CHAPTER 2

SUMMARIZING DATA

2.1 INTRODUCTION

In recent years, collection of statistical data has grown at such a rate that it would be impossible to comprehend this large volume of data unless they are presented and disseminated in a more convenient or easily interpretable form. The whole matter of putting these large masses of data into usable form has always been important, and it has multiplied greatly in the last few decades. This has been due, partly to the development and advancement of the modern computer and other electronic technologies, which have made it possible to accomplish in minutes what previously had to be done in months or years. Statisticians and scientists from all other disciplines use various methods to present these data in an organized, systematic and orderly manner. This chapter is devoted to discuss how the data can be summarized and presented using available methods.

2.2 MEANING OF DATA

Data are the observations or chance outcomes that occur in a planned experiments or scientific investigations. They are the raw materials of statistics and for all statistical purposes, we may define data as numbers whose common characteristic is variability or variation. The investigators generate data through some process of counting, observation or measurement. A statistical investigation may be conducted, for example,

among the male workers of an industry to know whether the workers smoke or not. The answer may be recorded as **'yes'** for those who smoke and **'no'** for those who do not smoke. Thus, all the workers of the industry may be classified into two categories: smokers and non-smokers. The number of smokers and non-smokers are numerical data, obtained through the process of **counting**. We may further attempt to **record** their ages or **measure** their height and thus obtain some numerical data on age and height. Some information may be obtained simply by **observing** whether a particular event occurs or does not occur. For example, we may observe whether a given day is rainy or sunny, a man has blue eyes or brown eyes. All these information constitute data.

All statistical data may be broadly classified into two broad categories, qualitative and quantitative. **Qualitative data** are generated by assigning observations into various independent categories and then counting the frequency of the occurrences within these categories. In effect, it is enumeration data, such as counting how many persons in a community are Muslims, and how many of them are of other religions. Or it might be categorizing them on the basis of their sex and then noting the frequencies or numbers of persons falling within the categories. Clearly the qualitative data are those which can be stated or expressed in qualitative terms. **Quantitative data**, on the other hand, are those which can be measured in quantitative units. Here we are able to measure or note the actual magnitude of some characteristics for each of the individuals or units under consideration. Measurement of height, weight, income, temperature, family size or the number of street accidents over a specified period will all result in quantitative data.

2.3 LEVEL OF MEASUREMENT

Statistical data, whether qualitative or quantitative, are generated or obtained through some measurement or observational processes. Measurement is essentially the task of assigning numbers to observations according to certain rules. The way in which the numbers are assigned to observations determines the scale of measurement being used. The rule chosen for the assignment process, then, is the key to which measurement scale is being used.

There are four levels of measurement. They are (a) Nominal level (b) Ordinal level (c) Interval level and (d) Ratio level.

Each type of measurement has unique characteristics and implications for the type of statistical procedures that can be used with it. We elaborate below these concepts in greater details.

2.3.1 Nominal Level of Measurement

All qualitative measurements are nominal, regardless of whether the categories are designated by names (red, white, male) or numerals (June 20, Room no. 10, bank account no., ID no. etc.). In nominal level of measurement, the categories differ from one another only in names. In other words, one category of a characteristic is not necessarily higher or lower, greater or smaller than the other category. Sex (male and female), religion (Muslim, Hindu, Christians etc) are the examples of nominal level of measurement. What one must ensure in nominal level of measurement is that the categories must be (a) homogeneous, (b) mutually exclusive, and (c) exhaustive. The nominal level of measurement gives rise to **nominal data**.

To work with such nominal data with statistical tools, we need to impose a numerical scheme on the data. For example, with gender, 0 might be assigned to males and 1 to females. With religion, the scheme might be to use 1 for Muslims, 2 for Hindus, 3 for Christians etc. In each of these cases, the numerical data have been artificially created, but none of the numbers have any numerical meaning. In measurement scale, nominal level of measurement is the lowest level of measurement.

2.3.2 Ordinal Level of Measurement

When there is an ordered relationship among the categories, we achieve what we refer to as the ordinal level of measurement. Unlike the nominal level, here we have the typical relations "higher", "more than" "less difficult", "more prejudiced", "more feminine" less favorable", more profitable", "less costly" and so on. More specifically, the relationships are expressed in terms of the algebra of inequalities: a is less than b ($a < b$) or is greater than b ($a > b$). Thus, the academic degrees (MA. BA, etc.), official designation (manager, deputy manager, accountant), socio-economic status (high, medium low), academic performance (outstanding, very good, good, poor), monthly frequency of visits of a physician in a clinic (frequently, occasionally, rarely, never), level of agreement among the common people on the issue of imposing VAT in food items (strongly agree, agree, disagree, strongly disagree) all belong to nominal level of measurement. Ordinal level of measurement of any characteristic gives rise

to **ordinal data** and ordering is the sole mathematical property applicable to ordinal data. Note that an ordinal scale is distinguished from a nominal scale by the **additional** property of **order** among the categories included on the scale. You can rate, for example, the level of agreement on the issue of VAT on a 4-point scale of 1 (strongly agree) to 4 (strongly disagree). Such ratings, however, have no real significance in the sense of usual arithmetic operations, but they certainly represent a way to introduce an ordering relation.

The chief properties of ordinal level of measurement are

(a) The categories are distinct, mutually exclusive and exhaustive
(b) The categories are possible to be ranked or ordered
(c) The distance or differences from one category to the other category is not necessarily constant.

2.3.3 Interval Level of Measurement

The interval level of measurement includes all the properties of the nominal and the ordinal level but an additional property that the difference (interval) between values is known and of constant size. A thermometer, for example, measures temperature in degrees, which are of the same size at any point in the scale. The difference between 20^0C and 21^0C is the same as the difference between 12^0C and 13^0C. The temperature 12^0C, 13^0C, 20^0C and 21^0C can be ranked and the differences between the temperatures can easily be determined. It is also important to note that 0 is just an arbitrary point on the scale. It does not necessarily represent the absence of heat, just that it is cold. In fact, 0 degrees Celsius is 32 degrees on the Fahrenheit scale. Owing to this, we cannot say that a temperature of 64^0F is twice as warm as a temperature of 32^0F. It is because of the fact that the Celsius equivalence of 32^0F (the freezing point of water) is 0^0C, while the equivalence of 64^0F is

$$\left(\frac{5}{9}\right)(64-32) = 17.8^0\text{C}$$

Obviously, 17.8^0 C is not twice as warm as 0^0C.

The Gregorian calendar is another example of an interval scale: 0 is used to separate B.C. and A.D. We refer to the years before 0 as B.C. and to those after 0 as A.D. Incidentally, 0 is a hypothetical date in the Gregorian calendar because there never was a year 0. The other examples are IQ,

calendar time (6 AM, 10 AM etc). The interval level measurement has the following properties:

(a) The data classification are mutually exclusive and exhaustive
(b) The data can be meaningfully ranked or ordered
(c) The difference between one data-classification to the next is known and constant.

A set of data in which we can form differences of measurements, but cannot multiply or divide, is known as **interval data**.

2.3.4 Ratio Level of Measurement

In practice, all quantitative data fall under the ratio level of measurement. It has all the ordering and distance properties of interval level. In addition, a 'zero point' can be meaningfully designated and thus ratio between two numbers is also meaningful. Examples of ratio level of measurement include wages, stock prices, sales values, age, weight and height. Thus it makes sense to speak of 0 sales, when there is no sales in the store. It is also quite meaningful to say that a 4-feet tall boy is twice as tall as a 2–feet tall boy. A family with 6 members is twice as large as of a family with 3 members.

In comparing the four levels of measurement, we can conclude that an ordinal measure is a nominal measure, and in addition, has the ordinality property, an interval measure is an ordinal measure plus it has a unit of measurement, and the ratio measure has all the properties of nominal, ordinal and interval measures, plus it has an absolute zero. This cumulative nature of the measures shows that a higher level of measure can be used as a lower level of measure, but the converse is not true. Thus, an interval variable, for example, can always be used as nominal or ordinal variable, but neither a nominal variable nor an ordinal variable can be used as an interval variable. The characteristic properties of the various levels of measurement are compared in the table shown in the next page:

For a quick and easy understanding of the characteristic properties of the four levels of measurement and for deciding whether a particular level of measurement qualifies as nominal, ordinal, interval or ratio, the following flow chart may be used:

- Do the numbers express a quantitative value or order?
  If no, then → nominal level

If yes, then ask:

- Do the differences between the numbers represent equal units of measurement (e.g. 3–2=4–3)?

If no, then → ordinal level

If yes, then ask:

- Does the measurement have an absolute zero?

If no, then → interval level

If yes, then → ratio level.

Table comparing the different levels of measurement

Scales	Characteristics	Examples
Nominal	Categories are homogeneous, mutually exclusive, and no assumptions about ordered relationships between categories made	▪ Sex of subject ▪ Eye color ▪ Religion ▪ Political affiliation ▪ Place of residence ▪ Room numbers etc
Ordinal	All of the above plus the categories can be rank-ordered	▪ Examination grade ▪ Health status ▪ Level of education ▪ Rank in job
Interval	All of the above plus exact differences between categories are specified and an arbitrary zero point is assumed	▪ Temperature ▪ IQ test score ▪ Calendar time
Ratio	All of the above with the exception that a true zero point is assumed	▪ Height ▪ Weight ▪ Fat consumed ▪ Wage

2.4 VARIABLE AND ATTRIBUTE

Variable

Throughout our discussion in this text, the word variable will keep appearing. The term **variable** refers to a characteristic or property of an object or individual that can be measured or observed to vary.

***Definition* 2.1** : *A variable is a characteristic or property, often but not always quantitatively measured, containing two or more values or categories that can vary from one individual to another.*

Religion, for example, is a characteristic of an individual person, which differs from one person to another and thus is a variable. Since religion is a

qualitative characteristic, it is referred to as **qualitative variable** and the resulting data are qualitative data. Religion has several non-overlapping categories, such as Muslim, Hindu, Buddhist and others. For this categorical nature, qualitative variables are also sometimes referred to as **categorical variables.** Further, it falls under nominal level of measurement, for which it is also called **nominal variable.**

***Definition* 2.2:** *A qualitative variable is a characteristic that is not capable of being measured but can be categorized to possess or not to possess some characteristics.*

A few examples qualitative variable are

- Color of a garment (red, white, etc.).
- Bank account type (savings, current, fixed).
- Place of birth (rural, urban, sub-urban etc.),
- Sex (male, female).
- Frequency of visits (frequent, occasional, rare, never).
- Examination grade (A, B, C).

Frequency of visits and examination grades also qualify as **ordinal variables,** since we can rank them in order of their magnitude.

***Definition* 2.3:** *A quantitative variable is one for which the resulting observations are numeric and thus possesses a natural ordering.*

When data refer to a quantitative characteristic, we achieve what we referred to as **quantitative data**. Thus age is a quantitative variable, because it is possible to express the differences between individuals on a quantitative scale of measurement. Other examples of quantitative variables are

- Sales volume in a department store,
- Years of teaching experience of an individual
- Income of individuals
- Longevity of lives
- Day temperature.

A quantitative variable may be either discrete or continuous. We can define a discrete variable as follows:

***Definition* 2.4:** *A variable can take on only values at isolated points along a scale of values, is called a discrete variable.*

Data for a discrete variable typically occur through the process of counting. They have equality of counting units as their basic characteristics. Mathematical operations, such as addition, subtractions, multiplication and division are meaningfully permissible with discrete variables. Discrete data thus represent both interval and ratio levels of measurement. Examples of discrete variables are:

- Family size
- Number of days absent from work for illness
- Number of shares in a business
- Number of automobiles imported during 1980–1990
- Number of units of an item in an inventory
- Number of assembled components found to be defective
- Number of typing errors in a document.

Do not assume, however, that discrete variable necessarily involves only whole numbers. But most of the discrete variables used by social scientists are expressed in terms of whole numbers

***Definition* 2.5:** *A continuous variable is one that may take on infinite number of intermediate values along a specified interval.*

Thus if the variable is height measured in inches, then 4 inches and 5 inches would be two adjacent values on the scale, between which an infinite number of values are possible: 4.5, 4.7, 4.78 etc. Some more examples of continuous variables are:

- Payoffs in business
- Waiting time in a bank counter
- Hourly average payment of factory workers
- Rainfall in millimeter recorded by meteorological office
- Height or weight of individuals.

Attribute

The distinct categories of the qualitative variable are sometimes called **attributes**. If one simply notes down for each individual whether he/she possesses or does not possess certain characteristic–a mobile set, smoking habit, or an opinion on certain political issue–this characteristic may be called an **attribute**. Quantification thus lies in counting how many individuals possess this attribute and how many do not and the proportion or percentage with this attribute provides a useful description of the population.

Constant

A variable is contrasted with a **constant**, the value of which never changes. For example, $\pi = 3.1416$, 1 foot=12 inches, e=2.718, the velocity of light=186,300 miles per second, total angle of a circle=360^0 are all constants.

2.5 SUMMARIZING AND PRESENTING DATA

A set of data even if modest in size, is often difficult to comprehend and interpret directly in the form in which it is collected. Suppose a sample of 50 workers was drawn from a business enterprise, which employed 500 workers. The researcher collected such data as the workers' age, level of education, wage, and their religion by directly interviewing the workers. These are some of the personal characteristics of the workers which the researcher needs to meet the objectives of a social research. Clearly, the information collected contains both qualitative and quantitative data. The table (Table 2.1) shown in page 30 displays these data. Having obtained the data, the most usual questions one might ask now:

(a) How many of the workers are below 30 years of age? Over 50?
(b) How many of them earn between 74 and 81 taka?
(c) How many of them have secondary level of education?
(d) Do most of the workers have large family size?
(e) How many workers belong to minority group?
(f) Are the workers frequent to remain absent from work?

The answers to the above questions can be given simply by counting the cases that appear in the table referred to above. But it will simply be a cumbersome job and sometimes impossible, if the number of cases is very large. What would then one expect us to do with this large volume of data?

Most of us would wish that someone had **classified**, **categorized** or **summarized** the data in a more convenient and readily interpretable form.

In this section, we shall discuss how this can be accomplished. Tabular and graphical procedures provide useful ways and means of organizing and describing the data such that they are more easily used and interpreted. The concept of frequency distribution is introduced here as a tabular method of summarizing data. We shall then show that this frequency distribution can also be displayed graphically employing a number of diagrams, charts, plots and curves.

