CHAPTER

2

SUMMARIZING AND PRESENTING DATA

LEARNING OUTCOMES

After completion of this chapter, you will learn

- Different ways and means to represent statistical data;
- To construct frequency tables for nominal, ordinal, discrete and continuous data;
- To present univariate and bivariate data in tabular and graphical forms;
- To employ stem and leaf plot for representing statistical data.

2.1 INTRODUCTION

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. After collecting and coding the desired data by the researchers the first step is to classify and tabulate the data. The classification and tabulation provide a clear picture of the collected data and on that basis the further processing is decided. It is a kind of sorting operation and can be repeated as many times as there are possible bases of classification. In another words we can say that it is a process of separating likes from the unlike with a view to present a condensed and homogeneous picture. Technically classification is a method of arranging data in groups or classes according to their similar attributes. Connor (1997) defined classification as: "the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals". Usually the data can be collected through questionnaire, or schedules which need to be consolidated further for the purpose of analysis and interpretation. We have discussed in length this aspect in our previous chapter. We can include a huge volume of data in a simple statistical table and one can easily get an overview about the sample by observing the statistical table rather than the raw data. It

is also essential to tabulate the data to construct **charts** also called **diagrams** and **graphs**. The need to summarizing and classifying the collected data is

- To simplify the complex data
- To economize space
- To identify omissions and errors
- To facilitate comparison
- To facilitate statistical analysis
- To depict trend
- To use it as future reference

The following example is designed to illustrate the various aspects of the need for tabulating and summarizing data.

Example 2.1: Suppose a sample of 50 workers was drawn from Beximco Pharmaceutical Company (BPC), which employed 500 workers. The researcher collected such data as the employees' age, level of education, level of skill, wage earning, and religious affiliation by directly interviewing the employees. These are some of the personal or background characteristics of the employees which the researcher should consider in order to meet the objectives of a research in any discipline. Clearly, the information collected are both qualitative and quantitative in nature. The accompanying table (Table 2.1) displays these data.

Having obtained the data, some of 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 have secondary level of education?
(c) Are most of the workers unskilled?
(d) What percentage of the workers belongs to minority group?
(e) Average number of days a worker remain absent from their work?
(f) Can we classify them by level of education and family size?

The answers to the above questions can be given simply by counting the event of interest that appears 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 done. Tabular and graphical procedures provide ways and means of neatly organizing and describing the data such that they are more easily used and interpreted. The concept of frequency distribution is introduced as a tabular method of summarizing data. We shall then show how this frequency distribution can be displayed graphically employing a number of diagrams, charts, plots and curves. It thus appears that there are largely two ways of presenting statistical data. They are

- Tabular presentation;
- Graphical presentation

We will discuss both of these methods in greater details in sections that follow.

Table 2.1: Raw Data on Background Characteristics of BPC Workers

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

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

2.2.1 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**.

***Definition 2.1**: Measurements in which the names or classifications are used to divide data into separate and distinct categories are **nominal measurements**.*

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.

2.2.2 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,

(a) The level of education (MA. BA, etc.).
(b) Official position (manager, deputy manager, accountant).
(c) Socio-economic status (high, medium low).

(d) Class performance (outstanding, very good, good, poor).
(e) Monthly frequency of visits of a physician in a clinic (frequently, occasionally, rarely, never).
(f) 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 the **nominal level of measurement**.

Definition 2.2: Measurements that rank observations into categories with a meaningful order are ordinal levels of measurements.

The chief properties of ordinal level of measurement are:

1. The categories are distinct, mutually exclusive and exhaustive.
2. The categories are possible to be ranked or ordered.
3. The distance or differences from one category to the other category is not necessarily constant.

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.

2.2.3 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).

Definition 2.3: Measurements on a numerical scale in which the value of zero is arbitrary but the difference between values is important are interval levels of measurement.

The interval level of data has the following properties:

(a) The data classifications 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.

In sum, a set of data in which we can form differences of measurements, but cannot multiply or divide, represents **interval data.**

2.2.4 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 are 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.

Definition 2.4: Numerical measurements in which zero is a meaningful value and is the difference between the values is important are ratio level of measurement.

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 below:

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 algorithm 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 2.2: Characteristics of Different Levels of Measurement: A Comparison

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 number 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

2.3 VARIABLE AND ATTRIBUTE

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. As we will see levels of measurement, discussed earlier, tell us how precisely variables are recorded or measured

2.3.1 The Variable

We begin this section with the definition of variable

***Definition 2.5 : 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.*

There are two major types of variable:

(a) Qualitative or categorical variable and
(b) Quantitative or metric variable

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 the nominal level of measurement, for which it is also called **nominal variable.**

Definition 2.6: A qualitative variable, also called categorical 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 more examples of 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).
- Blood group type (O, A, B, AB)

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

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. Quantitative variables are also called **metric variables**. Other examples of quantitative variables, among others, are

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

Definition 2.7: A quantitative variable, also called metric variable, is one for which the resulting observations are numeric and thus possesses a natural ordering. It is also called numerical variable.

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

Definition 2.8: A variable that 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 2015–2020
- 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 2.9: 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.

The types of variables we have discussed so far may be displayed in the form of a flowchart as shown below in Figure 2.1:

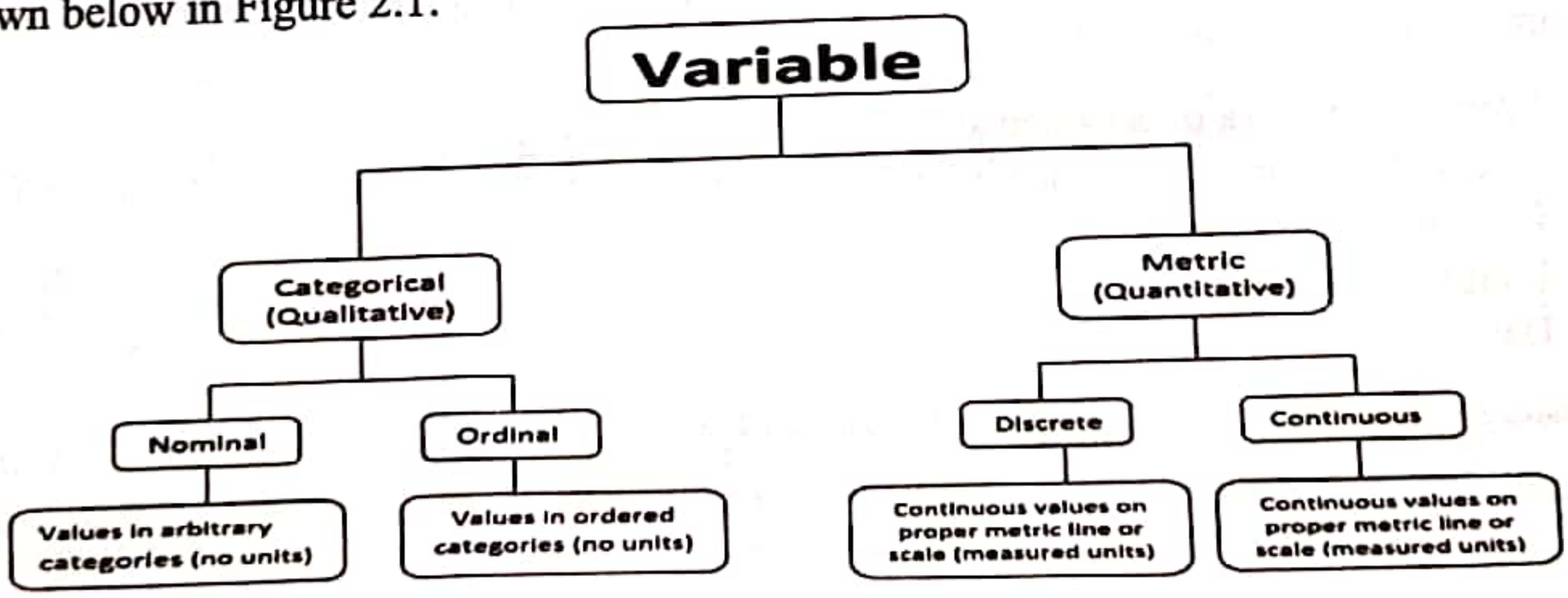


Figure 2.1: Flowchart showing the types of a variable

2.3.2 The 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–owns a mobile set, smokes or not, or holds an opinion on certain political issues–these characteristics may be called **attributes**. Quantification thus lies in counting how many individuals possess this attribute and how many do not and the proportion or percentage with these attributes provide a useful description of the population or sample.

***Definition 2.10: An attribute** is a quality, character or characteristic to someone or something.*

2.3.3 The 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.

Definition 2.11: A constant is a quantity which has a fixed value and does not change over time, circumstances or occasions.

2..4 FREQUENCY DISTRIBUTION AND ITS CONSTRUCTION

Before we construct a frequency distribution, it is necessary to define the term 'frequency' and frequency distribution. **Frequency,** also called **class frequency** refers to the number of observations falling within the confines of a particular class. It is the number of measurements or counts in a category or class.

The categories or classes along with the frequencies together constitute what is known as the **frequency distribution**. 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.

Definition 2.12: A frequency distribution is a set of mutually exclusive classes together with the frequencies of occurrences of the values in each class in a given set of data.

In most cases, a frequency distribution is presented in a tabular form, showing the frequency of measurements, observations or cases in each of the several non-over-lapping or mutually exclusive classes.

A frequency distribution may also be displayed graphically or by some statements or rules for pairing a class of observations with its frequency.

