....

.....

.....

.....

10.7 LET US SUM UP

In this unit, we discussed the nature of quantitative and qualitative data, the various methods of representing the quantified data graphically, and the qualitative data. The main points are as follows:

1) The data collected through the administration of various tools on the selected samples are of (i) quantitative and (ii) qualitative nature.

2) Quantitative data are expressed in nominal, ordinal, interval or ratio scales of measurement. These data are classified into two categories: (i) parametric and (ii) non-parametric. The parametric data are obtained by applying interval or ratio scales of measurement, whereas non-parametric data are either enumerated or ranked. In the enumerated data we make use of nominal scale and in the ranked one we apply ordinal scale.

3) The quantified data is tabulated in 'frequency distribution' and can be represented graphically with the help of a histogram, a frequency polygon, an ogive and stem and leaf plot and the box plot.

4) Graphs for representing nominal and ordinal data include the pie chart and the bar diagram.

5) Scatter diagram and the line diagram are the two graphs for studying the relations between two values.

6) If is some situations if it is difficult to measure or analyze a phenomenon into various components or variables in quantified terms, the researcher takes into consideration the phenomenon as a whole, in detail and depth. In other words, the researcher uses qualitative techniques of analysis. Qualitative data consist of detailed descriptions of situations, events, people, interactions, and observed behaviours. These data are also available in the form of direct quotations from people about their experiences, attitudes, beliefs, and thoughts. The excerpts or entire passages from documents, correspondence, records and case studies are also examples of qualitative data.

10.8 GLOSSARY

Quantitative Data : data which are expressed in nominal, ordinal, interval or ratio scales of measurement.

Qualitative Data : data which are available in the form of detailed descriptions of situations, events, people, interactions, and observed behaviour, direct quotations from people about their experiences, attitudes, beliefs, and thoughts, and excerpts from documents, correspondence, records and case histories.

Parametric Data : these are data which are got by applying interval or ratio scales of measurement.

Non-parametric Data : these are data which are got by applying nominal or ordinal scales of measurement. These types of data are either counted or ranked.

Median : the middle value in a distribution or set of ranked values; the point that divides the group into two equal parts.

Percentile Rank : the expression of an obtained test score in terms of its position within a group of 100 scores.

10.9 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress Exercise 1

1) Quantitative data is the description of an empirical event or phenomenon in a numerical system presented with the help of different scales of measurement such as nominal, ordinal, interval and ratio. The two major types of quantitative data are: parametric (obtained through interval or ratio scales) and non-parametric (counted by a nominal scale or ranked by a ordinal scale).

2)

Class Interval	Mid Point	f
159 - 199	197	1
190 - 194	192	2
185 - 189	187	4
180 - 184	182	5
175 - 179	177	8
170 - 174	172	10
165 - 169	167	6
160 - 164	162	4
155 - 159	157	4
150 - 154	152	2
145 - 149	147	3
140 - 144	142	1

3) Contingency table is a label of frequencies that show the observed frequencies of data elements classified according to two or more variables. Contingency tables are labels used to record and analyze the relationship between two or more values, usually categorical values.

Check Your Progress Exercise 2

1) A histogram is a graph in which class-intervals are represented along the horizontal axis and their corresponding frequencies are represented by rectangular bars on the interval. It differs from the bar graph in the sense that in the bar graph the bars are separated by spaces.

2) The various types of graphs used for representing frequency distribution include the histogram polygon, ogive, box plot and the stem and leaf graph.

3) Noninal and ordinal data are best represented by pie chart and bar diagram. Look sub-section 10.6.2 and enumerate on these graphs.

4) Scatter plot and line diagrams are best suited to depict the relationship between two values. Look up their advantages in sub-section 10.6.3 and answer in your own words.

Check Your Progress Exercise 3

1) Qualitative data describes a phenomenon which cannot be measured or quantified. The phenomenon is looked at in its totality. Detailed descriptions of situations, events, people, interactions, and observed behaviours constitute qualitative data.