Table 2.1: Raw data on background characteristics of workers

Worker	Wage	Age	Religion	Days absent	Family size	Education
1	93	25	Muslim	26	Small	Higher
2	66	29	Muslim	16	Large	None
3	93	32	Hindu	14	Small	Primary
4	69	39	Muslim	18	Medium	Primary
5	88	43	Christian	27	Large	Higher
6	76	40	Muslim	29	Medium	None
7	50	46	Muslim	23	Large	None
8	75	45	Muslim	33	Small	Higher
9	86	51	Christian	17	Large	Primary
10	97	37	Muslim	24	Medium	None
11	51	38	Muslim	17	Large	Primary
12	74	42	Muslim	18	Small	Primary
13	68	46	Muslim	21	Large	Higher
14	65	28	Muslim	11	Medium	Higher
15	89	30	Muslim	10	Medium	Primary
16	88	32	Muslim	12	Medium	Higher
17	77	36	Muslim	13	Large	None
18	87	37	Muslim	18	Large	Primary
19	85	41	Christian	15	Medium	Primary
20	84	35	Hindu	16	Medium	Higher
21	82	43	Muslim	22	Medium	Primary
22	83	44	Muslim	14	Small	Higher
23	82	42	Hindu	15	Small	Primary
24	81	42	Hindu	8	Medium	Higher
25	79	44	Muslim	9	Medium	Primary
26	80	47	Muslim	17	Small	Primary
27	65	46	Muslim	15	Large	None
28	74	36	Muslim	14	Small	None
29	69	37	Muslim	10	Medium	Higher
30	54	43	Muslim	11	Large	Primary
31	56	42	Muslim	10	Medium	None
32	73	45	Muslim	13	Large	Higher
33	75	34	Muslim	12	Medium	Primary
34	74	33	Muslim	17	Large	Higher
35	72	37	Christian	16	Medium	Primary
36	72	35	Muslim	20	Large	Primary
37	70	33	Hindu	19	Medium	None
38	63	36	Muslim	10	Large	Higher
39	70	38	Muslim	5	Medium	Higher
40	68	52	Muslim	12	Medium	Higher
41	59	54	Hindu	16	Medium	None
42	67	31	Muslim	18	Large	Primary
43	61	35	Hindu	21	Large	Higher
44	60	46	Muslim	19	Medium	Higher
45	56	44	Hindu	19	Medium	Primary
46	62	33	Muslim	15	Medium	Higher
47	72	36	Muslim	9	Small	Higher
48	73	38	Muslim	13	Small	Primary
49	71	32	Hindu	19	Medium	Higher
50	57	50	Christian	18	Medium	None

2.6 FREQUENCY DISTRIBUTION

Before we define what a frequency distribution is, it is necessary to define the term 'frequency'. **Frequency,** also called **class frequency,** refers to the number of observations or measurements falling within the confines of a particular class or category. For example, if an observation, say 5, is repeated 3 times (viz. 5, 5, 5), then we say that 3 is the frequency of 5.

***Definition* 2.6:** *A frequency distribution is a set of mutually exclusive classes or categories together with the frequency of occurrence of items, values or observations in each class or category in a given set of data, presented usually in a tabular form.*

A frequency distribution presents the data in a relatively compact form, provides a good overall picture, and contains information, which is adequate for many purposes. While a frequency distribution is presented usually in a tabular form, showing the frequency of measurements or observations in each of the several non-over-lapping classes, it may also be displayed graphically or by some statements, algebraic formula, or rules pairing a class of observations with its frequency.

A frequency distribution can be constructed for both qualitative and quantitative data. Quantitative data when grouped and organized in a frequency distribution result in a **grouped frequency distribution.** In contrast, for ungrouped data, every observed value of a variable is listed. This gives rise to what we call **ungrouped frequency distribution.**

We discuss below how a frequency distribution is constructed in practice.

2.6.1 Constructing Frequency Distribution for Qualitative Data

The construction of a qualitative distribution consists essentially of the following steps:

(a) Choose the category into which the data are to be grouped.
(b) Sort or tally the items or observation into appropriate categories.
(c) Count the number of items or observations falling in each category.
(d) Display the results in a table.

The resulting table represents the desired frequency distribution

Let us illustrate the construction of a frequency distribution by an example.

Example 2.1: Construct a frequency distribution for the family size data presented in Table 2.1

Solution: The data pertain to three distinct categories of the family size 'large', 'medium' and 'small'. Clearly, our data here are qualitative in nature and represent ordinal level of measurement. Our task now is to distribute the observations within the categories appropriately. To do this, we use a tally sheet and put one and only one tally mark for each item against each category simply by visual inspection. Then count the number of items falling in each category. The process follows the following steps:

- The first family in the order is 'small'. The category, 'small' appears in the first column of the table as a third entry. Put a tally mark against the family size 'small', which is simply a left-slashed off-diagonal stroke (/).
- Move on to the next entry, which is 'large'. Enter this again by a tally mark against the category 'large' appearing in the table.
- Repeat the above process until you have entered all the 50 items appearing in the observed set shown in Table 2.1.
- In the process of tallying, when you have completed four tallies in a category, put the fifth tally across the bunch of four by a diagonal slash to make a bunch of 5 tallies.
- Count the tallies for each category and put the number of tallies so counted in a tabular form.

The resulting tallies that appear below form our desired frequency distribution of family size shown in Table 2.2

Family size	Tally marks
Large	𝍸 𝍸 𝍸 /
Medium	𝍸 𝍸 𝍸 𝍸 ////
Small	𝍸 𝍸

Table 2.2: Frequency distribution of family size

Family size	Number of workers	Percent (%)
Large	16	32
Medium	24	48
Small	10	20
Total	50	100

The counts 16, 24, and 10 appearing in the second column of the table are the **class frequencies** for the categories large, medium and small respectively. The count 50 is the **total frequency**, which implies that we have listed all 50 cases. Since the data are grouped into non-numerical categories, the distribution is referred to as **qualitative distribution**. Note that the family size here is an ordinarily scaled variable.

You can construct a similar frequency distribution with religion as reported by the workers and shown in Table 2.1. Once you have done this, the resulting frequency distribution will appear as follows:

Table 2.3: Frequency distribution of workers by religion

Religion	Number of Workers	Percent (%)
Muslim	36	72
Hindu	9	18
Christian	5	10
Total	50	100

2.7 FREQUENCY DISTRIBUTION FOR QUANTITATIVE DATA

The construction of a frequency distribution with numerical or quantitative data is very similar to those for qualitative data, except that now the data have to be grouped into classes of appropriate intervals. The simplest device in doing so is to form an array first. An **array** is an ordering of values of the variable in order of their magnitude, usually in ascending order, i.e. from smallest to the largest. We illustrate the process of constructing such an array with the wage data in Table 2.1 before constructing a frequency distribution.

Table 2.4: Array of daily wage data

50	63	70	75	84
51	65	71	75	85
54	65	72	76	86
56	66	72	77	87
56	67	72	79	88
57	68	73	80	88
59	68	73	81	89
60	69	74	82	93
61	69	74	82	93
62	70	74	83	97

Such an array has the distinct advantage over the data in an unorganized form as in Table 2.1. It enables us to know at once that the minimum wage is Tk. 50 and the maximum is Tk 97.

The array, however, still offers only a cumbersome form of data organization, especially when the number of observations or values involved is very large. It is therefore more desirable to arrange the data into a number of mutually exclusive classes or categories with appropriate widths. We have called this arrangement a **frequency distribution**.

If we examine the data in Table 2.4, we see that some values have been repeated more than once. For example, the values 65, 68, 69, 73, 82, 88, 93 each has been repeated twice, 72 and 74 three times, and so on. One can then think of constructing a table with each value in one column and its frequency (number of times it occurs) in another column. By doing so, we have constructed an **ungrouped frequency distribution**. In this particular instance, the table will be of the following type:

Wage	Frequency
50	1
51	1
...	...
...	...
56	2
72	2
...	...
...	...
97	1
Total	50

The distribution in this form is still seen to be widely spread over a large number of cases and indicates visually no clear pattern and hence will not be very useful so far as the condensation of data is concerned. Secondly, most of the observations might have such low frequency counts associated with them that we are not justified in maintaining these observations as separate and distinct entities for economic reasons. Under these circumstances, it is customary for most researchers to group the data into several non-overlapping classes of reasonable width and then obtain a frequency distribution.

Although there is no "hard and fast" rule in constructing a frequency distribution from raw data, one must ensure that the frequency distribution so formed

- Contains as much information as possible and does not mislead the readers;
- Represents the complete range of possible values, whether or not they actually occurred;
- Enables the readers to visualize the extent to which the observed values are scattered over the range of possible values.

Grouping, however, has limitations too. One disadvantage of group distribution is the loss of information that inevitably results from grouping. For example, individual observations lose their identity when we group into classes, and some small errors in the calculated statistics (such as mean, variance, etc.) based on the grouped data, are inevitable. Carefully constructed frequency distributions, however, remove much of these limitations.

We now turn to discuss how a suitable frequency distribution can be constructed from raw data. The work is a simple one and will involve a few steps. But before we narrate these steps, we require to defining a few terms below, which are closely related to the construction of frequency distributions for numerical data.

Class: In the process of condensation, raw data are assigned to some chosen groups of appropriate size. These groups are called **classes**. A class is thus an interval containing observations, each observation being classified into one and only one class.

Frequency: The number of observations or values falling into each group or class is called **class frequency** or simply **frequency**. The frequency of a class thus shows how many times a particular value or observation is repeated in that class.

Class Limits: Ordinarily, for numerical data, the frequencies of a particular class are bounded by two values. The smaller value of the class is known as the **lower class limit**, while the larger value is known as the **upper class limit.** Class limits should be defined in such a way that no difficulty is experienced in assigning observed values to a class.

Class Boundary: Class boundaries, also called real limits or true limits, are the points which separate various classes rather than values being included in one of the classes. A class boundary is always located mid-way between the upper limits of a class and the lower limit of the next higher class.

Class Interval: The width or length of the class formed by the two boundary values is known as the **class interval** or **class width or class size.** A class interval represents the spread between the class boundaries.

Class Mark: The class mark is the value that lies in the middle of the class, and is obtained by averaging the two class boundaries. The class mark is also referred to as **class mid-point** or **mid-value** of the class.

Open Interval: An open interval is an interval with one of its limits (in either side) indeterminate. Thus an age of a person recorded as less than 45 years (*i.e.*<45) constitutes an open interval. Similarly, an age recorded as 75 and over (*i.e.* ≥75), also forms an open interval.

2.8 STEPS IN CONSTRUCTING GROUPED DISTRIBUTION

Having defined the above terms, we are now in a position to enumerate the principal steps of constructing a frequency distribution from raw data. The construction of such a distribution consists essentially of the following four steps.

(a) Decide on number of classes and the class widths in which the observations are to be grouped.
(b) Assign the observations to the appropriately chosen classes. This is called **tallying**.
(c) Count the number of observations falling in each class. These numbers are the **frequencies**.
(d) Display the results obtained in the above three steps in a table.

The resulting table is our desired **frequency distribution**.

The choice of the **class width** and the number of classes cannot be made independently. Larger class-width means fewer classes, and vice-versa. In deciding on the number of classes or class-width, one must examine the spread or variability of the individual observations within data set. Common sense suggests that the larger the variability, the larger is the class-width and hence fewer are the number of classes. Determining the exact number of classes and choosing a class-width however involve a tradeoff between too few and too many classes. However, the following points are important to remember in deciding the class width and number of classes.

❑ **The number of classes should be such that the true nature of the distribution is made manifest.**

When the number of classes is excessively large, and consequently the class intervals too small, it would not reveal the true pattern of the distribution, because each class would contain too few items or none at all. On the other hand, if the number of classes is excessively small and class intervals too wide, it may conceal the fact that relatively larger numbers of items are concentrated in a few classes with narrower intervals.

If the frequency distribution is depicted in graphic form, to facilitate visualization of the underlying pattern of the data, the number of classes should be fewer (say 5–15) than if the objective of the frequency distribution is to provide a more convenient framework for the performance of further statistical computations.

If the smallest value (S) and the largest value (L) in a data set are known, then as a rule of thumb, the range $R = L - S$, which shows the spread of the data, is divided by the class width (h) to determine the approximate number of classes desired (k). In other words

$$k = \frac{L-S}{h} = \frac{R}{h} \quad \ldots (2.1)$$

An empirical rule suggested by Sturge to determine the number of classes is the "2 to the k rule". This rule suggests that the number of classes should be the smallest whole number k that makes the quantity 2^k greater than or equal to the total number of observations (N) in the data set (i.e. $2^k \geq N$). Suppose a data set consists of N=50 observations. Then, since 2^5=32, which is smaller than N and 2^6=64, which is greater than N, the Sturge's rule dictates us to choose 6 classes, that is k=6.

❑ **The classes should preferably be of equal widths**

There are several reasons, why classes with uniform widths are desirable:

- Classes with equal width facilitate comparison of the frequency of occurrence between classes.
- If the frequency distribution serves as a framework for further computations, short-cut formula are available for distributions that have equal class widths.
- If the class widths are unequal, then we would have a distribution that is much more difficult to interpret than one with equal width.
- Unequal classes present difficulties in graphically presenting the distribution and in doing some computations of statistical measures.

Unequal class widths, however, may be necessary in certain situations to avoid a large number of empty or almost empty classes.

How to decide on the class interval? This depends primarily on how the data look like. The first thing is that it must not be too awkward to work with. If k is empirically determined, following Sturge, then a formula for h emerges from [2.1]:

$$h = \frac{\text{Range}}{\text{Number of classes desired}} = \frac{R}{k} \quad \ldots (2.2)$$

An empirical formula due to Sturge is also available which has been found to work well in many situations for choosing equal-spaced class interval (h):

$$h = \frac{R}{1 + 3.322 \log_{10} N} \quad \ldots (2.3)$$

❑ **Each observation or item should go into one and only one class**

Unless this is so, the total frequency will not add to the total number of observations. This calls for a careful and systematic reading of the observations during the tallying process.

❑ **Classes with zero frequencies should be avoided**

The principal reason for this avoidance is the awkwardness in the subsequent graphic and computational procedures.

❑ **Class mid-points should be so established that the mid-point is approximately equal to the arithmetic mean of the observations in the class**

If the frequency distribution is to be used for further computations, this characteristic is extremely important.

❑ **Class limits should be so established that the class mid-points fall on a whole number**

This is desirable simply because computations with whole numbers are simpler to perform and less subject to error.