Frequency distribution can be constructed for both categorical and numerical data. Numerical data when grouped and organized in a frequency distribution results in **grouped frequency distributions.**

In contrast, for **ungrouped data**, every observed value of a variable is listed. This gives rise to what we refer to as **ungrouped frequency distributions.**

2.4.1 Desirable Features of a Frequency Table

While construction a frequency distribution table, we should follow certain rules and principles that will result in a good table. A good table must contain the following parts:

1. Table number
2. Title of the table
3. The box head and the column caption
4. The row captions and the stub
5. The body of the table
6. Prefatory notes

7. Footnotes
8. Source notes

1. **Table number**: Each table should be numbered consecutively. For a book or a scientific report, you might have several chapters. In such cases, you might prefer to put table numbers to identify a chapter. For example, table number 2 of chapter 3 may be labeled Table 3.2, table 4 of chapter 5 as Table 5.4 and so on.
2. **Title of the table:** Every table must have a suitable title that succinctly describes the contents of the table. Table title should be as short as possible. The title must explain the contents of the table and communicate to us the "**what**", "**where**", "**how**" and "**when**" messages of the data.
3. **The column captions and box head**: The column captions refer to the vertical headings explaining what the columns represent. The spaces where these columns headings are written are called the box head.
4. **The row captions and the stub:** The horizontal headings and the sub headings of the row are called row captions and the space where these rows headings are placed is called the stub.
5. **The body of the table:** This is the main part of the table which contains the numerical information classified with respect to row and column captions.
6. **Prefatory notes:** A statement given below the title and enclosed in brackets usually describes the units of statement and is called the prefatory notes.
7. **Footnotes:** Anything presented in a table which the readers may find difficult to understand from the title, or captions, should be explained in footnotes.
8. **Source notes**: The source notes are given at the end of the table indicating the source the information has been taken from. It includes, among others the compiling agency, publication etc.

The general sketch of a table displaying its different parts appears below as a dummy table:

Table 2.3: Format of a Statistical Table (Dummy Table)

(Prefatory Notes)

	BOXHEAD			
	COLUMN CAPTIONS			
ROW CAPTIONS	A	B	C	D
STUB	**THE BODY**			

Foot Notes……

Source Notes……

We discuss below how a frequency distribution table or simply a frequency distribution is constructed practically.

2.5 CONSTRUCTION OF FREQUENCY TABLE: CATEGORICAL DATA

2.5.1 Construction of Univariate Tables

The construction of a univariate frequency table for categorical (also called qualitative) data with a single categorical variable consists essentially of the following steps:

(a) Choose the category into which the data are to be grouped.
(b) Sort or tally the data into appropriate categories.
(c) Count the number of items or measurements falling in each category
(d) Display the results in a table.
(e) The resulting table represents the desired frequency distribution.

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

Example 2.2: A market researcher conducted an inventory of 25 firms and categorized them as 'large', 'medium' and 'small' depending on the investment, floor space and number of employees. The categories of the 25 listed firms were as follows:

Small	large	small	medium	large
medium	large	small	large	medium
large	small	large	medium	medium
medium	large	large	medium	medium
medium	small	small	medium	medium

Present the data in a frequency distribution.

Solution: The data pertain to three distinct categories of the firm: 'large', 'medium' and 'small'. 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 firm in the order is 'small'. The category, 'small' appears in the first column of the table. Put a tally mark against the firm size 'small', which is simply a left-slashed off-diagonal stroke (/).
- Move to the next entry, which is 'large'. Enter this again by a tally mark against the category 'large' appearing in the first raw as before.
- Repeat the above process until you have entered all the 25 firms appearing in the observed set.
- 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 as under.

Firm size	Tally	Count
Large	𝍸 ///	8
Medium	𝍸 𝍸 //	12
Small	𝍸	5

- The resulting table that appears below represents a frequency distribution of the firm size.

Table 2.4: Frequency Distribution of Firm Size

Firm size	Number of firms	Percent
Large	8	32.0
Medium	12	48.0
Small	5	20.0
Total	**25**	**100.0**

The counts 8, 12, and 5 appearing in the second column of the table are the **class** or **category frequencies** for the categories large, medium and small respectively. The count 25 is the **total frequency**, which implies that we have listed 25 cases. Since the data are grouped into non-numerical categories, the distribution is referred to as **qualitative** or **categorical distribution**. Note that the firm size here is an ordinally scaled variable. The usual analysis of this table is limited to only finding percentage values as shown in the last column of the table.

You can construct a similar frequency distribution table with religion as reported by the workers of BPC as shown in Table 2.1. The table with tally marks will be as follows

Religion	Tally marks	Count
Muslim	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	36
Hindu	𝍸 ////	9
Christian	𝍸	5

Once you have done this, the distribution in tabular form will appear as follows.

Table 2.5: Frequency Distribution of Workers by Religion

Religion	Number of workers	Percent
Muslim	36	72.0
Hindu	9	18.0
Christian	5	10.0
Total	**50**	**100.0**

These tables are of immense importance to get a descriptive overview of the variable under study in terms of the composition of the population or sample. The simplest way to achieve this is to form a percentage column as before. Looking at the last column of Table 2.4, for example, it is possible to say that of the total workers, 72 percent are Muslims and so on. Note that the percentages in the last column must add to 100.

2.5.2 Construction of Cross Tables

Frequently, we need to describe relationship between categorical or ordinal variables. Market research organizations, for example, describe attitudes toward products, measured on an ordinal scale, as a function of educational levels, social status measures, geographical areas, and other ordinal or categorical variables. Production analysts study relationships between departments or production lines and performance measures to determine reasons for product change, reasons for interruption of production, and quality of output. These situations are usually described by **cross tables** and portrayed by bar-type charts. The entries in such cross tables 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 worker by his/her education level and family size simultaneously. This attempt produced the following table:

Table 2.6: Family Size and Education Level of 50 Workers

	Family size			
Education	**Large**	**Medium**	**Small**	**Row total**
None	4 (8%)	6 (12%)	1 (2%)	**11 (22%)**
Primary	6 (12%)	8 (16%)	5 (10%)	**19 (38%)**
Higher	6 (12%)	10 (20%)	4 (8%)	**20 (40%)**
Column total	**16 (32%)**	**24 (48%)**	**10 (20%)**	**50 (100%)**

Note: The values in the parentheses represent the percentages

Table 2.6 as shown above, 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**. A cross-table with r rows and columns is referred to as an r by c table, written as $\boldsymbol{r \times c}$ **table.**

One can form a percentage distribution of this cross-table just by dividing each cell frequency by 50, the total frequency. These percentages are then placed beneath the cell frequencies under parentheses. The explanation of these cell frequencies is simple: Of the total 50 workers, 4 workers (8%) who come from large families are illiterate (have never gone to school). This percent is found as follows:

$$\text{Illiterate workers belonging to large families} = \frac{4}{50} \times 100 = 8\%$$

Similarly, 16 percent of the workers who come from medium sized families completed primary level of schooling.

$$\text{Primary level passed workers of medium sized families} = \frac{8}{50} \times 100 = 16\%$$

The table on family size and level of education just cited above 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 (**the problem**) and education is an independent variable (**the factor**). The totals in the columns and rows are called the **marginal frequencies.**

These frequencies constitute what we call 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. Obviously, the marginal distributions are uni-variate distributions. One such distribution with education is shown in Table 2.7 below.

Table 2.7: Marginal Distribution of Workers by Level of Education

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

A cross table may be of mixed-type in terms of variables. One variable may be categorical or ordinal in nature while the other a continuous variable. Here is an example of such a table (based on the BPC workers data in Table 2.1)

Table 2.8: Distribution of Workers by Their Age and Religion

	Religion			
Age	Muslim	Hindu	Christian	**Row total**
25–34	9	3	–	**12**
35–44	19	5	3	**27**
45–54	8	1	2	**11**
Column total	**36**	**9**	**5**	**50**

Example 2.3: A market research team conducted a survey among 350 boys and 250 girls of about the same age on their preference of cold drink available in the market. The result is shown in the accompanying table:

	Type of drink				
Sex	Coca-Cola	Sprite	Mountain Dew	7-Up	**Total**
Boys	95	90	85	80	**350**
Girls	65	50	80	55	**250**
Total	**160**	**140**	**165**	**135**	**600**

(a) Of the total children, how many of the girls prefer sprite?
(b) How many of the children like mountain dew?
(c) Among the total children, what is the percentage of boys who prefer 7-up?
(d) Among the total children, how many do prefer sprite?
(e) Of the total boys, what percentage of them likes mountain dew?
(f) Of those who prefer coco-cola, what is the percentage of boys?
(g) Construct marginal distributions for boys and girls and verify your answer with (d)

Solution: We answer the above questions below.

(a) Examine that, 50 girls out of 600 like sprite. In terms of percentage, it works out to 8.3 percent:

$$\frac{50}{600}\times 100 = 8.3\%$$

(b) A total of 165 children out of 600, irrespective of sex, like mountain dew. This results in 27.5 percent.

$$\frac{165}{600}\times 100 = 27.5\%$$

(c) Eighty boys out of the total children like 7-up. In percentage, it is 13.3:

$$\frac{80}{600}\times 100 = 13.3\%$$

(d) Of the 600 children, 140 or 23.3percent prefer sprite.

$$\frac{140}{600}\times 100 = 23.3\%$$

(e)There are 350 boys and 85 of them prefer mountain dew. The percentage being

$$\frac{85}{350}\times 100 = 24.3\%$$

(f) Of the total children 160 prefer coco-cola and 95 of them are boys. Hence percentage of boys in this category is

$$\frac{95}{160}\times 100 = 59.4\%$$

(g) The marginal distribution of drink types is shown below:

Table 2.9: Drink–type Liked by the Children

Drink type	No of children	Percent
Coca–Cola	160	26.7
Sprite	140	23.3
Mountain Dew	165	27.5
7–up	135	22.5
Total	**600**	**100.0**

As the marginal distribution of the children by drink-type shows, 140 children (second value in column 3) or 23.3 percent prefer sprite.

$$\frac{140}{600} \times 100 = 23.3\%$$

The marginal distribution of the children by sex is displayed in Table 2.10 below:

Table 2.10: Distribution of Children by Sex

Sex	No of children	Percent
Boy	350	58.3
Girl	250	417
Total	**600**	**100.0**

For more details, a combination of Table 2.9 and Table 2.10, along with the percentage values may be displayed in Table 2.11 below:

Table 2.11: Percentage Distribution of the Children by Sex and Cold Drink Preference

Drink type	No. boys	No. of girls	Row total
Coca–Cola	95	65	**160**
	(27.1%)	(26.0%)	**(26.7%)**
	(59.4%)	(40.6%)	**(100)**
Sprite	90	50	**140**
	(25.7%)	(20.0%)	**(23.3%)**
	(64.3%)	(35.7)	**(100)**
Mountain Dew	85	80	**165**
	(24.3%)	(32.0%)	**(27.5%)**
	(51.5%)	(48.5%)	**(100)**
7–up	80	55	**135**
	(22.9%)	(22.0%)	**(22.5%)**
	(59.3%)	(40.7%)	**(100)**
Column total	**350**	**250**	**600**
	(100)	**(100)**	**(100)**

The first percentages in each cell are based on the column totals, while the second percentages are based on the row totals. Let us explain this. We have 160 children who prefer coca-cola, of whom 95 (27.1%) are boys and the remaining 65 (26.0%) are girls. These percentage values are arrived as follows:

$$\text{Percentage of boys preferring Coca-Cola} = \frac{95}{350} \times 100 = 27.1\%$$

$$\text{Percentage of girls preferring Coca-Cola} = \frac{65}{250} \times 100 = 26.0\%$$

$$\text{Percentage of children preferring Coca-Cola} = \frac{160}{600} \times 100 = 26.7\%$$

Likewise, of 160 children who prefer coca-cola, 95 are boys and the remaining 65 are girls. Thus, of the 160 children:

$$\text{Percentage of boys preferring Coca-Cola} = \frac{95}{160} \times 100 = 59.4\%$$

$$\text{Percentage of girls preferring Coca-Cola} = \frac{65}{160} \times 100 = 40.6\%$$

These percentages appear beneath the first percentages in each cell.

With three variables, 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 2.6, 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

Religion		Family size			
		Large	Medium	Small	Sub total
Muslim	**Education**				
	None	4	3	1	**8**
	Primary	5	5	3	**13**
	Higher	4	7	4	**15**
	Sub total	**13**	**15**	**8**	**36**
Non-Muslim	**Education**				
	None	0	3	0	**3**
	Primary	1	3	2	**6**
	Higher	2	3	0	**5**
	Sub total	**3**	**9**	**2**	**14**
	Grand total	**16**	**24**	**10**	**50**

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

The table is simple to interpret. Of the 50 workers, 36 (72%) are Muslims and the remaining 14 (28%) are Non-Muslims. Moving along the row totals, we find that, of the 36 Muslim workers, 13 (36.1%) belong to large families, 15 (41.7%) belong to medium size families and the remaining 8 (22.2%) belong to small families. These percentages make 100 percent.

You can similarly compute the column percentages to say the percentages of workers attaining different levels of education. For example, moving along the columns, we see that 8 (22.2%) Muslim workers have never gone to school. Similarly 13 (36.1%) and 15 (41.7%) Muslim workers have respectively completed primary and secondary level of education. The table may similarly be interpreted for Non-Muslims.

The cell frequencies may similarly be explained. Of the 36 Muslim workers, who came from large families, 4 (11.1%) of them had never been to school.

For numeric (quantitative) variables, we can construct bi-variate table too. The accompanying table is such a table constructed from the wage and age data of Table 3.1. Since the data spread over a long range of 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 reclassified as 50–65, 66–81 and 82–97. The resulting bi-variate table is as follows:

Table 2.13: Distribution of Workers by Their Age and Wage

	Age in years			
Wage (Tk.)	25–34	35–44	45–54	**Row total**
50–65	2	6	5	**13**
66–81	6	13	5	**24**
82–97	4	8	1	**13**
Column 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 categorical.

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 are 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. 6 CONSTRUCTION OF FREQUENCY DISTRIBUTION: NUMERICAL DATA

The process of constructing 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. What is an array? We provide a formal definition of arrow below:

***Definition 2.13:** 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 below the process of constructing such an array with the wage data in Table 2.1 before we construct a frequency distribution.

Such an array as appears below 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 intervals with appropriate class widths. This results in a descent arrangement of the raw data. We have called this arrangement a **frequency distribution.** The construction of a frequency distribution is much facilitated once the data have been arranged in the array as mentioned above. The required data appear below once arranged as described above.

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

If we examine the data arranged above in an ordered array, we see that some values have been repeated more than once. For example, the values 65, 68, 69, 73, 82, 88, 93 each have 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**. Such a distribution is constructed below in a summary form below for an understanding of an ungrouped frequency distribution format to comprehend the idea.

Table 2.14: A Layout of an Ungrouped Frequency Distribution on Wage Data

Wage	Frequency
50	1
51	1
54	1
56	2
…	…
…	…
63	1
65	2
66	1
…	…
…	…
93	2
97	1
Total	**50**

As you see, the distribution in this form is 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, some 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 and then obtain a frequency distribution of grouped data.

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 the next higher class.

Class interval: The width (w) or length of the class, formed by the two boundary values is known

as the **class interval** or **class width**. 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..7 STEPS IN CONSTRUCTING GROUPED FREQUENCY 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 a numerical or quantitative 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 or a chart.
(e) The resulting table is our desired frequency distribution.

Determining the exact number of classes and choosing a class-width however involves a tradeoff between too few and too many classes. Data should be thoroughly examined to see if there is any outlier (extreme values). Also check whether the observations differ much from one to another. However, the following points are important to remember in deciding the class width and number of classes.

- The class intervals should preferably be of equal width.
- The number of classes should be such that the true nature of the distribution is made manifest.

The number of classes (k) vis-à-vis the widths (w) in a frequency table is somewhat arbitrary. In general, your table should have between 5 and 20 classes. The width should also be of reasonable length. Both w and k depend on the spread of the data set. Too few classes would not reveal any details about data; too many would prove as confusing as the list of raw data itself. Similarly, too wide a w will also be of no use for estimation of any parameter of the frequency distribution. On the other hand too short a w will fail to condense the data.

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

$$k = \frac{L - S}{w} = \frac{R}{w} \qquad \cdots (2.1)$$

An empirical formula is available for the determination of k. This is of the following form:

$$k = 1 + 3.322 \log N \qquad \cdots 2.2)$$

If you have somehow decided about k, the number of classes, you can use the following relation to determine w given the range R and N:

$$w = \frac{R}{1 + 3.322 \log_{10} N} \qquad \cdots (2.3)$$

This approach has been proposed by Sturge.

Another 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 the total number of observations (N) in the data set, that is $2^k \geq N$. 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 also dictates us to choose 6 classes, that is k=6.

The above rules in forming class widths (w) or number of classes (k) may be used interchangeably. Since R is always known, you are in a position to determine k knowing w or set w knowing k using any of the rules suggested above.

In forming a frequency distribution, a few points should always be adhered to:

- Each observation or item should go into one and only one class.
- Classes with zero frequencies should be avoided
- Class mid-points should be so established that the mid-point is approximately equal to the arithmetic mean of the observations in the class.
- Class limits should be so established that the class mid-points fall on a whole number.
- Avoid open-end classes.
- Use of unequal class intervals is recommended certain situations to avoid a large number of empty or almost empty classes.

2.8 CONSTRUCTION OF FREQUENCY DISTRIBUTION: DISCRETE DATA

We have earlier shown how a frequency distribution is constructed from non-numerical data. Construction of cross-tabulation or contingency tables has also been demonstrated. We now turn to show to construct a grouped frequency distribution from discrete (numerical) data is illustrated with the following example:

Example 2.4: The following data refer to the number of days the workers were absent from their work due to illness during the year preceding the inquiry (see Table 2.1 for data).

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

Clearly, the observations shown above have been arranged in ascending array (from smallest to the largest). The smallest number in the data set is 5 and the largest number is 33 so that the range is 28. If we use formula (2.3) to decide on the class widths, then w works out approximately to 6.6:

$$w = \frac{R}{1+3.322\log_{10}N} = \frac{28}{1+3.322(1.69897)} = \frac{28}{6.64} = 4.2 \approx 5$$

To be in the safe side, we round the w value to the next higher digit, so that the width of each class is 5. Note that, we have employed the following empirical formula to determine the value of k to use it in the denominator:

$$k = 1+3.322\,log\,50 = 1+3.322(1.69897) = 6.64$$

Note that we do not need to go beyond 6 classes as there is no value beyond the class 30–34.

The accompanying table shows the result of the tallying process. It is constructed by reading down the data in Example 2.4 above 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 intervals	Tally marks	Frequency
5–9	////	4
10–14	𝍸 𝍸 ///	15
15–19	𝍸 𝍸 𝍸 𝍸 /	21
20–24	𝍸 /	6
25–29	///	3
30–34	/	1
Total	–	**50**

The data in Example 2.4 are discrete vis-à-vis numerical and hence the resulting distribution is a **discrete frequency distribution**. Look at the above table which contains a tally column. To construct the table under reference, we first remove the tallies. This will result in a table which will take the form as shown in the accompanying table (Table 2.14) without the tallies in the second column.