❑ **Open-end classes should be avoided**

It is desirable to avoid open-end classes because a mid-point cannot be determined for such a class. We will see that mid-point is a pre-requisite for certain computations.

Occasionally, something unique in the data will cause us to deviate from the above guidelines, but they are appropriate most of the times.

We encounter relatively less problem in the formation of frequency distribution with the discrete data. So far as the formation of class intervals is concerned with continuous data, an important thing to note is whether the data are given to the nearest taka or to the nearest dollar, whether they are recorded to the nearest inch or the nearest tenth of an inch or the nearest hundredth of an inch, and so forth. In other words, we must examine the extent to which the numbers are rounded. The following example is designed to illustrate this point.

Suppose that we wanted to classify the heights of children. We would use first of the following three classifications, if the heights are recorded to the nearest inch. But if the heights are given to the nearest tenth of an inch, we would use the second classification, and the third, if the heights are recorded to hundredth of an inch.

Height (nearest inch)	Height (nearest 10th of an inch)	Height (nearest 100th of an inch)
25–29	25.0–29.9	25.00–29.99
30–34	30.0–34.9	30.00–34.99
35–39	35.0–39.9	35.00–39.99
40–44	40.0–44.9	40.00–44.99
45–49	45.0–49.5	45.00–49.99
etc	etc	etc

To illustrate what we learned so far, let us now go through the actual steps of constructing a frequency distribution for a given set of data. First, we will deal with discrete data.

2.9 FORMATION OF DISCRETE FREQUENCY DISTRIBUTION

The construction of a grouped frequency distribution from discrete data is illustrated with the following example.

Example 2.2: As shown in Table 2.1, the number of complete days the workers were absent from their work during the year preceding the inquiry are arranged below in an ascending array:

5	8	9	9	10	10	10	10	11	11
12	12	12	13	13	13	14	14	14	15
15	15	15	16	16	16	16	17	17	17
17	18	18	18	18	18	19	19	19	19
20	21	21	22	23	24	26	27	29	33

The smallest number in the data set is 5 and the largest number is 33 so that the range is 28. If we decide to use a class width of 10, then k works out approximately to 3 by (2.1).

$$k = \frac{R}{h} = \frac{28}{10} \approx 3$$

The three classes then would be: 5–14, 15–24 and 25–34. If we use as few as three classes as above, we would find that almost 55% of the observations fall into the second class. As a result, we will be losing too much information. What would be then our choice for the number of classes? Let us try with the Sturge's rule. Since $2^5=32$, which is less than N and $2^6=64$, which is more than N, the rule suggests a value of $k=6$. We therefore opt for the second alternative, viz. the 6 classes: 5–9, 10–14, 15–19, 20–24, 25–29, and 30–34. The implied class interval here is 5.

What does the Sturge's empirical formula suggest?

$$h = \frac{R}{1+3.322\log_{10}(50)} = \frac{28}{6.64} = 4.21$$

A class interval of 4.21 would be very awkward to work with and therefore we would round it to 5 for convenience. This is compatible with our previous choice.

Having decided on the class intervals and the number of classes, our next task is to tally the observations in their respective classes.

The accompanying table shows the result of this tallying process. It is constructed by reading down the data array vertically and on reading the first entry 5, entering a diagonal stroke (/) in the class 5–9, for the second entry 12, another stroke in the class 10–14, for the third and fourth entries 15 and 17, two more strokes are entered in the class 15–19 etc. After the tallying process is completed, the strokes in each class are counted and the number is entered in the last column of the table as the frequency of the class. The resulting table is as follows:

Class interval	Tally	Frequency
5–9	////	4
10–14	//// //// ////	15
15–19	//// //// //// //// /	21
20–24	//// /	6
25–29	///	3
30–34	/	1
Total	–	50

The data in Example 2.2 are discrete and hence the resulting distribution is a **discrete frequency distribution**. The numbers in the last column of this table are called the **class frequencies**, and they simply give the number of items in each class. Also in our example, 5, 10, 15, 20, 25, and 30 are the lower limits and 9, 14, 19, 24, 29, and 34 are the upper limits. For the class 15–19 (say), the **class mark or class mid-point** is 17 and the **class width** is 5. This is read as 15 to 19 **inclusive** and is obtained as (19–15)+1=5 or a difference of the lower limits of the two consecutive class limits. For discrete distributions, the class limits are **always** inclusive in nature.

Had the data in Example 2.2 been continuous, such as age, height, etc., the classes so formed would not have been proper. It is because of the fact that although this classification does not make any confusion as to which class a value is to belong, it does not maintain the continuity of the data. One way of maintaining the continuity is to reconstruct the frequency distribution in a manner so as to maintain the continuity of the data. We illustrate below how a frequency distribution is constructed with continuous data.

2.10 FORMATION OF CONTINUOUS FREQUENCY DISTRIBUTION

The method of constructing a frequency distribution for continuous data is illustrated below using the data in Table 2.1.

Example 2.3: The ages of the 50 workers appearing in Table 2.1 are reproduced below in ordered array. Construct a frequency distribution with appropriately chosen class widths and number of classes.

25	33	37	42	45
28	34	37	42	46
29	35	37	42	46
30	35	38	43	46
31	35	38	43	46
32	36	38	43	47
32	36	39	44	50
32	36	40	44	51
33	36	41	44	52
33	37	42	45	54

Solution: For the given data, $N=50$ and following Sturge's rule, k works out to 6. The choice of class limits and hence the class widths will be based on the range of the values. Since the highest observation is 54 and

the lowest is 25, the range is 29; we may approximate the width of the class by dividing the range R by k. The class-width determined in this manner is often not an integer and must be rounded up or down to an integer. We have in the present instance h=(54–25)/6=4.83, which is approximated to 5. The Sturge's empirical rule for h also suggests such a value for the class interval:

$$h = \frac{R}{1+3.322\log_{10}(50)} = \frac{29}{6.64} = 4.36 \approx 5$$

With the values of k and h specified above, we form a frequency distribution that appears below:

Table 2.5: Frequency distribution of the workers by age

Age in years	Number of workers
25–29	3
30–34	9
35–39	15
40–44	12
45–49	7
50–54	4
Total	50

Although the classification above does not make any confusion as to which class a value is to belong, it does not maintain the continuity of the data, despite the fact that the age data themselves are continuous. We describe below how to reconstruct the frequency distribution of the same data preserving their continuity.

The choice of the class limits reflects the extent to which the values being grouped are rounded off. The worker's ages in the present example are rounded to the nearest year. Thus a worker between 34 and 34.5 would be counted in the second class (i.e.30–34), whereas one who is between 34.5 and 35 would be counted in the third (i.e. 35–39). Thus 34.5 is really the boundary between the second and the third classes. Similar boundaries between the other classes may be determined. These are sometimes called **true class limits, real class limits** or **class boundaries.** The true limits of a value of a continuous variable are equal to that number plus or minus one half the unit of measurement. This adjustment consists in finding the difference between the lower limit of the second class and the upper limit of the first class, dividing the difference by two, subtracting the value so

obtained from the lower limit and adding the value to all upper limits. The correction factor (C) can be expressed as

$$C = \frac{\text{Lower limit of second class} - \text{Upper limit of first class}}{2}$$

so that C for the distribution in Table 2.5 is 0.5. Thus subtracting 0.5 from lower limits and adding 0.5 to all upper limits of the distribution, the frequency distribution with the true limits is as follows:

Table 2.6: Distribution showing true class limits

Age in years	Number of workers
24.5–29.5	3
29.5–34.5	9
34.5–39.5	15
39.5–44.5	12
44.5–49.5	7
49.5–54.5	4
Total	50

In this classification, the classes are so formed that the upper limit of one class is the lower limit of the next class. This type of classification is known as the **exclusive method** of classification. As a result, the upper and lower class limits are now renamed as upper and lower **class boundaries** respectively. Note that the method ensures continuity of data in as much as the upper limit of one class is the lower limit of the next class.

Class boundaries constructed in this fashion have certain advantages. For one thing, it ensures that each item falls within an interval, and not on any of the boundaries. This will not lead to ambiguities as to whether an item should go into one class or another, so long as we are careful in giving the class limits to a sufficient number of decimals. For another, any rounded figure will be put in the same class, as it would have been if it had not been rounded.

Under this modification, the width of any class is equal to the difference between lower boundary and the upper boundary of the class. It may also be obtained by finding the difference either between two successive lower boundaries, or between two successive upper boundaries or between two successive class mid-points so long as the class widths are equal.

Although the method is widely followed in practice, it raises confusion as to which class a value 29.5 (say) falling on the boundary belongs to.

An alternative way of expressing the classes that does not lead to such confusion is to read the class '24.5–29.5' as '24.5 to less than 29.5', '29.5–34.5' as '29.5 to less than 34.5' and so on. The distribution thus is as follows:

Age in years	Number of workers
24.5 to less than 29.5	3
29.5 to less than 34.5	9
34.5 to less than 39.5	15
39.5 to less than 44.5	12
44.5 to less than 49.5	7
49.5 to less than 54.5	4
Total	50

Example 2.4: For the daily wage data in Table 2.1, the highest wage is Tk 97 and the lowest is Tk 50. To decide on the number of classes, we examine the '2 to the k rule'. Since N=50, and since 2^5=32 is less than 50 and 2^6=64 is greater than 50, the rule suggests a value of k=6, implying that there will be 6 classes. With this choice of k, we estimate the interval with the formula (2.2) giving h=8. The resulting distribution is

Wage	Frequency
50–57	6
58–65	7
66–73	14
74–81	10
82–89	10
90–97	3
Total	50

Since the wage data are continuous, the distribution, when adjusted for this continuity, appears as follows:

Wage	Frequency
49.5–57.5	6
57.5–65.5	7
65.5–73.5	14
73.5–81.5	10
81.5–89.5	10
89.5–97.5	3
Total	50

In constructing the distribution, we note that the largest observation '97' is just barely included in the last class. In cases where the largest observation

is not contained in the last class, we are free to add one more class so as to contain the last observation. This inclusion is unlikely to do much harm in the accuracy of the estimates calculated from the given data.

In the determination of the number of classes, one of the most important considerations that we ignore is the **spread** of the data. Although range is taken into consideration while determining the class intervals, it is a very weak measure of variability of data. The more spread the data, the larger will be the class-width to obtain valid statistical measures, such as mean, variance and the like.

The grouping process may also sometimes lead to improper choice of the class intervals. This may result in a distorted and misleading frequency distribution. There is little substitute for common sense here, but it may be necessary to change the class intervals if we suspect the information is being hidden by a poor selection of the class intervals. Suppose, production statistics of an electric company is available for the last 21 days. The production manager wants to assess whether the production of electric bulbs is showing any increasing trend over time or not. If you present the weekly data, the manager will have an impression that the production is increasing. This is shown below:

	Quantity produced on							
Week	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Total
First	17	19	21	20	21	27	26	151
Second	29	30	32	34	36	34	37	232
Third	40	42	48	49	35	30	27	271

The table shows that the company produced 151 thousand bulbs during the first week, which increased to 232 thousand during the second week and finally to 271 thousand during the third week. The manager is satisfied since the production is showing an overall increasing trend. But if we show him the daily production, he will rather be disappointed noting that during the second half of the third week, the production has started to decline from 49 thousand to 27 thousand. Grouping the data thus tends to suppress the fact or the real trend. It does not, however, mean that grouping is unnecessary or not recommended, it simply reminds us the fact that we must be very careful in forming class intervals so that the real nature of data is not suppressed.

Example 2.5: The data below specify the longevity of 80 electric bulbs produced in an hour's run by an electric company recorded to the nearest tenth of a month. Construct a frequency distribution.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.0	3.3
3.4	1.5	3.1	3.3	3.8	3.1	4.7	3.7	3.2	3.2
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1	3.1	3.4
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4	3.2	3.3
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5	3.3	3.0
1.6	1.7	2.2	2.5	2.6	2.8	2.9	3.9	3.4	3.2
3.5	3.7	3.7	3.9	3.8	3.7	3.6	3.8	3.2	3.1
4.1	4.0	4.4	4.3	4.3	4.6	4.7	4.8	3.3	3.8

Here we have 80 observations, each of which has been recorded to the nearest tenth of a month. The 2^k rule suggests a value 7 for k and hence a value $(4.8-1.5)/7 \cong 0.5$ for h. Sturge's rule for estimating h also results in a value that can be approximated to 0.5 for h:

$$h = \frac{4.8-1.5}{1+3.322\log_{10} 80} = 0.45 \approx 0.5$$

With these choices, the frequency distribution is

Class	Frequency
1.5–1.9	4
2.0–2.4	2
2.5–2.9	8
3.0–3.4	30
3.5–3.9	20
4.0–4.4	10
4.5–4.9	6
Total	80

The distribution above displays the inclusive type classification of the data and hence needs to be adjusted to account for the continuity in the limits.

Since the data have been recorded with one place of decimal, it needs to be carried out to one more place of decimal employing a correction factor as discussed earlier. The correction factor here is

$$C = \frac{2.0-1.9}{2} = 0.05$$

Incorporating this correction factor, we have the following distribution.

Class boundaries	Frequency
1.45–1.95	4
1.95–2.45	2
2.45–2.95	8
2.95–3.45	30
3.45–3.95	20
3.95–4.45	10
4.45–4.95	6
Total	80

Generally speaking, the guidelines we have given for forming classes are not meant to be inflexible rules. Rather, they are intended to help us find reasonable classes.

2.11 OTHER FORMS OF FREQUENCY DISTRIBUTION

2.11.1 Percentage Distribution

It is sometimes convenient and rewarding to deal with percentage distribution rather than the absolute ones. A **percentage distribution** is formed by dividing the number of cases attributable to a category or class by the total number of cases and multiplying the resulting value by 100. Thus if f_i is the frequency of the i-th class of a frequency distribution and N the total frequency, then the percentage of cases falling in the i-th class is

$$P_i = \frac{f_i}{N} \times 100 \quad \ldots (2.4)$$

The total frequency in a percent distribution will add to 100.

One advantage of presenting the frequency distribution in percentage form is that it facilitates evaluating the relative importance of each of the classes. The analyst relates to a familiar 100-percent base instead of total frequency, thus standardizing the data by a common base. It is more awkward, for example, to compare 36 to 75 than to compare 48 percent to 100 percent. Percentage distributions are particularly useful where comparison must be made between two different frequency distributions that are similar with respect to class breakdown but differ in their total frequency when class intervals are of equal sizes.