Table 2.15: Distribution of Workers by the Number of Days Absent

Days absent	Number of workers
5–9	4
10–14	15
15–19	21
20–24	6
25–29	3
30–34	1
Total	**50**

The table is simple to interpret. Of the total 50 workers, 4 were absent for 5 to 9 days, 15 for 10 to 14 days and so on. One problem of grouping the data here is that we loose the identity of individuals. We cannot say, how many of these 4 workers were absent for 5 days. Was there any worker who was absent for 6 days? We have no way to say how these four workers were distributed by the number of days they were absent. Yet then such summarization is very important to make condensation of data for further statistical analyses.

A full description of the table so constructed is seen in the accompanying table showing the class frequencies, lower limits, upper limits, class marks and class widths.

Table 2.16: Different Column Headings of a Table: An illustrative Example

Class interval	Class widths	Class frequencies	Lower limits	Upper limits	Class marks
5–9	5	4	5	9	7
10–14	5	15	10	14	12
15–19	5	21	15	19	17
20–24	5	6	20	24	22
25–29	5	3	25	29	27
30–34	5	1	30	34	32
Total	–	**50**	–	–	–

Note that for discrete distributions, the class limits are **always** inclusive in nature.

The number of days absent in Example 2.4 were all isolated numbers, representing the **discrete data**, but if they had been weights, rounded to the nearest pound or age, rounded to the nearest year, or temperature rounded to nearest degree, it would have been a case of **continuous data** and 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 discuss below how a frequency distribution is constructed with continuous data obtained from measuring a continuous variable, in the present case, the age data of the workers presented in the beginning of this chapter in Table 2.1.

2.9 CONSTRUCTING FREQUENCY DISTRIBUTION: CONTINUOUS DATA

The construction of a grouped frequency table from continuous data is illustrated with the following example:

Example 2.5: The ages of the 50 workers appearing in Table 2.1 are reproduced below in an ordered array. Construct a suitable frequency distribution with class width of appropriate size.

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 the formula, $k = 1 + 3.322 \log N$, k works out to 6.64, which we approximate to 6 for practical reason:

$$k = 1 + 3.322 \log 50 = 1 + 3.322(1.69897) = 6.64$$

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 R 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

$$w = \frac{(54 - 25)}{6} = 4.83$$

-which is approximated to 5.

With the values of k and w specified above, we form a frequency distribution that appears below. Note that the tallying process may also be followed here.

Table 2.17: Age Distribution (in years) of 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 of the same data preserving the continuity of data.

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, whereas one who is between 34.5 and 35 would be counted in the third. 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 (CF) can be expressed as

$$CF = \frac{\text{Lower limit of second class} - \text{Upper limit of first class}}{2} \quad \ldots (2.4)$$

The resulting frequency table with the true limits is as follows

Table 2.18: Revised Version of Table 2.16 with True 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. 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. With this conversion, we can also speak of **upper** and **lower class boundaries.** 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:

Table 2.19: An Alternative Way of Representing Table 2.17

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 but less than 49.5	7
49.5 to less than 54.5	4
Total	**50**

Example 2.6: For the daily salary data, the highest wage is Tk 97 and the lowest wage is Tk 50. To decide on the number of classes to be chosen, 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. With this choice of k, we estimate w with the formula $w=R/k$ giving w=8. The resulting distribution is shown in Table 2.20 below:

Table 2.20: Distribution of Workers by Wages Earned

Wage (Tk.)	Number of workers
50–57	6
58–65	7
66–73	14
74–81	10
82–89	10
90–97	3
Total	**50**

Solution: Since the wage data are continuous, the distribution, when adjusted for this continuity, appears in Table 2.21:

In constructing the distribution under reference, 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.

Table 2.21: Revised Form of Table 2.19 Adjusted for Continuity

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**

So far as the formation of class intervals is concerned, one more important thing to note is whether the data are given to the nearest unit, or to the nearest tenth of a unit, or to the nearest hundredth of a unit. The following example explains further how a frequency distribution is constructed from such data.

Example 2.7: The data below specify longevity of 80 electric bulbs in months. 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

Solution: 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 w. With these choices, the frequency distribution is as follows:

Table 2.22: The Longevity Distribution of Electric Bulbs

Longevity (in years)	Number of bulbs
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 classification of the data and hence needs to be adjusted to reflect 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. The correction factor is (2.0–1.9)/2=0.05 here. The distribution incorporating this correction factor is as follows:

Table 2.23: Revised Version of Table 2.22 with Class Boundaries (True Class Limits)

Class boundaries (Longevity in years)	Frequency (Number of bulbs)	Percent (%)
1.45–1.95	4	5.0
1.95–2.45	2	2.5
2.45–2.95	8	10.0
2.95–3.45	30	37.5
3.45–3.95	20	25.0
3.95–4.45	10	12.5
4.45–4.95	6	7.5
Total	**80**	**100.0**

The interpretation of the results in the above table is that out of 80 bulbs, 4 burnt out within a period of 1.45 months to1less than 1.95 months. For as many as 30 bulbs, the longevity varied between 2.95months and less than 3.45 months

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.

Example 2.8: The accompanying table shows the statistics of toll (revenue) collected from Bangabandhu Bridge for 20 FYs from 1998–99 through 2017–18 in crore taka (source: Bangladesh Economic Review 2019). Construct a frequency distribution with appropriate class widths.

Fiscal year	Toll (in crore taka)	Fiscal year	Toll (in crore taka)
1998–99	58.81	2008–09	212.45
1999–00	64.77	2009–10	243.93
2000–01	81.15	2010–11	267.66
2001–02	91.99	2011–12	304.66
2002–03	107.02	2012–13	325.20
2003–04	129.30	2013–14	323.28
2004–05	150.43	2014–15	349.08
2005–06	156.06	2015–16	402.43
2006–07	171.50	2016–17	484.42
2007–08	199.55	2017–18	543.80

Solution: The primary task in constructing a frequency distribution is to determine the number of classes (k) and the class width (w).

Employing the 2^k rule, we note that $2^4=16$ which is less than N, where $N=20$. Again $2^5=32$, which is greater than N. Hence the 2^k rule suggests a k value of 5, i.e. $k=5$. The empirical rule provides with $k=1+3.322 \log_{10} N=1+3.322(1.3010)=5.3$. Comparing these two, wed decide to adopt $k=6$ an hence

$$w=\frac{R}{k}=\frac{543.80-58.81}{6}=80.8$$

To ensure the coverage of all the values, we choose w to be 85.

Check that with the estimated number of classes and class width, the resulting table will be as follows:

Table 2.24: Revenue Collection from Bangabandhu Bridge for 20 Fiscal Years

Revenue	No. of years	Percent
55.5–140.5	6	30.0
140.5–225.5	5	25.0
225.5–310.5	3	15.0
310.5–395.5	3	15.0
355.5–480.5	1	5.0
480.5–565.5	2	10.0
Total	**20**	**100.0**

As the table shows, the govt. received the highest revenue between 480.5 and 565.5 crore taka only for two years out of a total span of 20 years, the lowest being for 6 years when the revenue collection was between 55.5 to 140.5 crore of taka..

Example 2.9: The data refer to the GDP growth (real) for Bangladesh for a period of 40 consecutive years from 1980 to 2019 in percentages (Source: Bangladesh Economic Review, 2019). Construct a frequency distribution table with appropriate class limits.

3.1	5.6	3.2	4.6	4.2	3.7	4.0	2.9	2.4	4.3
4.6	4.2	4.8	4.3	4.5	4.8	5.0	5.3	5.0	5.4
5.6	4.8	4.8	5.8	6.1	6.3	6.9	6.5	5.5	5.3
6.0	6.5	6.3	6.0	6.3	6.8	7.2	7.6	7.9	8.1

Solution: Adopting the 2^k rule for $N=40$, we choose k to be 6, while the empirical rule gives us $k=1+3.322 \log_{10} N=1+3.322(1.6021) = 6.3$. This rule suggests $k=7$. Finally we choose 7 class intervals. The width w is determined as follows:

$$w=\frac{R}{k}=\frac{L-S}{k}=\frac{8.1-2.4}{7}=0.81$$

For better coverage, we choose $w=1.0$. You can check by tallying process as before that the distribution can be constructed as follows:

Table 2.25: GDP Growth in Bangladesh for 40 Years

(GDP growth %)	Number of years	Percent
1.95–2.95	2	5.0
2.95–3.95	3	7.5
3.95–4.95	12	30.0
4.95–5.95	9	22.5
5.95–6.95	10	25.0
6.95–7.95	3	7.5
7.95–8.95	1	2.5.
Total	**40**	**100.0**

2.10 OTHER FORMS OF FREQUENCY DISTRIBUTION

There are varieties of way to present a frequency distribution depending on the requirement and needs. A few of them are

a) Percentage frequency distribution
b) Relative frequency distribution
c) Cumulative distribution

2.10.1 Percentage Distribution

It is sometimes convenient and rewarding to deal with percentage distribution rather than the absolute ones. 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.