2.11.2 Relative Frequency Distribution

Instead of presenting the frequencies in absolute terms, it is sometimes convenient to express the frequencies in relative terms. The resulting

distribution is then called **relative frequency distribution**. The relative frequency is simply the fraction or proportion of the total number of items belonging to the class or category. For a data set having a total of N observations, or items, the relative frequency of i-th class is f_i/N. The total relative frequency in such a case will add to 1.0.

For the wage data, the percentage frequency distribution and relative frequency distribution are shown in a single table as below:

Table 2.7: Relative frequency distribution based on wage data

Class boundaries (Wage)	Absolute frequency	Percentage frequency	Relative frequency
49.5–57.5	6	12.0	0.12
57.5–65.5	7	14.0	0.14
65.5–73.5	14	28.0	0.28
73.5–81.5	10	20.0	0.20
81.5–89.5	10	20.0	0.20
89.5–97.5	3	6.0	0.06
Total	50	100.0	1.00

2.12 CUMULATIVE FREQUENCY DISTRIBUTION

In many occasions, the analyst is interested not in the number of observations falling in each class, but rather: how many of the observations in the distribution have a value less than or more than some benchmark values. Referring to Table 2.7, we might ask a series of questions, a few of which are:

- How many of the workers earn less than taka 73.5?
- How many of the workers earn taka 57.5 or more?

The answers to these questions can conveniently be given by constructing a distribution what we refer to as a **cumulative frequency distribution**. Two forms of cumulative distributions are in use: less than type and more than type.

The **less than type** cumulative frequency provides the total (cumulative) frequency below the upper class limits or boundaries for each class. This can be formed from the frequency distribution, the relative frequency distribution or from the percent distribution. For the wage data, a less than type cumulative frequency distribution appears in Table 2.8 below:

Table 2.8: Less than frequency distribution for wage data

Wage	Frequency	Cumulative frequency	%Cumulative frequency
49.5–57.5	6	6	12.0
57.5–65.5	7	13	26.0
65.5–73.5	14	27	54.0
73.5–81.5	10	37	74.0
81.5–89.5	10	47	94.0
89.5–97.5	3	50	100.0
Total	50	–	–

The cumulative frequencies in column 3 represent the wage less than the upper boundaries of the indicated wage in column 1. Thus the cumulative frequency 27 in column 3 states that 27 workers received less than taka 73.5. A percentage cumulative distribution can also be developed from the percentage distribution in the same way that a cumulative frequency distribution is developed from a frequency distribution. Such percentages are easier to interpret than the absolute frequencies. Thus instead of saying that 27 workers received less than taka 76.5 as above, it is more convenient to say that 54 percent workers received such an amount. These percentages are shown in the last column of the table under reference.

A useful way of presenting a less than cumulative frequency distribution is to add a new class with the occurrence of **zero** frequency and put the distribution as follows:

Wage	Cumulative frequency	% cumulative frequency
Less than 49.5	0	0
Less than 57.5	6	12.0
Less than 65.5	13	26.0
Less than 73.5	27	54.0
Less than 81.5	37	74.0
Less than 89.5	47	94.0
Less than 97.5	50	100.0

Addition of a new class does not have any bearing on the distribution; rather it has the advantage, as we will see later, of drawing a mathematically meaningful frequency curve.

The more than type cumulative distribution or **decumulative distribution** is employed when the question is to ascertain how many observations or

items in the distribution have a value greater than or equal to the value of the lower limit or lower boundary of certain class. The term **decumulative** is used to describe the "more than" frequency distribution because movement through the distribution is accompanied by a decumulation in frequency. Referring to the wage distribution in Table 2.7 again, we observe that all 50 of the workers have wage more than or equal to the lower class boundary of the first class, viz. 49.5. The accompanying table shows a decumulative distribution for the wage data.

Table 2.9: More than type frequency distribution for wage data

Wage	Decumulative frequency	% Decumulative frequency
49.5 or more	50	100.0
57.5 or more	44	88.0
65.5 or more	37	74.0
73.5 or more	23	46.0
81.5 or more	13	26.0
89.5 or more	3	06.0
97.5 or more	.0	.0

The table is simple to read. The second entry '44' in column 2 simply states that out 50 workers, 44 (or 88%) receive a sum of taka 57.5 or more. The other entries can similarly be interpreted. Here also note that an additional class has been added at the end of the distribution with zero frequency of occurrence for the sole purpose of drawing easily comprehensible frequency curve.

2.12.1 Bi-variate Frequency Distribution

Thus far we have been discussing construction of frequency tables with a single variable, such as age, sex or religion. These tables represent uni-variate frequency distribution. They are called uni-variate because they involve only one variable. Any statistical analysis based on these tables will be called **uni-variate analysis**. Frequently, more than one variable are studied simultaneously to establish causal relations among the variables. For example, we might be interested to study the relationship between education of the workers and their family size. Such an analysis is facilitated through constructing what is called **bi-variate frequency distribution**. The rows of such table identify the categories of one variable and the columns identify the categories of the other variables. The entries in the table are called the **cell frequencies**, which are essentially the

number of times each value of one variable occurs with each possible value of the other. When such tables are constructed with qualitative data, the resulting table is called **contingency table**. The simplest method of looking at relations between variables in a contingency table is to do a percentage comparison based on the row totals, column totals or the overall totals.

Suppose, it is possible to classify a workers by his/her education and family size simultaneously. This attempt produced the following table:

Table 2.10: Family size and education level of 50 workers

	Family size			
Education	Large	Medium	Small	Total
None	4	6	1	11
Primary	6	8	5	19
Higher	6	10	4	20
Total	16	24	10	50

Table 2.10 is a **contingency table** of two categorical (qualitative) variables: family size and level of education both measured on ordinal scale. Since two variables are involved in the construction of the above table, the table is also known as **bi-variate** (or 2-way) table. Also since both the variables have three levels, it is also known as a 3×3 (read three by three) **cross table**. This table is intended to answer the question of the type: does education have any effect on the family size? Here family size is a dependent variable (**thè problem**) and education is an independent variable (**the factor**). The totals in the columns and rows are called the **marginal frequencies**, which in fact, constitute the uni-variate frequency distributions of the variables education status and family size respectively. Thus 16, 24, and 10 in the row total are the marginal frequencies, so are the column totals 11, 19 and 20. The distributions formed with these frequencies are known as the marginal distributions. One such distribution with education is shown in Table 2.11 below.

Table 2.11: Marginal distribution of respondents by level of education

Education	Number	Percent
None	11	22
Primary	19	38
Higher	20	40
Total	50	100

In a similar way, one can think of a **tri-variate table**, in which three variables are involved: one dependent and two independent variables. The concept can be extended to multivariate cases, which involves several variables. We illustrate here the case of three variables, the third variable being respondents' religious affiliation (Muslim, Hindu, ..).

Suppose that from Table 2.10, we infer by some statistical testing that workers with more education are more likely to have smaller family than those who are less educated. One might argue that this difference is due to religion. To test his claim, one can form a table that displays the relationships of education and family size for each religion category as shown in the accompanying table. This table is a tri-variate table. Religion, in this particular instance is known as **controlled** variable.

Table 2.12: A Tri-variate table for education and family size by religion

		Family size			
Religion		Large	Medium	Small	Total
Muslim	Education				
	None	4	3	1	8
	Primary	5	5	3	13
	Higher	4	7	4	15
	Total	13	15	8	36
Non-Muslim	Education				
	None	0	3	0	3
	Primary	1	3	2	6
	Higher	2	3	0	5
	Total	3	9	2	14

In constructing this table, we have merged Hindus and Christian into a new category "Non-Muslim". This table is of order 3×3×2.

For numeric variables, we can construct bi-variate table too. The accompanying table is such a table constructed from the wage and age data of Table 2.1. Since the data spread over a long range values, we recode the data into some suitable groups and then put the frequencies in the respective cells. For illustrative purposes, the age data were recoded as 25-34, 35–44, and 45–54, while the wage data were recoded as 50–65, 66–81 and 82–97. The resulting bi-variate table is as follows:

	Age			
Wage	25–34	35–44	45–54	Total
50–65	2	6	5	13
66–81	6	13	5	24
82–97	4	8	1	13
Total	12	27	11	50

When a table of this type is constructed with numerical data, it is called **correlation table** in contrast to contingency tables constructed when the data are qualitative.

The procedure of constructing such a table follows the same principle as for constructing a uni-variate table. Look at the age and wage data in Table 2.1 simultaneously. For the first worker, we have the pair (93, 25) for wage and age. Look at the wage column vertically. You find that 93 is located within 82–97. The age 25 is located in the range 25–34. Put a tally mark in the intersection of 82–97 and 25–34. Proceed in the same way with other pairs of values and complete the table.

2.13 DESIRABLE FEATURES OF A FREQUENCY TABLE

We enumerate below a few desirable features of a frequency distribution

(a) A frequency table should have a title that explains the contents of the table clearly.
(b) A table should be as simple as possible and avoid unnecessary details.
(c) Each row and column of the table should be labeled clearly and concisely.
(d) The column and row totals should be provided if necessary.
(e) Units of data should be provided when appropriate.
(f) Sources of data, whenever possible, should be indicated.
(g) Any explanation of the data inside the table should be given beneath the table as footnote.

2.14 PRESENTING DATA BY GRAPHS AND DIAGRAMS

In addition to presenting statistical data through tabular form, one can present the same through some visual aids. This refers to **graphs** and **diagrams**. This is one of the most convincing and appealing ways in which statistical data may be presented. Such a presentation gives a bird's eye view of the entire data and therefore the information presented is easily understood. When frequency distributions are constructed primarily to condense large sets of data into an **easy to digest** form, graphical and diagrammatic presentations are preferred. The most common forms of diagrams for presenting categorical (qualitative) data are (a) Bar diagram (b) Pie diagram (c) Multiple bar diagram (d) Component bar diagram and (e) Pareto diagram

2.15 PRESENTATION OF QUALITATIVE DATA

2.15.1 Bar Diagram

A bar diagram, also known as bar chart, is a form of presentation in which the frequencies are represented by rectangles usually separated along the horizontal axis and drawn as bars of convenient widths. Its lengths are equal to the frequency or proportional to the magnitudes they represent. The widths of these bars have no significance but are taken to make the chart look attractive. In presenting the bars, there is no necessity of having a continuous scale.

Example 2.6: The accompanying table shows the stock position of finished goods in metric tons as of June 2004 of the Bangladesh Chemical Industries Corporation. Represent the data by a bar diagram.

Finished goods	Quantity
TSP	8916
SSP	18455
Paper	2660
Cement	7048
Sanitary ware	862
Insulator	1462
Tiles	17335

The vertical bar diagram constructed from these data is shown in Figure 2.1 below.

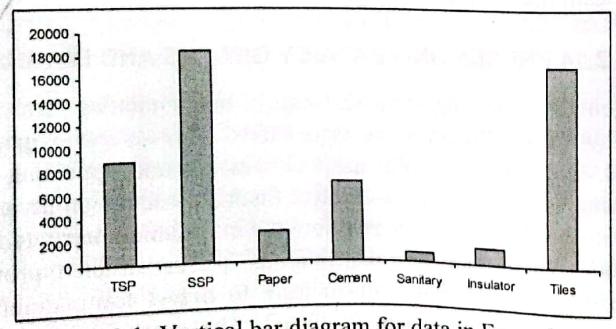

Figure 2.1: Vertical bar diagram for data in Example 2.6

The above chart shows a conventional form of bar diagram. We can also form a relative frequency distribution and present the same in a bar diagram. You will note that the bar charts in these 2 cases will be identical.

The bar diagram presented above is a vertical bar diagram. You can construct a horizontal bar diagram with the same set of data. Such a diagram appears in Figure 2.2.

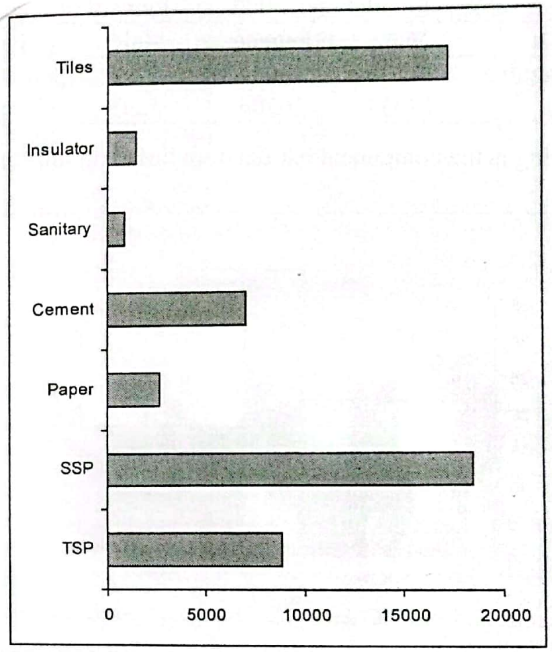

Figure 2.2: Horizontal bar diagram for data in Example 2.6

2.15.2 Component Bar Diagram

Component bar diagram, also called stacked bar diagram, is a good device to display categorical data. In such a diagram, the total values as well as the various components constituting the total are shown. The bar is sub-divided into as many parts as there are components. Because of this, the diagram is also known as **sub-divided bars or sub-divided rectangles.**

Each part of the bar in the component bar diagram represents each component, while the whole bar represents the total value. The component parts are variously colored or shaded to make them distinct. Instead of using absolute values, one can use also percentage values to construct the component bar diagram. The following example is designed to illustrate the construction of component bar diagram.

Example 2.7: Given below are the 1991census population of Chittagong and Dhaka divisions by sex. Display them by a component bar diagram.

	Population in '000		Percent of population	
Region	Male	Female	Male	Female
Chittagong	11228	10637	51.3	48.7
Dhaka	17634	16306	52.0	48.0

The following is the component bar diagram based on the absolute values:

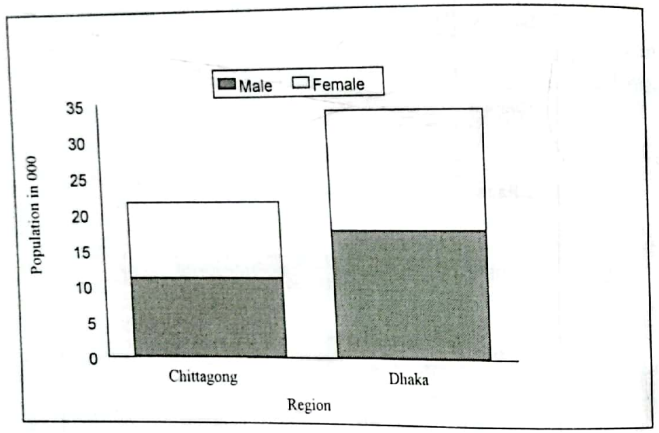

Figure 2.3: Component bar diagram for data in Example 2.7

2.15.3 Multiple Bar Diagram

Another diagram, which is frequently used to present statistical data, is th multiple bar diagram. This is primarily used to compare two or mo characteristics corresponding to a common variate value. Multiple b charts are grouped bars, whose lengths are proportional to the magnitud of the characteristics. The bars of a multiple chart are usually put adjace to each other without allowing any space between them. Different shadi or color can be used to distinguish one group of bars from other groups.

Data on population values for different regions, literacy rates by se volume of exports by type of production etc. can be represented b multiple bar chart. Data in Table 2.13 are used to illustrate the constructio of a multiple bar chart (see Figure 2.4).

Table 2.13: Education level of female population of Bangladesh

	Percent of females with		
Division	No education	Primary education	Secondary education
Barisal	43.9	34.4	21.7
Chittagong	41.8	37.0	21.2
Dhaka	45.9	35.3	18.8
Khulna	39.6	41.2	19.2
Rajshahi	48.5	37.8	13.7
Sylhet	52.6	36.1	11.3

Source: Bangladesh Demographic and Health Survey: 1996-97

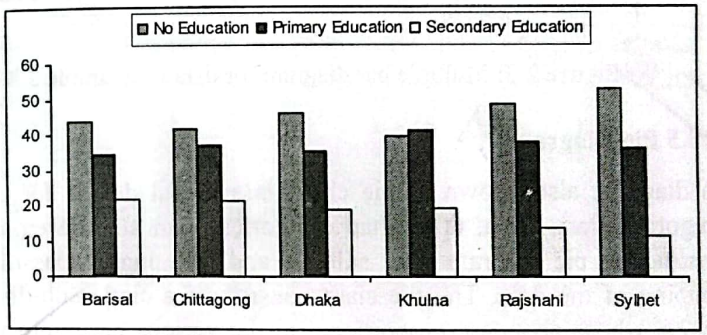

Figure 2.4: Multiple bar diagram for data in Table 2.13

Example 2.8: The accompanying table shows the per capita income of SAARC countries for the period 1999–2000, 2000–2001, and 2001–2002. Display the data by a multiple bar chart.

Countries	Per capita income in US$		
	1999–2000	2000–2001	2001–2002
Bangladesh	381	374	378
Bhutan	510	560	600
India	450	460	470
Maldives	2130	2130	2170
Nepal	230	240	230
Pakistan	450	420	420
Sri Lanka	890	840	850

Source: National Accounts Statistics, BBS (July 2004)

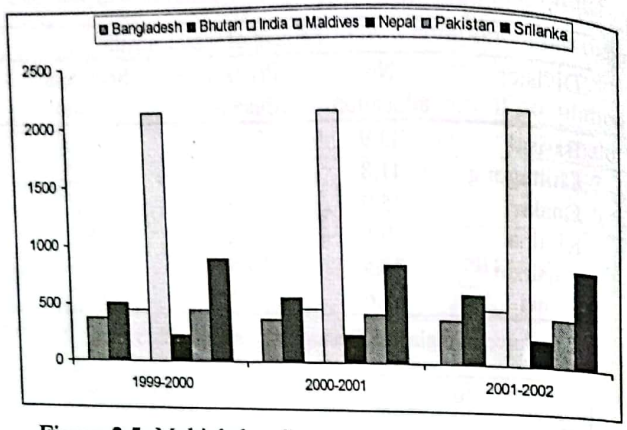

Figure 2.5: Multiple bar diagram for data in Example 2.8

2.15.5 Pie Diagram

Pie diagram, also known as pie chart, is a useful device for presenting categorical data. Data, other than categorical, can also be employed for constructing pie diagram after suitable and meaningful classification or grouping of the data. The pie chart consists of a circle sub-divided into sectors, whose areas are proportional to the various parts into which the whole quantity is divided. The sectors may be shaded or colored differently to show their individual contributions to the whole. The following steps are involved in constructing a pie chart

(a) Convert the absolute frequencies into relative frequencies for each category of the variable.
(b) Multiply the relative frequencies so converted by 360 for each category. The resulting values are the angles expressed in degrees.
(c) Check that the column obtained in step (b) adds to 360.
(d) Draw a circle of appropriate radius.
(e) Present the figures obtained in step (b) in the circle with the help of a protractor.

The resulting figure is the desired pie diagram of your data.

Example 2.9: The table below shows the production of major crops as assessed by BBS for the period 2003–2004

Major crops	Production (Lac metric tons)
Aus	18.32
Amon	115.2
Boro	124.5
Wheat	15.07
Potato	31.53

Source: Statistical Yearbook 2001

Draw a pie chart to represent the data.

Following the steps outlined above, we construct a table to facilitate the drawing of the desired pie for the data in Example 2.9.

Major crops (a)	Production (b)	Relative frequency (c)=(b)÷304	Angles in degrees (d)=(c)×360
Aus	18	.060	22
Amon	115	.378	136
Boro	124	.408	147
Wheat	15	.049	18
Potato	32	.105	37
Total	304	1.00	360

Now draw a circle with a suitable radius and present the figures in column (d) in the circle with the help of a protractor. The pie diagram appears in Figure 2.6.

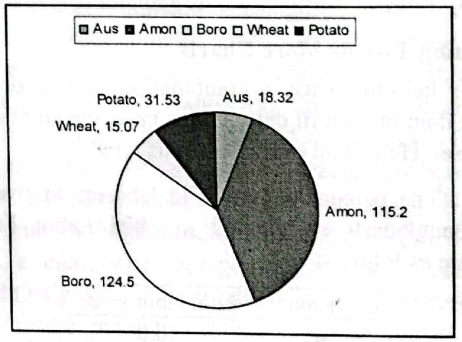

Figure 2.6: Simple Pie diagram for data in Example 2.9

The pie and bar diagrams presented here can be put in various other forms. Due to the recent development in the electronic computers, the task of presenting data by graphs and diagrams has become much easier and the

scope is wider than before. We show one more type of pie diagram in Figure 2.7 that displays the data in Example 2.10.

Example 2.10: The accompanying table shows the percentage of Bangladeshi expatriates in different countries by categories in 2002. Display the data by a pie chart

Category	Percent of expatriates
Unskilled	52.61
Skilled	24.98
Semi-skilled	15.99
Professional	6.41

Source: Bangladesh Economic Review–2003.

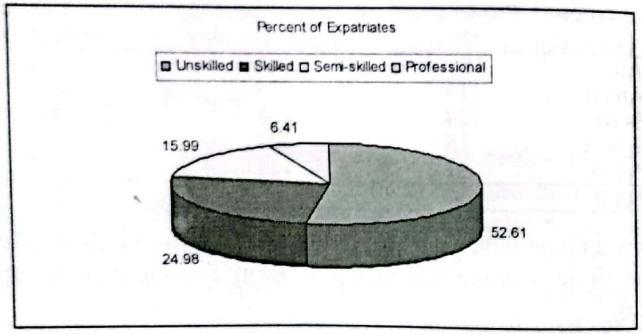

Figure 2.7: Three dimensional pie diagram for data in Example 2.10

2.15.6 Combining Two or More Charts

Sometimes it is helpful to use a combination of two or more charts to represent more than one set of data. In this case, the additional set is a part of the original set. Here is an example of this type:

Example 2.11: The percentages of child laborers in five administrative divisions of Bangladesh as reported in 1996 Labor Force Survey of Bangladesh were as follows:

Division	% in labor force
Barisal	10.0
Chittagong	25.1
Dhaka	28.2
Khulna	11.5
Rajshahi	25.2
Total	100

Among the 28.2% child laborers in Dhaka division, there were 10.1% seasonal, 82.7% temporary and 7.2% permanent child labor in labor force. Present the data by a suitable diagram.

We use Figure 2.8 to represent the above data.

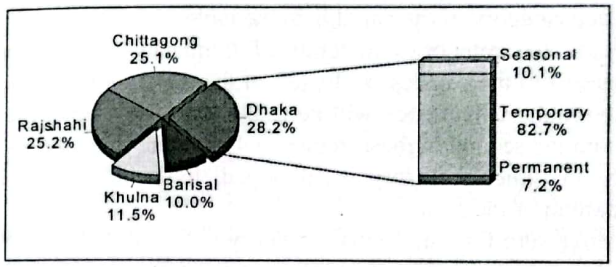

Figure 2.8: Pie-cum-component bar diagram for data in Example 2.11

We could however, present the component part (i.e.28.1%) by a bar or even a second pie diagram. But present representation appears to be more revealing in this particular instance.

2.15.7 Pareto Diagram

Pareto chart is a special type of graphical device used to help identify important quality problems in manufacturing industries. They help to point out opportunities for process improvement. By using these charts, we can place priorities on problem-solving activities. The chart is named after Vilfredo Pareto (1848–1923), who was an Italian economist. Pareto suggested that, in many economies, most of the wealth is held by a small minority of the population. This feature of the economy is well represented by a Pareto chart. It is represented by vertical bars in which the categorical responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same scale. The main principle behind this graphical technique is its ability to separate the "vital few" and the "trivial many" enabling us to focus on the important responses. The vital few are the small number of responses that account for the large percentage of the total, while the trivial many are the large number of responses that account for the small remaining percentage of the total. Pareto's concept, often called the 80–20 rule, is that 80 percent of the problems is caused by 20 percent of the problems. The chart achieves its

greatest utility when the categorical variable of interest contains many categories.

To develop a Pareto chart, we proceed as follows:

(a) Construct a frequency distribution of the variable of interest
(b) Display the frequency, percentage and cumulative frequency for each category of the variable in the table
(c) Rank the categories in terms of frequency of occurrence from largest to the smallest at the left of the table, i.e. the category with the highest frequency will be at the top of the table, the category with the second-highest frequency below the first, and so forth.
(d) If an 'other' category is employed, it should be placed at the bottom of the table.
(e) Make sure that the 'other' category does not make up 50 percent or more of the total frequency and the frequency of the 'other' category remains below the frequency of the category at the top of the table.

The vertical axis of the Pareto chart on the left contains the frequencies or percentage frequencies, the vertical axis on the right contains the cumulative percentages (from 100 on the top to 0 on bottom), and the horizontal axis contains the categories of interest. The equally spaced bars are of equal width. The point on the cumulative percentage polygon for each category is centered at the mid-point of each respective bar. Hence when studying a Pareto chart, we should be focusing on two things: the magnitude of the differences of the bar lengths, corresponding to adjacent descending categories and the cumulative percentages of these adjacent categories.

Example 2.12: The accompanying table provides data on the number of defects on the labels being placed on 16–ounce jars of grape jelly by type of defects. Display the data by a Pareto chart.

Type of defect	Frequency
Printing error	33
Crooked label	78
Wrinkled label	14
Smudged label	6
Loose label	23
Missing label	45
Others	12
Total	211

(Source: Bowerman and O'Connel, 1997)

To develop a Pareto chart, the following table is constructed:

Labeling defect	Frequency	%	Cumulative frequency	% Cumulative frequency
Crooked label	78	36.96	78	36.97
Missing label	45	21.33	123	58.30
Printing error	33	15.64	156	73.94
Loose label	23	10.90	179	84.84
Wrinkled label	14	6.64	193	91.48
Smudged label	6	2.84	199	94.32
Others	12	5.69	211	100.1
Total	211	100.0	–	–

Looking at the chart, we observe that the heights of the bars on the vertical scale represent the frequency of occurrence of the labeling defects. The bars are arranged in decreasing height from left to right. Thus, the most frequent defect (here the crooked label) is at the far left, the next most frequent defect (here the missing label) to its right and so forth. The chart graphically illustrates that crooked label, missing label and printing errors are the 'vital few', which account for about three-fourths of the labeling defects.

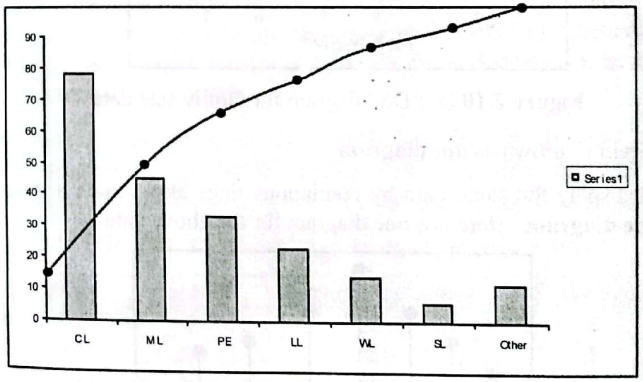

Figure 2.9: Pareto chart for data in Example 2.12

The Pareto chart is widely used in the statistical control of the process and product quality. Thus in the process of identifying the number and type of defects in manufacturing products or services, if the vital few defects are very costly to an organization, it may wish to work on eliminating their causes before working to solve other problems.

2.16 PRESENTATION OF QUANTITATIVE DATA

Quantitative data are available in two forms: discrete and continuous. As such, the diagrams to be used to represent them also vary. Although the frequency distributions with discrete data can be presented by bars, it is desirable to present them by some separate diagrams to make them distinct from diagrams employed for qualitative data.

2.16.1 Diagrams for Discrete Data

Discrete data may be presented by either dotted or continuous lines. Suppose we have a survey data on number of children in 65 families as shown below:

Family size:	1	2	3	4	5	6	7
Number of families:	1	3	4	5	4	3	2

A dot diagram for these data will be as follows:

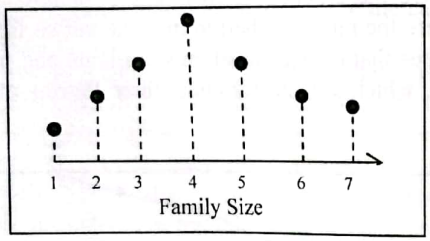

Figure 2.10 (a): Dot diagram for family size data

The diagram is known as **dot diagram**.