A percentage distribution is formed by dividing the number of cases attributable to a category 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

$$P_i = \frac{f_i}{N} \times 100 \quad \ldots (2.5)$$

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

$$\sum \left(\frac{f_i}{N} \right) \times 100 = 100. \quad \ldots (2.6)$$

2.10.2 Relative Frequency Distribution

Instead of presenting the frequencies in absolute figures, 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

$$\text{Relative frequency of the } i\text{-th class} = \frac{f_i}{N} \qquad \cdots (2.7)$$

The total relative frequency in such a distribution will add to 1.0. That is

$$\sum\left(\frac{f_i}{N}\right) = 1.0 \qquad \cdots (2.8)$$

Note that the relative frequencies are essentially **proportions**, which when multiplied by 100, result in percentage frequencies and hence the percentage distribution.

For the wage data in Example 2.6, the percentage frequency distribution and relative frequency distribution are shown in a single table as below:

Table 2.26: Absolute, Percentage and Relative Frequency Distributions Based on Wage Data in Example 2.6

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**

Thus, for example the percentage frequency for the 3rd class is

$$\frac{f_3}{N} \times 100 = \frac{14}{50} \times 100 = 28\%$$

And the relative frequency for the same class is

$$\frac{f_3}{N} = \frac{14}{50} = 0.28$$

2.11 CUMULATIVE FREQUENCY DISTRIBUTION

In many occasions, the analyst is interested not in the number of observations falling in eac class, but rather: how many of the observations in the distribution have a value less than more than some benchmark values. Referring to Table 2.1, we might ask a series of question a few of which are:

- How many of the workers receive wage less than taka 73.5?
- How many of the workers receive wage taka 83.5 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 boundaries** for each class. This can be formed from the frequency distribution, the relative frequency distribution or from the percent distribution. We illustrate the procedure with a few examples below.

Example 2.10: Refer to the wage data in Table 2.26. Construct a less than type cumulative frequency distribution.

Solution: The necessary column headings are displayed in row 1 of the table under reference.

Wage (in taka)	Class frequency	Less than type cumulative frequency	% Less than type 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 interpretation of the values in column 3 is simple. The cumulative frequencies in column 3 represent the wage less than the upper boundaries of the indicated wages 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 73.5 as above, it is more convenient to say that 54 percent of the workers received such an amount. These percentages are shown in the last column of the table.

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 shown in Table 2.27:

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

Table 2.27: An Alternative Way of Presenting Less than Frequency Table: Wage Data

Wage (in taka)	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

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 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 2.21, we construct a decumulative table below, and observe that all 50 of the workers have wage greater 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 2.28: Decumulative Frequency Distribution of 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 fifth entry '13' in column 2 simply states that out 50 workers, 13 (26%) receive a sum of taka 81.5 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 sake of constructing a meaningful graph at a later section.

2.12 GRAPHICAL PRESENTATION OF DATA

In addition to presenting a frequency distribution in tabular form, one can present the same through some visual aids. This refers to **graphs** and **diagrams or chart.** 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. In this section, we will discuss how a frequency distribution can be represented by such graphical devices. As we are aware that a frequency distribution can be

constructed either from categorical (qualitative) or quantitative (numerical) data, the graphs and diagrams to be constructed will also differ accordingly: **categorical (qualitative)** or **quantitative (numerical)**. First, we will deal with the presentation of categorical (qualitative) data followed by presentation of quantitative (or numerical) data.

2.12.1 Presentation of Categorical Data

The most common forms of diagrams suitable for presenting categorical data are the following:

(a) Bar diagram
(b) Stacked bar diagram
(c) Cluster or multiple bar diagram
(d) Pie diagram
(e) Pareto diagram
(f) Dot plot

(a) Bar Diagram

Bar diagrams also called bar charts are commonly used to describe categorical data. A bar diagram is a form of presentation in which the frequencies against the categories are represented by rectangles separated usually along the horizontal axis and drawn as bars of convenient widths. The widths of these bars have no significance but are taken to make the chart look attractive.

Example 2.11: The accompanying table shows the stock position of finished goods in Metric tons as of June 2004 of the Bangladesh Chemical Industries Corporation (BCIC). Represent the data by a suitable bar diagram.

Finished goods	**Quantity** (in Metric tons)
TSP	8916
SSP	18455
Paper	2660
Cement	7048
Sanitary ware	1620
Insulator	3520
Tiles	17335
Total	**54928**

Source: MIS Report of BCIC, June 2004, p: 13.

Solution: There are a number of ways to represent the data by bar diagram. We represent the data by two different bar diagrams. These are

(a) Vertical bars
(b) Horizontal bars

Figure 2.2 and Figure 2.3 represent these bars

যেকোনো একটা draw করলেই হবে।

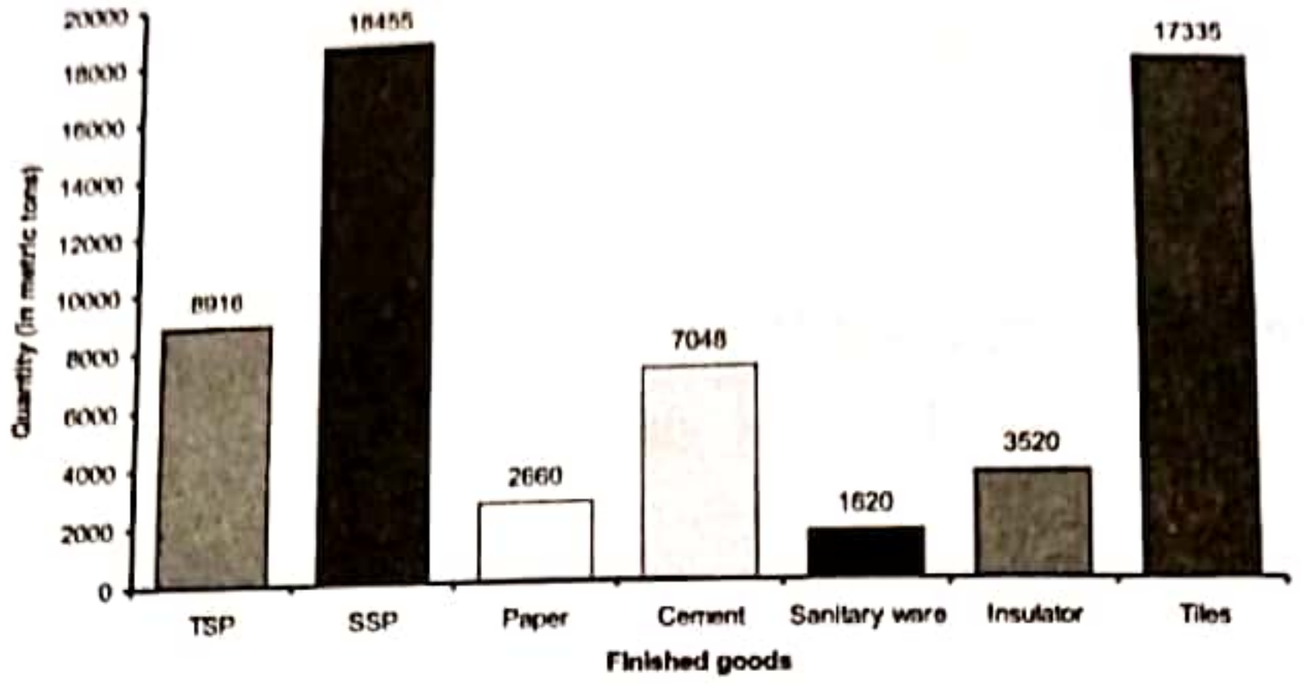


Figure 2.2: Vertical bar chart showing the stock positions of goods

The diagram clearly shows that BCIC had the highest stock of SSP followed by tiles while sanitary wares showed the least.

An alternative way of representing the same data may be accomplished by what is known as the horizontal bar diagram. Figure 2.2 below is such a diagram.