You can display the same data by continuous lines also. Such a diagram is called **line diagram**. Here is a line diagram for the above data:

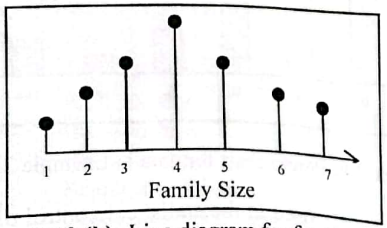

Figure 2.10 (b): Line diagram for family size data

2.16.2 Diagrams for Continuous Data

Histogram

The most common form of graphical presentation of a frequency distribution is the histogram. A histogram is constructed by placing the class boundaries on the horizontal axis of a graph and the frequencies on the vertical axis. Each class is shown on the graph by drawing a rectangle whose base is the class boundary and whose height is the corresponding frequency for the class. When the class boundaries are required to be unequal because of some particular feature of the data set, the method of constructing a histogram should be modified accordingly.

To avoid the ambiguity of equal and unequal class widths, it is advisable to construct histograms with proportional frequencies relative to the width of the class boundaries. This ensures that, it is the area, not the height that represents frequency of a class. Thus, if the frequency of a class is 30 and the class-width is 5, then the height of the vertical bar should be taken as 30/5=6 so that the area of the rectangle 6x5=30 represents the frequency. If the width of another class is different from 5, say 6 and the frequency is 24, then the height of the rectangle will be 24/6=4. Thus if a histogram of frequency distribution with unequal class-widths is to be constructed, necessary modification must be made to adjust the vertical height of the rectangle, so that the area of the rectangle represents the frequency. We now illustrate below how a histogram is constructed when we have equal as well as unequal class widths.

Consider first the case of equal class-width. We illustrate the case with the following frequency distribution constructed from data on the weekly expenditure in taka for 80 students of an elementary school of Dhaka City:

Table 2.14: Data for histogram for equal class widths

Expenditure	Frequency	Height of rectangles	Class width
04.5–9.5	8	8	5
9.5 –14.5	29	29	5
14.5–19.5	27	27	5
19.5–24.5	12	12	5
24.5–29.5	4	4	5
Total	80	-	-

The histogram when constructed with the above data will look like as in Figure 2.11.

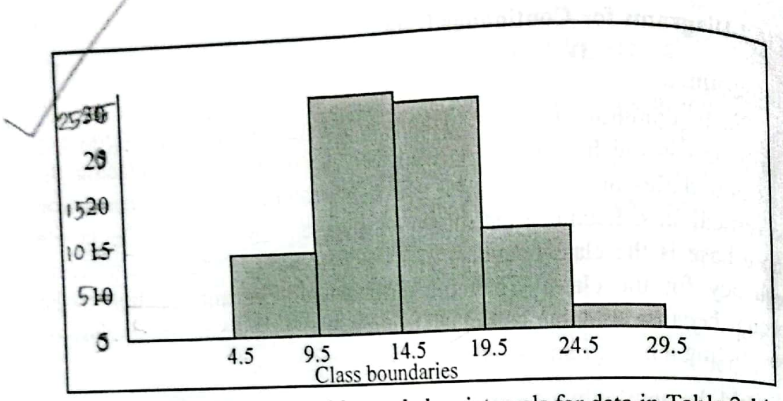

Figure 2.11: Histogram with equal class intervals for data in Table 2.14

To construct a histogram for a frequency distribution with unequal class widths, consider the following frequency distribution

Table 2.15: Data for histogram for unequal class widths

Class boundary	Class Frequency	Class width	Height of rectangles
48.5–58.5	4	10	4÷10=0.4
58.5–68.5	8	10	8÷10=0.8
68.5–73.5	5	5	5÷5=1.0
73.5–78.5	5	5	5÷5=1.0
78.5–98.5	28	20	28÷20=1.4
Total	50	–	–

In Table 2.15, the widths of class intervals are made unequal, varying from 5 to 20. Now to construct histogram, heights of the rectangles are obtained by dividing the frequency of each class by the corresponding class widths. These are in essence, proportional heights and are often called **frequency densities**. These values are shown in column 4. The resulting histogram appears in Figure 2.12.

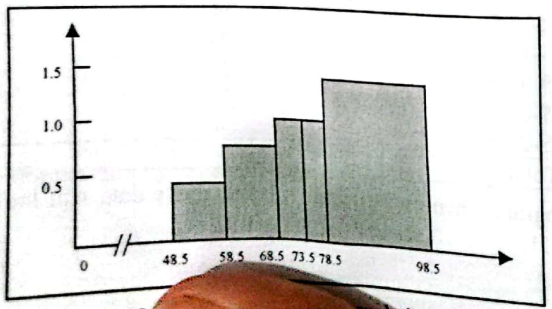

Figure 2.12 ...ual class intervals

2.16.3 Difference between a Bar Diagram and a Histogram

How does a histogram differ from a bar diagram? We enumerate below a few points of differences between a histogram and a bar diagram:

(a) In a histogram, the rectangles are adjacent to each other, while the choice of the spacing in a bar diagram is arbitrary.

(b) In a histogram, the areas of the rectangles represent the frequencies, but in a bar diagram, heights of bars represent the frequencies

(c) A histogram is a two dimensional figure, while a bar diagram is a uni-dimensional figure.

(d) A histogram is constructed for a continuous frequency distribution, while a bar diagram is usually constructed for categorical (qualitative) data.

2.16.4 Stem and Leaf Plot

The stem and leaf plot is a simple device to construct a histogram-like picture of a frequency distribution. It allows us to use the information contained in a frequency distribution, to show the range and concentrations of the scores, the shape of the distribution, presence of any specific values or scores not represented and whether there are any stray or extreme values (outliers) in the distribution. This is, in essence, a display technique taken from the area of statistics called **exploratory data analysis** (EDA). Compared to other graphical techniques presented thus far, stem and leaf plot is an easy and quick way of displaying data. The technique was first proposed by Tukey (1977) as an aid to understanding and exploring data through statistical analysis. We now illustrate the technique by an example.

Example 2.13: The following data represent the marks obtained by 20 students in a statistics test.

84	17	38	45	47	53	76	54	75	22
66	65	55	54	51	33	39	19	54	72

Use a stem and leaf plot to display the data.

Solution: We note that the lowest score is 17 and the highest score is 84. For stem and leaf plots, classes must be of equal lengths. We will use the first or leading digit (tens) of score as the stem and the **trailing** (units) digits as the leaf. For example, for the score 84, the leading digit is 8, and the trailing digit is 4; for 72, the leading digit is 7 and the trailing digit is 2 and so on. In a frequency distribution, as you might recall, a class interval determines where a

measurement or observation is to be placed. The stem and leaf plot follows the same principle, in which a leading digit (stem of a score) determines the row in which the score is placed. The trailing digits for a score are then written in the appropriate row. In this way each score is recorded in the stem and leaf plot. With the given data now, let us take the "stem" to represent the tens (leading digits) and the "leaf" the units (trailing digits). Thus for the first 5 scores 84, 17, 38, 45, and 47, the plot is

Stem	Leaf
1	7
2	
3	8
4	57
5	
6	
7	
8	4

The complete diagram in unordered sequence is

Stem	Leaf
1	79
2	2
3	839
4	57
5	345414
6	65
7	652
8	4

We then arrange the leaves in ascending order in order to make the plot a bit neater and give an **explanatory message** or a **key** beneath the plot. The final figure in ordered sequence is

Stem	Leaf
1	79
2	2
3	389
4	57
5	134445
6	56
7	256
8	4

Key: 1|7 represents 17.

To read the score from the above figure, start at the first row and read the scores 17 and 19. These scores are shown as 1|7, 9. The key beneath the table helps to understand this presentation. The second row contains 22, while the third row contains three scores: 33, 38 and 39 and so on. Note that the number of leaves must be equal to the number of observations. From the figure, the largest score (84) and the smallest score (17) can be readily located. In addition, an entire picture of how the scores are distributed (or scattered) emerges. For example, it is readily apparent that there are more scores in the fifties than any other group; 8 scores are less than 50, and only 4 scores are above 70.

Note that the plot looks like a horizontal histogram. It turns out to be a usual histogram if the plot is rotated 90 degrees counterclockwise. The advantage of a stem and leaf plot over the histogram is that it reflects not only frequencies, concentrations of scores and shape of the distribution, but also the actual score from which we can determine whether there are any values not represented and whether there are stray or extreme values (outliers). Another advantage of a stem and leaf plot is that it retains the original data.

In the above example, we have shown that stems represent the tenth digits, but it does not necessarily represent the tenth digit always. The following examples demonstrate this feature:

Further note that each stem defines a class interval and limits of each interval are the largest and the smallest possible scores for the class. The values represented by each leaf must be between the lower and the upper limits of the interval. The chosen classes in this particular instance are seen to be 10–19, 20–29, …,and 80–89. In other words, the stem and leaf display shown above represents the following frequency distribution:

Class	Frequency
10–19	2
20-29	1
30–39	3
40–49	2
50–59	6
60–69	2
70–79	3
80–89	1
Total	20

Example 2.14: Display the values 6, 8, 12, 14, 14, 15, 15, 16, 18, 19, 19, 23, 23, 24, 26, 26 by a stem and leaf diagram.

Stem	Leaf
5	13
10	244
15	001344
20	334
25	11

Key: 5|1 means 6, 15|3 means18

Example 2.15: Display the values 161, 163, 163, 166, 168, 168, 169, 169, 170, 170, 171, 173, 173, 174, 175, 177, 179, 180 by stem and leaf diagram.

Stem	Leaf
16	133
16	68899
17	001334
17	579
18	0

Key: 16|8 means168, 17|0 means170

Example 2.16: Using the data on longevity displayed in Table 2.5 (which is reproduced below for ease of representation), construct a stem and leaf plot.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.0	3.3
3.4	1.5	3.1	3.3	3.8	3.1	4.7	3.7	3.2	3.2
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1	3.1	3.4
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4	3.2	3.3
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5	3.3	3.0
1.6	1.7	2.2	2.5	2.6	2.8	2.9	3.9	3.4	3.2
3.5	3.7	3.7	3.9	3.8	3.7	3.6	3.8	3.2	3.1
4.1	4.0	4.4	4.3	4.3	4.6	4.7	4.8	3.3	3.8

We display the whole numbers 1, 2, 3 etc, as the stems and the digits after the decimals as the leaves. The stem and leaf display in the unordered and ordered forms appears below:

Unordered display:

The unordered display is a kind of display in which the observed values in the leaf segment appear irrespective of the magnitude of the values, while in the ordered display, the values usually appear in ascending order.

Stem	Leaf
1	5967
2	2659625689
3	52700341381722463911431724423829053094257798768213 8
4	1573417210433678

Ordered display:

Stem	Leaf
1	5679
2	2255666899
3	00001111112222222233333334444445556677777888889999
4	0111233344567778

Key: 1|5 means 1.5, 3|0 means 3.0

The first class corresponds to the stem 1 and consists of all values in the range 1.5–1.9 with the frequency 4. The second class corresponds to the stem 2 and contains all values in the range 2.0–2.9 with frequency 10. The other two classes are formed in a similar manner and these two classes contain the frequencies 50 and 16.

The plot of data presented in Example 2.16 contains only four stems and consequently does not provide an adequate picture of the distribution. To overcome this problem, we need to create more classes by using stems labeled as 1, 2^*, 2, 3^*, 3, 4^* and 4. Here, for instance, in the row labeled 2^*, we place the values from 2.0 to 2.4, and in the row labeled 2, we place the values from 2.5 to 2.9. Doing this, we obtain a modified stem and leaf display that appears below:

Stem	Leaf
1	5679
2^*	22
2	55666899
3^*	00001111112222222233333333444444
3	55566777777888889999
4^*	0111233344
4	567778

The display now represents the following frequency distribution:

Class	Frequency
1.5–1.9	4
2.0–2.4	2
2.5–2.9	8
3.0–3.4	30
3.5–3.9	20
4.0–4.4	10
4.5–4.9	6
Total	80

This is the distribution that we exactly constructed with the same data i Example 2.5.

Example 2.17: The Bangladesh Telephone Corporation Limited (BTCL while billing for their call charges, imprints a date on the bill form b which the telephone users must pay their bills to avoid disconnection o their telephone lines. It has been observed over the years that a larg majority of the customers fail to pay their bills by the date specified b BTCL. The company makes an attempt to see the extent of this delay an selects a random sample of 65 customers and records the delay times i days in making their payment. The data were as follows:

22	29	16	15	18	17	12	13	17	16	15	19	17
10	21	15	14	17	18	12	20	14	16	15	16	20
22	14	25	19	23	15	19	18	23	22	16	16	19
13	18	24	24	26	13	18	17	15	24	15	17	14
18	17	21	16	21	25	19	20	27	16	17	16	21

Construct a stem and leaf plot and comment on the skewness of t distribution.

Solution: Note that the payment delays range from 10 days to 29 da from the date specified by BTCL and are expressed in whole numbers. construct a stem and leaf plot, we will use these whole numbers as ste Noting that a whole number, say 18, can be written as 18.0, we will use tenth-place digit, 0, as the leaf value. Therefore the stem and leaf plot w be as follows:

Stem	leaf
10	0
11	
12	00
13	000
14	0000
15	0000000
16	000000000
17	00000000
18	000000
19	00000
20	000
21	0000
22	000
23	00
24	000
25	00
26	0
27	0
28	
29	0

Key: 18|0 means18, 20|0 means 20.

Looking at the display, we see that the 'tail, of the distribution of the higher delay times in payments is longer than the 'tail' of the distribution of the smaller delay times. This shows that a few of the payment times are somewhat larger than the rest of the payments times. Here we say that the distribution is positively skewed. Given the above data, we can form a frequency distribution of the following form:

Class boundary	Frequency
09.5–12.5	3
12.5–15.5	14
15.5–18.5	23
18.5–21.5	12
21.5–24.5	8
24.5–27.5	4
27.5–30.5	1
Total	65

When comparison of two sets of data is desired, we can draw stem and leaf diagram by putting two data sets side by side. Such a plot can be called

back to back stem and leaf plot. Here is an example that displays the back to back plot. Back to back plot is in essence a type of diagram that can be compared with population pyramid which is most commonly used to represent the age structure of a population by sex.

Example 2.18: Two groups of students Group A and Group B obtained the following scores out of 100 in an examination. Display the data by a back to back stem and leaf plot.

Scores for group A:

22, 35, 47, 55, 58, 55, 63, 65, 66, 68, 69, 61, 71
45, 46, 52, 75, 73, 72, 81, 82, 85, 86, 94, 91, 67

Scores for group B:

90, 91, 92, 94, 21, 34, 34, 40, 41,
44, 45, 51, 52, 54, 55, 57, 59, 64,
64, 65, 66, 66, 67, 67, 67, 78, 73,
75, 77, 79, 71, 81, 82, 83, 83, 84.