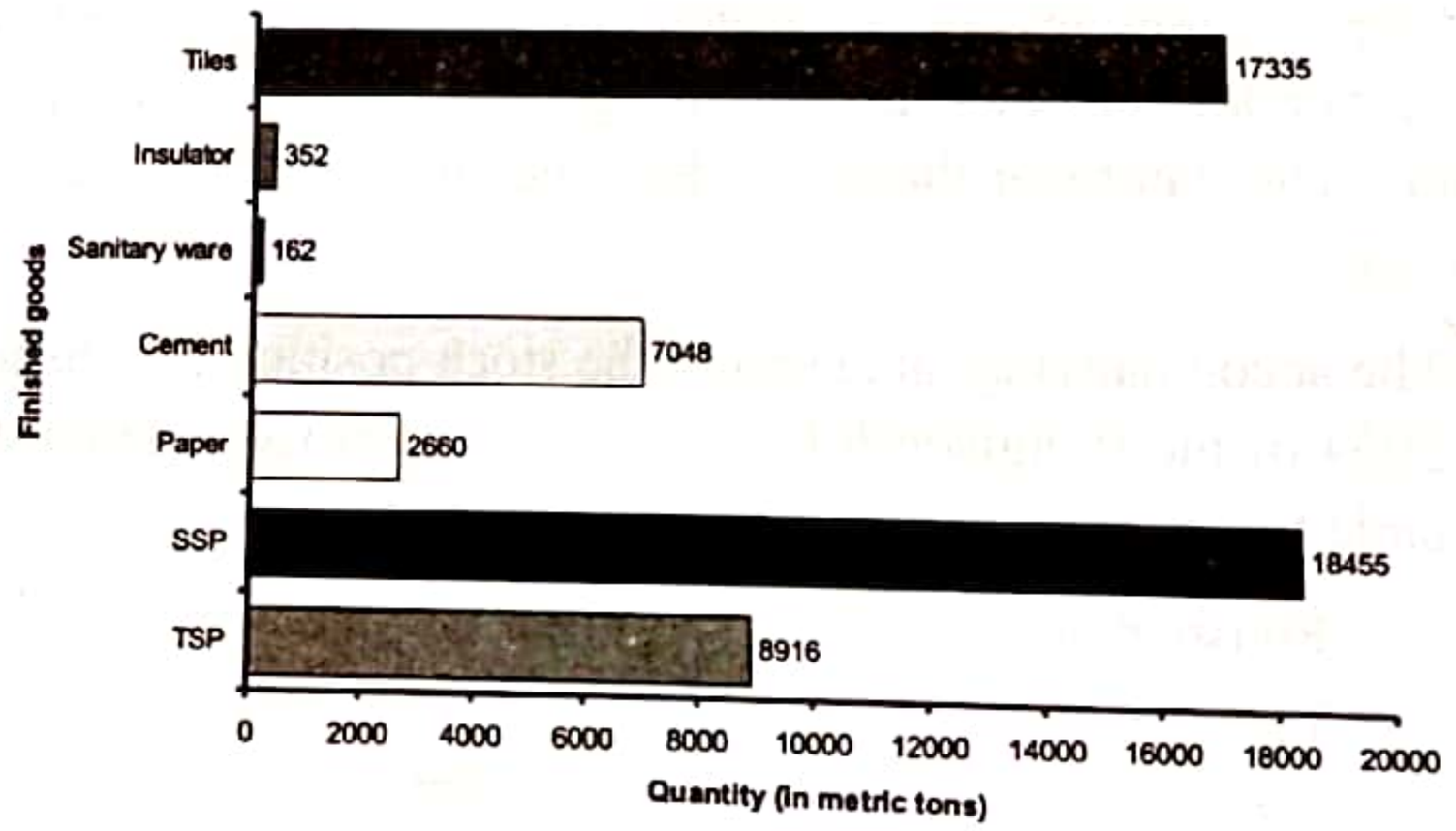


Figure 2.3: Horizontal bar chart showing the stock positions of goods

Example 2.12: Table below compares the average labor productivity internationally acros key trade competitors with Bangladesh. Display the data by a horizontal bar diagram

Country	Average labor productivity, 2010 (1990 PPP$)
Vietnam	5877
Thailand	16352
China	13056
South Asia	8038
Sri Lanka	16085
India	8515
Bangladesh	3841

Solution: The diagram is constructed using appropriate scales and is shown below:

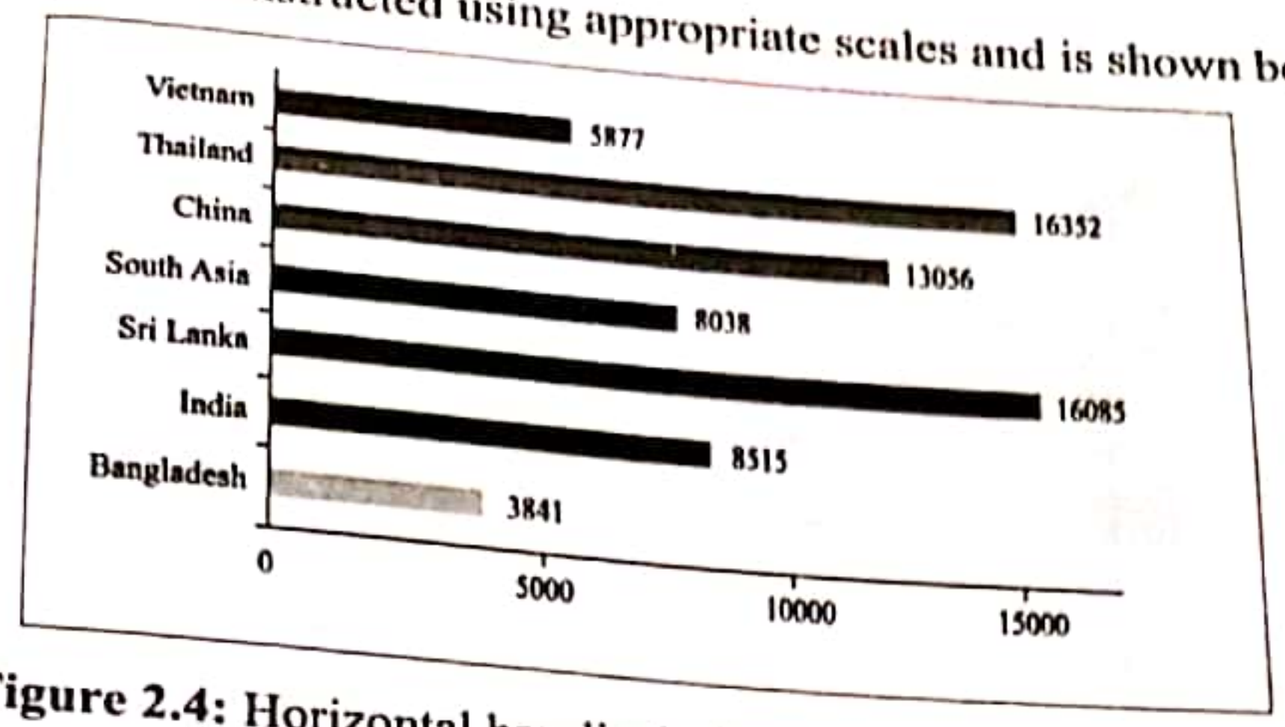


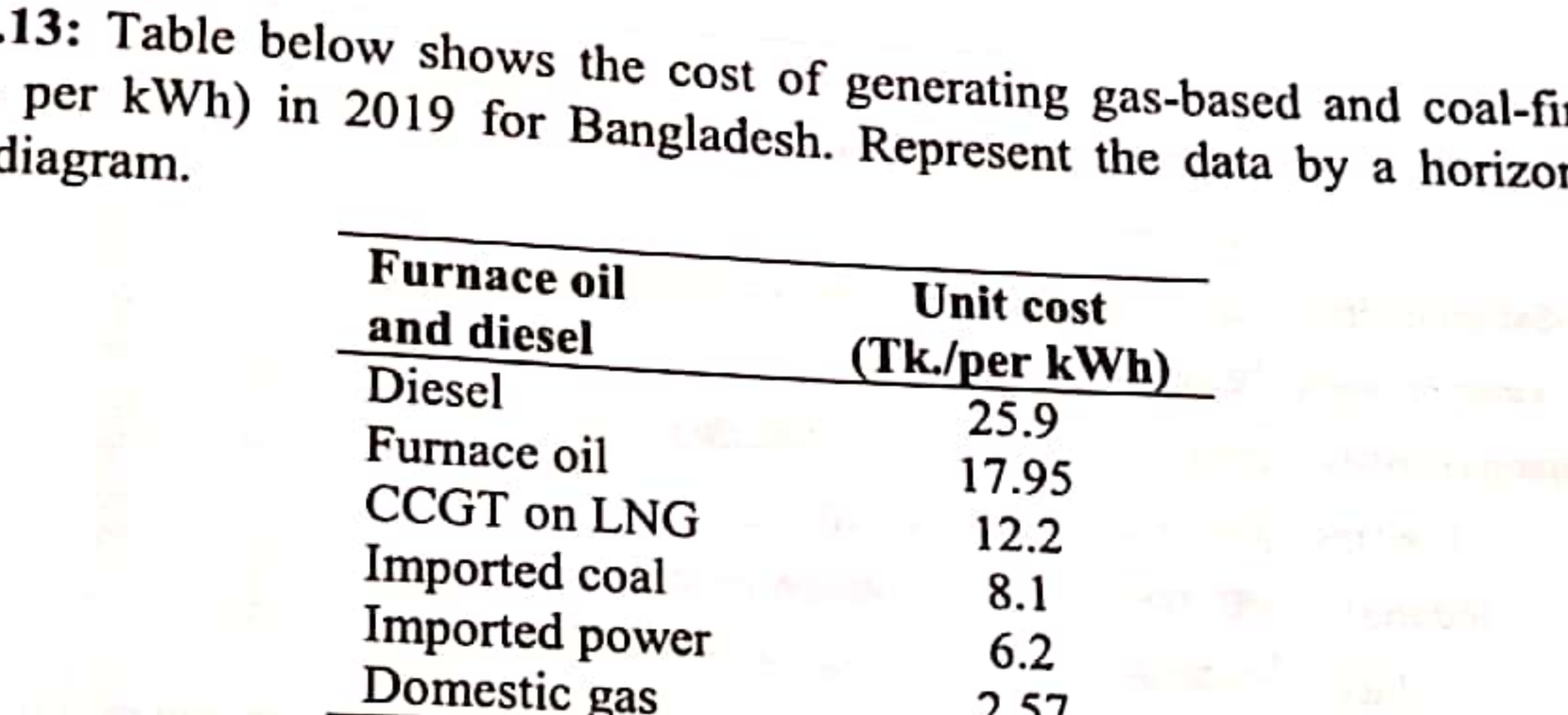
Figure 2.4: Horizontal bar displaying the data in Example 2.12

As the figure shows, even after allowing for purchasing power differentials of the US dollar across countries, the average labor productivity in Bangladesh is far behind its main trade competitors.

Example 2.13: Table below shows the cost of generating gas-based and coal-fired plants energy (Tk. per kWh) in 2019 for Bangladesh. Represent the data by a horizontal and a vertical bar diagram.

Furnace oil and diesel	Unit cost (Tk./per kWh)
Diesel	25.9
Furnace oil	17.95
CCGT on LNG	12.2
Imported coal	8.1
Imported power	6.2
Domestic gas	2.57

Solution: Figures 2.5 and 2.6 respectively display the data on both horizontal and vertical scales.