The diagram is as follows:

Group A		Group B
Leaf	Stem	Leaf
2	2	1
5	3	44
765	4	0145
8552	5	124579
9876531	6	44566777
5321	7	135789
6521	8	12334
41	9	0124

Key: 6|4 means 64

You can now compare the performance of the two groups in terms of the marks obtained by the students. You can also combine the two groups and draw a stem and leaf plot for all the students.

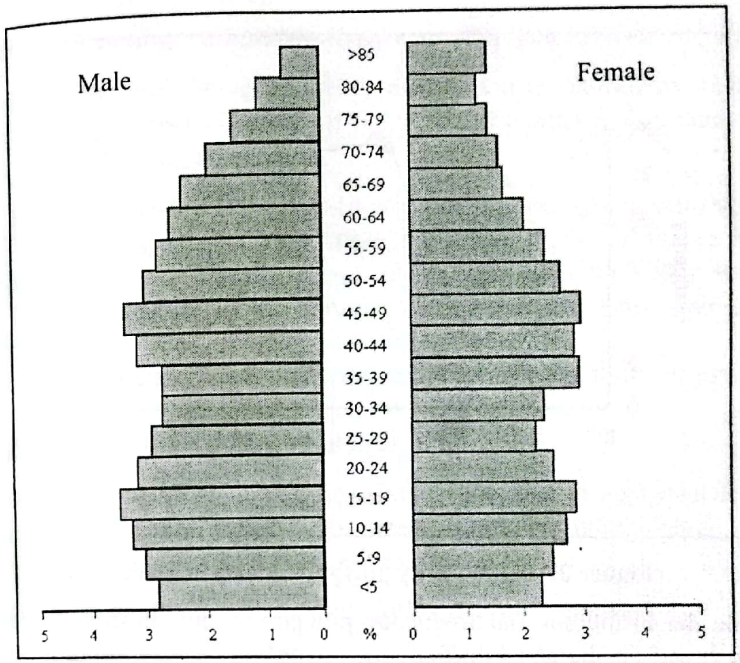

Figure 2.13: A typical age-sex pyramid for Bangladesh population 2010

2.16.5 Frequency Polygon

A **frequency polygon** provides an alternative to a histogram as a way of graphically presenting a distribution of a continuous variable. The presentation involves placing the mid-values on the horizontal axis and the frequencies on the vertical axis. However, instead of using rectangles, as with the histogram, we find the class mid-points on the horizontal axis and then plot points directly above the class mid-points at a height corresponding to the frequency of the class. Classes of zero frequency are added at each end of the frequency distribution so that the frequency polygon touches the horizontal axis at both ends of the graph. This makes the frequency polygon a close figure. The frequency polygon is then formed by connecting the points with straight lines. The frequency polygon for the frequency distribution of expenditure of the elementary school of the students shown in Table 2.14 is displayed in Figure 2.14.

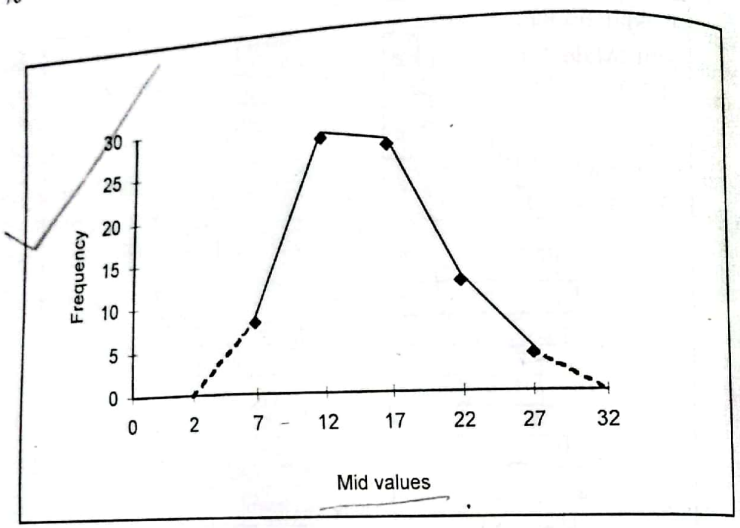

Figure 2.14: Frequency polygon for data in Table 2.14

While the histogram and frequency polygon in our illustrated examples were based on the absolute frequency distribution, they could have been based just as easily on the relative frequency distribution. The graphical presentations based on the relative frequency distributions would have looked identical to those constructed with absolute frequencies with the exception that the vertical axis would have been measured in terms of relative frequency rather than the absolute/actual frequency.

The histogram and frequency polygon are equally good techniques for presenting continuous data. A histogram is more often used when single distributions are presented, while the frequency polygon is largely used for comparison of two or more distributions.

In a continuous frequency distribution, if the number of observations is large, then the number of classes can be increased so as to make the magnitude of class intervals smaller and smaller. And in such a case, the graph representing the distribution will approach a smooth curve. The same is true in the case of a frequency polygon too. Such a curve is called a **frequency curve**. That is, when a frequency polygon is smoothed; the resulting curve will be a frequency curve.

2.16.6 Constructing Cumulative Frequency Polygon (ogive)

A graph of the cumulative frequency distribution or cumulative relative frequency distribution is called an **ogive**. To construct a less than type ogive, follow the following steps:

(a) Put the upper class limits (precisely the upper boundaries) on the horizontal axis and cumulative frequencies on the vertical axis.
(b) Plot a point directly above each upper class limit at a height corresponding to the cumulative frequency at that upper class limit.
(c) Plot one additional point above the lower class limit for the first class at a height of zero.
(d) Connect these points by straight lines.

The straight lines drawn in step (d) above allow one to approximate the cumulative frequency between the class limits by interpolating. The resulting graph is a **less than type ogive**.

To construct a **more than type ogive**, follow the steps below:

(a) Plot a point against each lower class limit at a height corresponding to the cumulative frequency at that lower class limit.
(b) Plot an additional point above the upper class limit for the terminal class at a height of zero.
(c) Connect these points by straight lines.

The resulting graph is a **more than type** ogive.

Example 2.19: The following table is constructed from data collected on the life length of 40 rats in years for a laboratory experiment. Display the data by a less than type and a more than type ogive.

Life length (in years)	Number of rats
1.45–1.95	2
1.95–2.45	1
2.45–2.95	4
2.95–3.45	15
3.45–3.95	10
3.95–4.45	5
4.45–4.95	3
Total	40

Table 2.16 is constructed to draw the required ogives and the resulting ogives are sketched in Figures 2.15 and 2.16.

The ogive or cumulative frequency polygon has the advantage of providing a convenient way to estimate the median and the percentiles of a sample, which will be discussed in the next chapter. In addition, it has the advantage that the number of items between two values can be readily ascertained. The ogive allows us to see how many observations in a data set fall at or below a given point on the scale. This is most useful when we have a distribution of scores and we are interested in finding out how one score compares to the rest of the scores.

Table 2.16: Cumulative frequency distributions for less than and more than type ogives based on the rat life data.

(a) Less than type distribution

Age	Cumulative frequency
Less than 1.45	0
Less than 1.95	2
Less than 2.45	3
Less than 2.95	7
Less than 3.45	22
Less than 3.95	32
Less than 4.45	37
Less than 4.95	40

(b) More than type distribution

Age	Decumulative frequency
1.45 or more	40
1.95 or more	38
2.45 or more	37
2.95 or more	33
3.45 or more	18
3.95 or more	8
4.45 or more	3
4.95 or more	0

It is also possible to present the two graphs (more than type and less than type) together in a single graph on the same scale. Such a graph is helpful in identifying the central value of a distribution. This is discussed in Chapter 3.

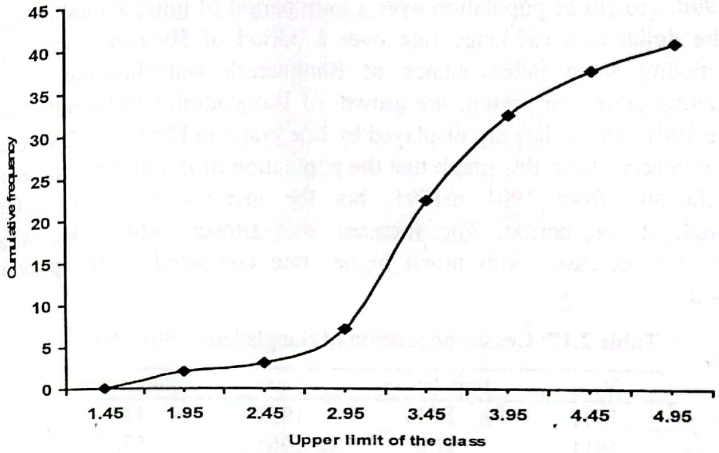

Figure 2.15: Less than type ogive for longevity data in Table 2.16

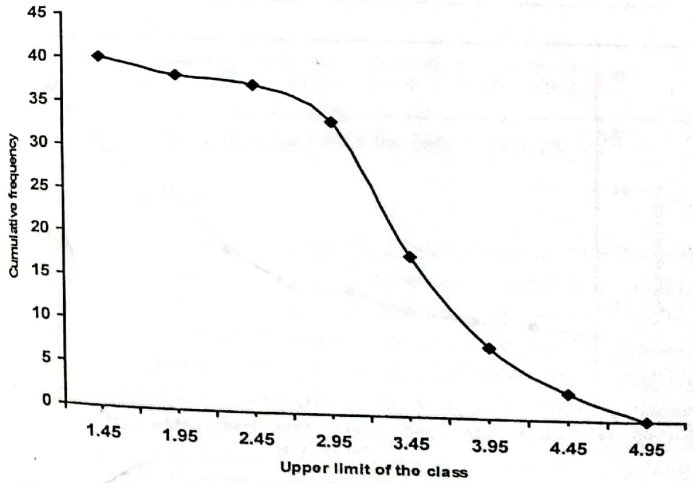

Figure 2.16: More than type ogive for longevity data in Table 2.16

2.16.6 Line Graph

A line graph is particularly useful for numerical data if we wish to show time series data such as production of jute for a period of 20 years, export

of raw materials from Bangladesh for a period of say 40 years from 1950 to 1990, growth of population over a long period of time, annual changes in the dollar-taka exchange rate over a period of 50 years, price of a commodity since independence of Bangladesh and the like. In the following table, we present the growth of Bangladesh census population since 1901. These data are displayed by line graph in Figure 2.17. One can easily observe from this graph that the population of Bangladesh increased significantly from 1901 to1991, but the increase was not uniform throughout the period. The increase was slower until 1941, which thereafter increased with much higher rate compared to the previous period.

Table 2.17: Census population of Bangladesh: 1901-1991

Year	Population	Year	Population
1901	28.9	1951	44.2
1911	31.6	1961	55.2
1921	33.2	1974	76.4
1931	35.6	1981	89.9
1941	42.0	1991	111.5

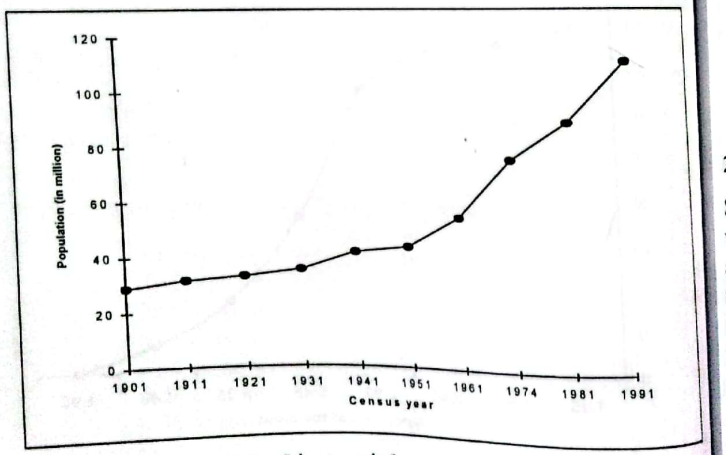

Figure 2.17: Line graph for data in Table 2.17

Example 2.20: The data below are the remittances (in million US dollars) from UK and USA of Bangladeshi expatriates for the period FY2001–FY2007. These data are compared graphically in Figure 2.18 below.

Year	Remittances from USA	Remittances from UK
2000–01	225.6	55.7
2001–02	356.2	103.3
2002–03	458.1	220.2
2003–04	467.8	297.5
2004–05	557.3	375.8
2005–06	760.7	555.7
2006–07	930.3	886.9

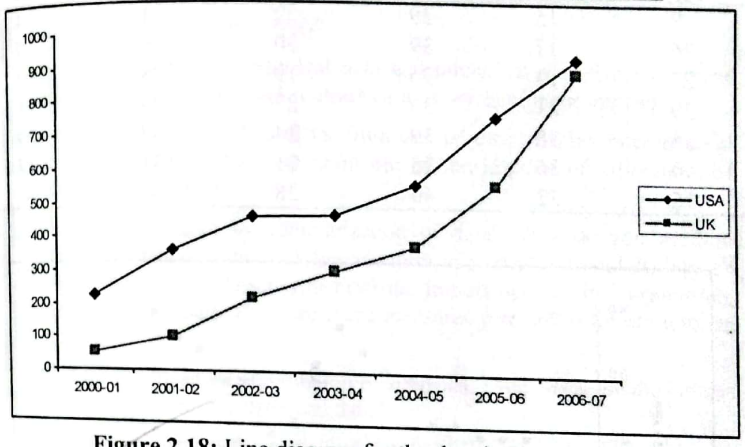

Figure 2.18: Line diagram for the data in Example 2.20

2.16.7 Scatter Diagram

Scatter diagrams are useful for displaying information on two quantitative variables, which are believed to be inter-related. Height and weight, age and height, income and expenditure, are examples of some of the data sets that are assumed to be related to each other. Such data can be displayed by scatter diagrams. Data in Table 2.18 below relate to the age at marriage of 20 couples obtained in a survey conducted in an area. A scatter diagram in Figure 2.17 displays these data. The diagram clearly demonstrates that as age of the husband increases, the wife's age also increases, thus implying that a positive relationship exists between husband's age and wife's age.

The list of graphs and diagrams presented in this chapter is not necessarily exhaustive. There are numerous types of diagrams and charts, which are frequently used to present statistical data. The preference for a particular type of graphs or diagrams over the other largely depends on the nature of

data and the choice and taste of the users and the presenters. With the advancement of computer technology, a good number of graphic packages are now available, which are very easy to handle with. Harvard Graphics is a commonly used graphic package. Excel, Minitab and Microsoft word also have graphic facilities.

Table 2.18: Age at first marriage of 20 couples in years

Husband's age	Wife's age	Husband's age	Wife's age	Husband's age	Wife's age
19	15	39	32	21	14
26	17	39	30	26	19
27	19	26	19	28	19
29	21	33	27	29	21
36	28	39	34	37	29
40	36	25	21	31	27
35	32	40	38	-	-

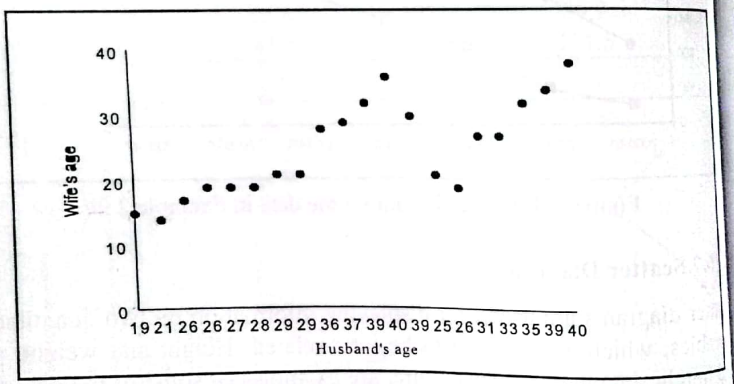

Figure 2.19: Scatter diagram for the data in Table 2.18

2.17 USE OF COMPUTER IN SUMMARIZING DATA

Use of computer as an aid to present statistical data has increased considerably in recent years. Many statistical packages and softwares such as SPSS, SAS, R, and STRATA are some of the statistical packages that have enormous facilities and options to prepare tables of any dimension, charts and diagrams in addition to their extensive use in statistical data analysis. Microsoft Excel and even Microsoft word have also such facilities.

EXERCISES 2

1. What do you mean by statistical data? How are they generated? Why do you need statistical data at all? What are the usual sources of statistical data in Bangladesh?
2. What is a variable? How do you distinguish a dependent variable from an independent variable, a discrete variable from a continuous variable, a dummy variable from a categorical variable? What is an attribute? Give example in each case.
3. Define data. How are they collected? What are the various types of data available for statistical analysis?
4. Distinguish between categorical data and numerical data. Describe their various types. Compare and contrast various levels of measurement.
5. In what way nominal data differ from ordinal data? Define interval level data and ratio type data and point out the basic point of differences of these two types of data.
6. What do you mean by summarization of data? Why do you need to summarize statistical data? Define the term 'classification' and distinguish it from 'tabulation'. Discuss in brief the importance of ratio, proportion, percentage and rates. How are these measures formed? Illustrate with an example.
7. What do you mean by presentation of statistical data? What are the various methods of presenting statistical data?
8. How do you present the statistical data when the data are (i) categorical and (ii) numerical?
9. Define a frequency distribution. What is the purpose of constructing a frequency distribution? Set out the important steps involved in the construction of a frequency distribution from raw data.
10. What is frequency table? Distinguish between a univariate frequency table and bivariate frequency table. What purposes do they serve?
11. What are the various types of charts and diagrams used in presenting statistical data? State the importance and utility of these devices. What are advantages of diagrammatic presentation of statistical data over tabular presentation?
12. What are the diagrams that would be suitable for representing data generated through nominal and ordinal level of measurement? How would you present numerical data graphically?
13. Give a brief description of various graphs and diagrams for representing statistical data. State the advantages and disadvantages of these methods giving illustrations whenever possible.

14. What is an array? Define classification and distinguish it from tabulati... How will you proceed to classify the observations made and what fact... will you take into consideration in tabulating them?

15. Define the following in connection with a frequency table:

(a) Class interval (b) Class mark or class mid-point (c) Class limits ... Class boundaries and (e) Class frequency.

16. What types of diagram would you prefer to represent a freque... distribution of variable measured on interval scale? Compare a histogr... with a bar diagram and distinguish them from a frequency polygon.

17. What are the main considerations that lead to the choice of class interv... in constructing a frequency distribution from raw data? How do the cl... limits differ from class boundaries? How do you determine the class ma... of a distribution?

18. The following figures refer to the employment of the Bangladesh nation... abroad in 1996 as given by the Bangladesh Bureau of Statistics:

Profession	Number employed
Professionals	3188
Skilled	64301
Semi-skilled	34689

Use the data to construct a (i) pie chart and a (ii) horizontal bar diagram

19. Describe how an ogive can be constructed for a frequency distributi... How does a less than type ogive differ from a more than type ogi... Illustrate with an example.

20. Suppose that 100 students are enrolled in a statistics class and ... following are the test scores received by them:

77 44 49 33 38 33 76 55 68 39 44 59 36 55 47 61 53 32 65 51
29 41 32 45 83 58 73 47 40 26 59 43 66 44 41 25 39 72 37 55
34 47 66 53 55 58 49 45 61 41 55 92 83 77 45 62 45 36 78 48
54 50 51 66 80 73 57 61 56 50 45 82 71 48 46 69 38 72 56 64
38 45 51 44 41 68 45 92 43 12 37 16 44 57 63 71 40 64 57 51

(a) Form an array in ascending order.

(b) Organize the data in 10 classes such as 10.5 to 19.5, 19.5 to 28.5, 8... to 91.5 and 91.5 to 100.5 and complete the frequency distribution.

(c) Using the above data, draw a histogram, a frequency polygon and ... ogive.

(d) Draw a stem and leaf diagram to represent the data.

21. The number of speeding tickets issued per day in a certain city are group... into a table having the classes 2–7, 8–13, 14–19, 20–25, 26–31 a... 32–37. Find

(a) the lower class limits;
(b) the upper class limits;
(c) the class boundaries;
(d) the class marks;
(e) the class interval.

22. The following is a distribution of the final examination scores which 200 students obtained in a 3-week course in Economics.

(a) Convert this distribution into a percentage distribution.
(b) Convert this distribution into a cumulative "less than type" and a "more than type" distributions.
(c) Draw a frequency polygon and an ogive (both less than and more than type) of the original distribution.
(d) Draw a histogram of the original distribution.
(e) Obtain a relative frequency distribution and hence a histogram.

Scores	Number of students
0–19	24
20–39	55
40–59	76
60–79	32
80–99	13
Total	200

23. The following table shows the percent distribution of female population of Bangladesh in five administrative divisions as obtained in 1993-94 Bangladesh Demographic and Health Survey (BDHS).

Division	Population (%)
Barisal	6.3
Chittagong	26.0
Dhaka	30.6
Khulna	12.8
Rajshai	24.3

Construct a pie chart of this distribution.

24. A sample of 9640 ever-married women in 1993-94 Bangladesh Demographic and Health Survey (BDHS) showed the following distribution according to their level of education:

Education	Number of women
No education	5598
Primary incomplete	1681
Primary complete	922
Secondary/Higher	1439

Use pie chart and bar diagram to represent the data

25. A study of air pollution in a city yielded the following daily readings the concentration of sulfur dioxide (in parts per million):

0.04 0.14 0.17 0.11 0.18 0.20 0.13 0.17 0.10 0.07
0.15 0.09 0.10 0.16 0.12 0.27 0.10 0.12 0.05 0.06
0.08 0.15 0.05 0.09 0.11 0.13 0.07 0.14 0.09 0.17
0.10 0.08 0.14 0.02 0.14 0.08 0.11 0.08 0.01 0.18
0.11 0.13 0.06 0.19 0.12 0.15 0.01 0.05 0.11 0.04
0.14 0.17 0.11 0.18 0.20 0.13 0.17 0.10 0.07 0.15
0.09 0.14 0.19 0.18 0.19 0.21 0.16 0.20 0.16 0.18

(a) Group these data into a frequency table having the classes 0.00–0.
0.05–0.09, 0.10–0.14, 0.15–0.19, 0.20–0.24, and 0.25–0.29.
(b) Convert the distribution in (a) into a cumulative "less tha
percentage distribution and hence an ogive
(c) Construct a histogram of the distribution obtained in (a)

26. The following data refer to the trend in infant, child and under-f mortality rates per 1000 live births for the periods 1979-83, 1984-88 a 1989-93. Use a suitable diagram to show the trend in mortality rates the period under reference.

Reference period	Infant mortality rate	Child Mortality rate	Under-five mortality rate
1979-83	116.5	71.8	180.0
1984-88	111.9	59.0	164.3
1989-93	87.4	50.1	133.1

27. Table below shows the age specific fertility rates (ASFRs) per 1 women in Bangladesh for the year 1993-94 as obtained in 1996- BDHS.

Age group	ASFRs
15-19	147
20-24	192
25-29	150
30-34	96
35-39	44
40-44	18
45-49	6

Use an appropriate diagram to represent the ASFRs.

28. The **Bhorer Kagoj** (A daily in Bangladesh) reported the following figu on number of suicides that were committed in Comilla and Jhenaid districts over the period 1993-1997.

Year	Comilla	Jhenaidah
1993	332	901
1994	336	707
1995	385	725
1996	372	808
1997	415	726
Total	1840	3867

Display the data in a multiple bar chart and comment on the trend in suicides of these two districts. What other alternative diagrams could be used to represent the data?

29. The following data refer to the ages of 80 children recorded in months, obtained in a nutrition survey by a research team:

11	20	16	10	22	07	13	14	23	11
15	05	10	12	14	18	15	13	11	06
26	16	14	18	27	29	23	33	17	24
17	18	21	11	10	12	13	18	15	16
22	14	15	08	09	17	15	14	10	11
10	13	12	17	16	20	19	10	07	12
16	18	21	19	19	15	19	13	19	18
06	11	15	20	13	24	24	17	12	25

(a) Construct a frequency distribution displaying the class intervals, class marks and class boundaries.

(b) Draw a histogram of the data.

(c) Set up a relative frequency distribution.

(d) Construct a frequency polygon and an ogive.

(e) Present the data by a stem and leaf plot.

From the frequency distribution, find the number of children who were older than 15 months. How many of them were of age 25 months or more? How many were between 5 and 14 months?

30. In which scale of measurement will you put the following variables?

family size, Religion, Race, level of satisfaction, day temperature, length of schooling, happiness, age, room number, telephone number, opinion, hair color, work status of women.

Which type of graphs and diagrams would be suitable for representing the above variables?

31. In a study of sources of air pollutant emissions, 61% of nitrogen oxides were attributed to transportation, 24% were attributed to stationary fuel combustion sources, 8% were attributed to industrial processes, and 7% were attributed to other miscellaneous sources. Illustrate the results of this study with a pie chart.

32. The following values represent the numbers of participants in a study on recreation:

Baseball	190
Rowing	259
Golf	83
Sailing	291
Skating	140
Swimming	638
Tennis	148
Football	740
Volley	145
Cricket	422

Draw a horizontal bar diagram to illustrate the result of this study.

33. Define cross-tabulation, dummy table, bi-variate table and a two-way table.

One hundred individuals were interviewed and were classified according to their age, sex and level of education. The variables were recorded as follows:

Age: in complete years (18, 19, ...50+)

Sex: Male (M) and Female (F)

Level of education: No schooling (N), Primary (P) and Above primary (A)

The information was recorded in the order (Sex, Age, and Level of education). Thus, an entry (M, 35, P) for example, represents that the person so interviewed was a male, 35 years old and had primary level of education. The data obtained were as follows:

(F 20 P), (M 30 P) , (F 24 P), (M 30 A) , (M 20 P), (M 40 P) , (M 27 N), (M 34 P), (M 20 P), (M 29 A), (M 26 P), (F 37 P), (M 20 P), (M 44 A), (F 39 P), (M 26 P), (M 31 N), (M 44 A), (F 22 A), (M 26 N), (F 55 P), (F 53 N), (M 20 P), (M 29 P), (M 30 P), (F 61 P), (M 30 A) , (F 29 P), (M 40 P) , (M 27 N), (M 36 N), (F 47 P), (F 59 A), (M 26 A), (M 37 A), (F 20 P), (M 47 N), (F 53 A), (M 29 P), (M 37 A), (F 49 A), (M 22 N), (M 24 N), (M 50 N), (F 53 N), (M 46 N), (M 37 A), (M 49 A), (M 42 N), (M 27 N), (M 33 P), (F 30 N) , (F 28 P), (F 30 N) , (M 20 P), (M 40 P) , (M 37 A), (M 34 P), (M 20 P), (F 29 A), (M 46 P), (F 47 P), (M 25 A), (M 44 N), (M 39 P), (M 26 A), (F 31 N), (M 49 A), (F 21 A), (F 26 N), (F 55 P), (F 53 N), (M 20 P), (M 29 P), (F 30 P) , (F 61 P), (M 30 A) , (M 29 P), (M 40 P) , (M 27 N), (F 36 N), (F 47 P), (F 59 A), (M 26 A), (M 37 A), (F 20 P), (M 47 P), (F 53 P), (M 29 P), (F 37 A), (F 59 A), (M 22 P), (F 24 N), (M 50 N), (F 53 N), (M 47 N), (M 47 A), (M 29 A), (F 42 A), (M 27 N), (M 55 P)

Prepare a table of the following format to display the age, sex and level of education of the respondents. How many of the males are above primary? How many are primary or below taking both sexes together? Compute the percentage of females 30 years and over with primary or higher level of education.

	Males			Females			Both sexes		
Age	N	P	A	N	P	A	N	P	A
< 20									
20-29									
30-39									
40-49									
50+									

34. A study on alleged grounds for marriage breakdown in Bangladesh during 1998 produced the following percentage values based on 6000 cases:

Grounds	Percentage of grounds cited
Adultery	37.4
Dowry	13.8
Physical cruelty	23.8
Mental cruelty	16.5
Addiction to alcohol	2.7
Separation for long time	5.0
Others	0.8

Display the data by an appropriate diagram. If the total number of marriages is 500, how many marriages ended in divorce? Due to dowry? Physical cruelty?

35. The following table shows the exports of principal commodities (in million taka) from Bangladesh during 1990-94. Present the data by a suitable diagram:

Commodities	1990-91	1991-92	1992-93	1993-94
Prawn & shrimp	5017	5359	6967	9105
Tea	1544	1290	1555	1670
Raw jute	3231	3474	2795	1896
Jute yarn	1049	1091	1048	1387
Leather products	4422	4981	5274	6111
RMG	29941	39770	51117	53070
Handicraft	146	180	181	157