Source: Compiled by GED

Figure 2.5: Horizontal bar diagram representing the data on cost of oil

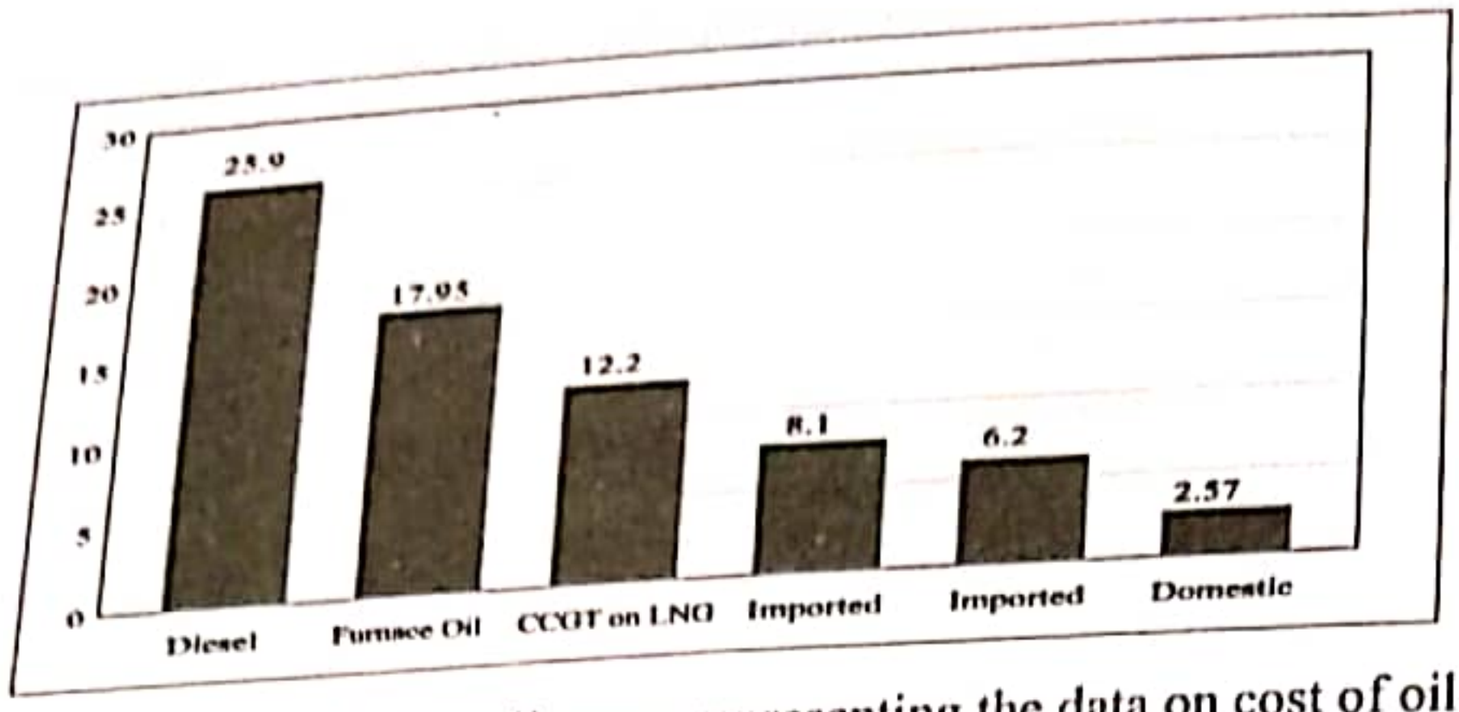


Figure 2.6 Vertical bar diagram representing the data on cost of oil

Example 2.14: The growth performance of Bangladesh in the sixth plan compared to the best performing nations globally is shown in the accompanying graph. Comment on the performance of Bangladesh in relation to the countries and areas shown on the top the bars.

Solution: The graph is drawn below from the original data (data not shown)

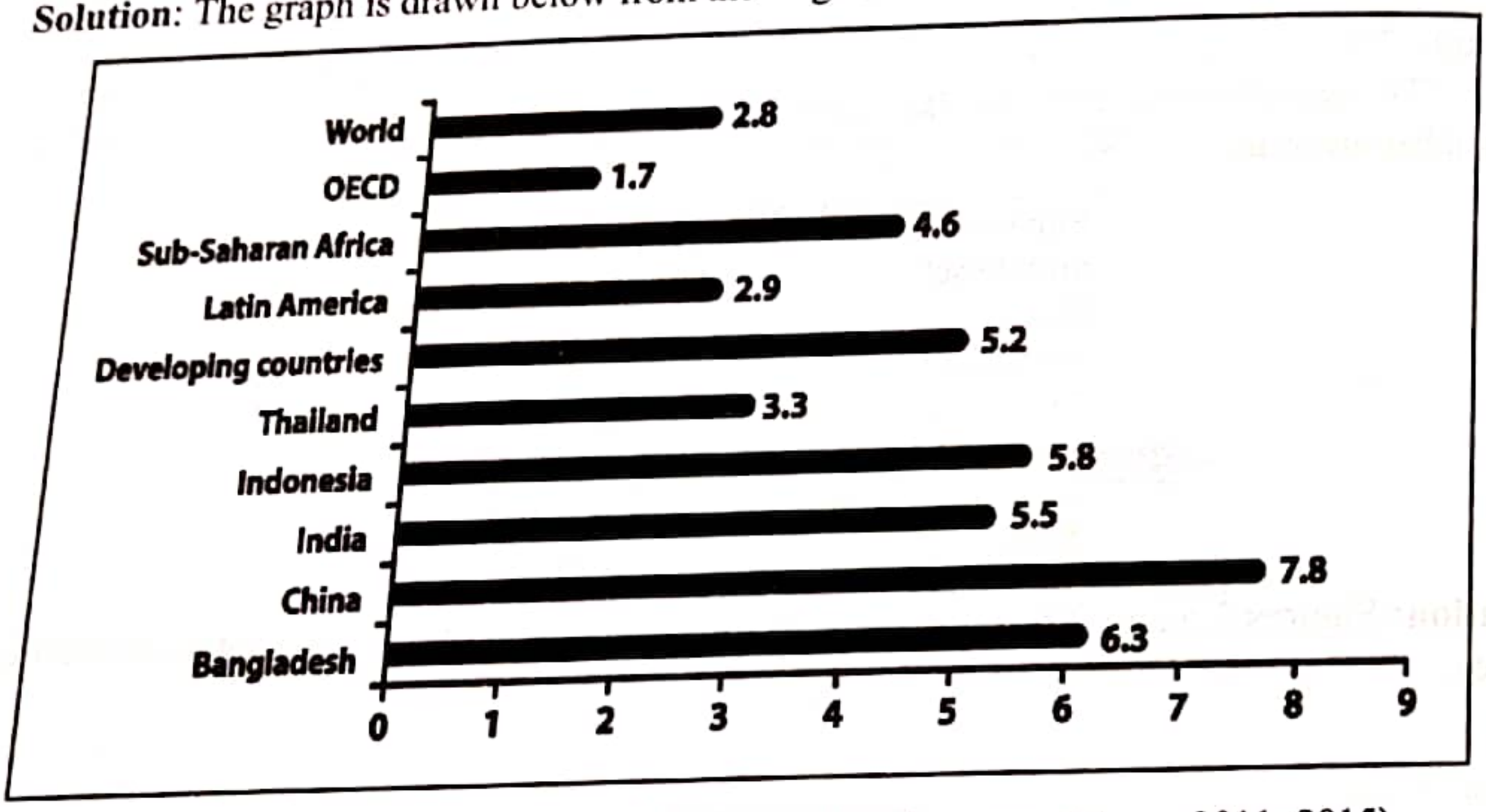


Figure 2.7: *Sixth FYP performance in international context (%p.a.: 2011–2015)*

Conclusion and interpretation: Growing at an average pace of 7.8 percent per annum over the five years of 2011–2015, China remains the global growth leader. Bangladesh grew at 6.3 percent ahead of India, Thailand Indonesia, and the average for all developing countries. The performance is even more impressive when compared with average growth in Sub-Saharan Africa., Latin America, and the global average.

Example 2.15: Table below shows the number of employments for a period of ten years from FY 2008-09 to FY2017–18 as reported in Bangladesh BMET, Bangladesh Bank. Display the data by a bar chart of both horizontal and vertical type.

FY	Employment abroad (in '000)
2008–09	650
2009–10	427
2010–11	439
2011–12	591
2012–13	441
2013–14	409
2014–15	441
2015–16	625
2016–17	905
2017–18	900

Source: Bangladesh Bank: 2019

Solution: The vertical bars in Figure 2.8 below shows the number of expatriate employees abroad during the stated period.

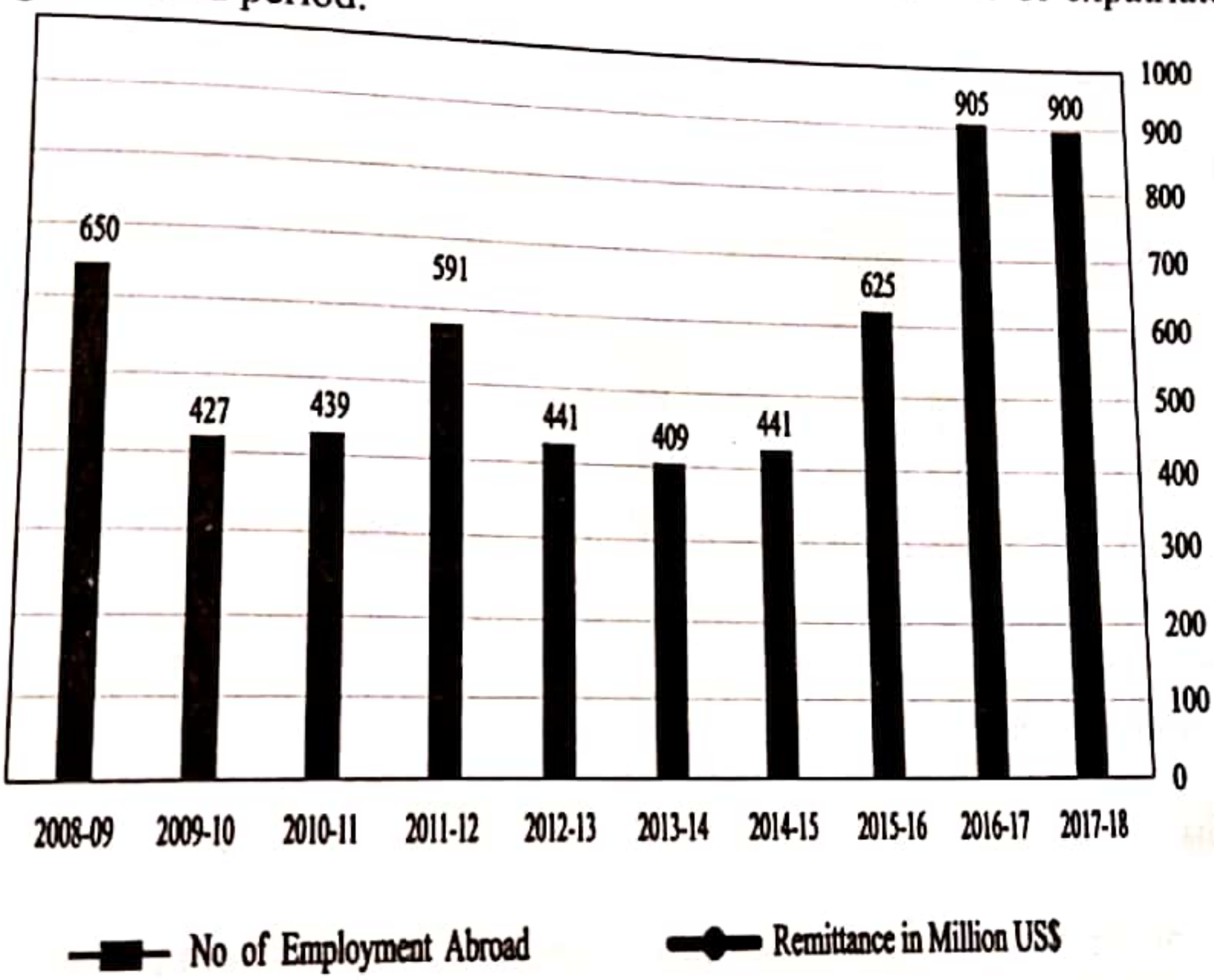


Figure 2.8: Manpower export and remittance inflow: 2008–2018

Conclusion: As the graph shows, the number of employees going abroad remained almost static till 2014–15, which thereafter showed an upward trend and reached to 900 thousand in FY 2017–18.

Example 2.16: The data below present the percentages of remittance receiving households by the number of dwelling rooms as obtained in 2011 Population Census of Bangladesh and reported in Population Monogram on Housing conditions prepared by BBS.

Represent the data by a 3-dimensional cylindrical bar diagram.

Number of dwelling units	Percentage of remittance receiving households
1	14.2
2	25.8
3	25.1
4	18.5
5	8.9
6	4.5
7	3.1
Total	100.1

Solution: Cylindrical charts are column or bar chart that use cylinder shaped items to show data. Although cylinder charts do not add any additional data, sometimes using this shape allows us to achieve better visual appearance of data. Here is the diagram (Figure 2.9) representing the above data.

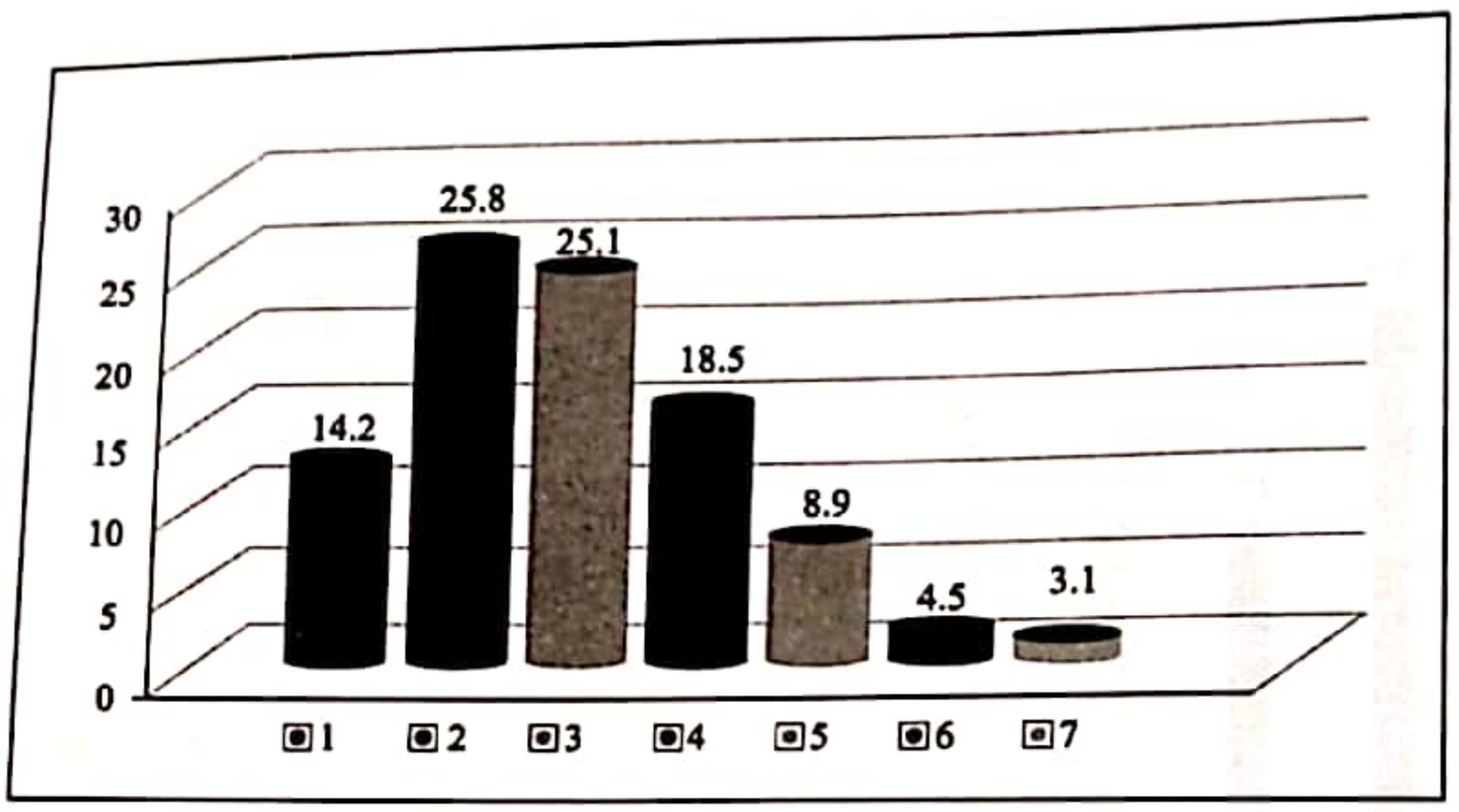


Figure 2.9: Cylindrical bar chart displaying remittance receiving households in Bangladesh

2.12.2 Pictogram

A **pictogram** is one of the simplest (and most popular) forms of data visualization. Also known as "pictographs", "icon charts", "picture charts", and "pictorial unit charts", **pictograms** use a series of repeated icons to visualize simple data. In such charts, icons represent numbers to make it more interesting and easier to understand. A key is often included to indicate what each icon represents. All icons must be of the same size, but a fraction of an icon can be used to show the respective fraction of that amount.

What is the difference between pictogram and histogram?

Pictogram is the way of expressing the data with the help of pictures. On the other hand, **histograms** are used to show "continuous data", or the data that is distributed in small intervals within a certain range. Figure 2.10 below is an example to illustrate a pictogram displaying how the population of New Zealand grew over the period 1901–1996.

Figure 2.10: Pictograms displaying New Zealand population: 1901–1996

Pictograms may also be viewed as **cone charts.** Cone charts are just usual bar charts that use conical shaped items instead of instead of rectangular bars to represent data. Like cylindrical charts, such charts also do not add any additional data, but use of this shape allows us to achieve an attractive visual appearance of the data in hand. We illustrate below the use of cone chart by an example.

Example 2.17: Table below shows the average expenditures in selected heads from remittance by eight administraive divisions in Bangladesh, 2016. Display the data by a cone diagram.

Division	Average expenditure (in Taka)
Barisal	82064
Chottagong	54274
Dhaka	90712
Khulna	84166
Rajshahi	86551
Rangpur	112243
Sylhet	56962
Natonal	73879

Solution: Cone chart in essence is pictogram. Find below the diagram (Figure 2.11) constructed ussing the given data. As you can note Rangpur divion topped otehr divisons in expenditure followd by bfollowed by dhaka. The lowest expenditure was reported in Chittagong division. Four divisions. Barisal, Dhaka, Khulna and Rajshahi shoed the expendture above the national average.

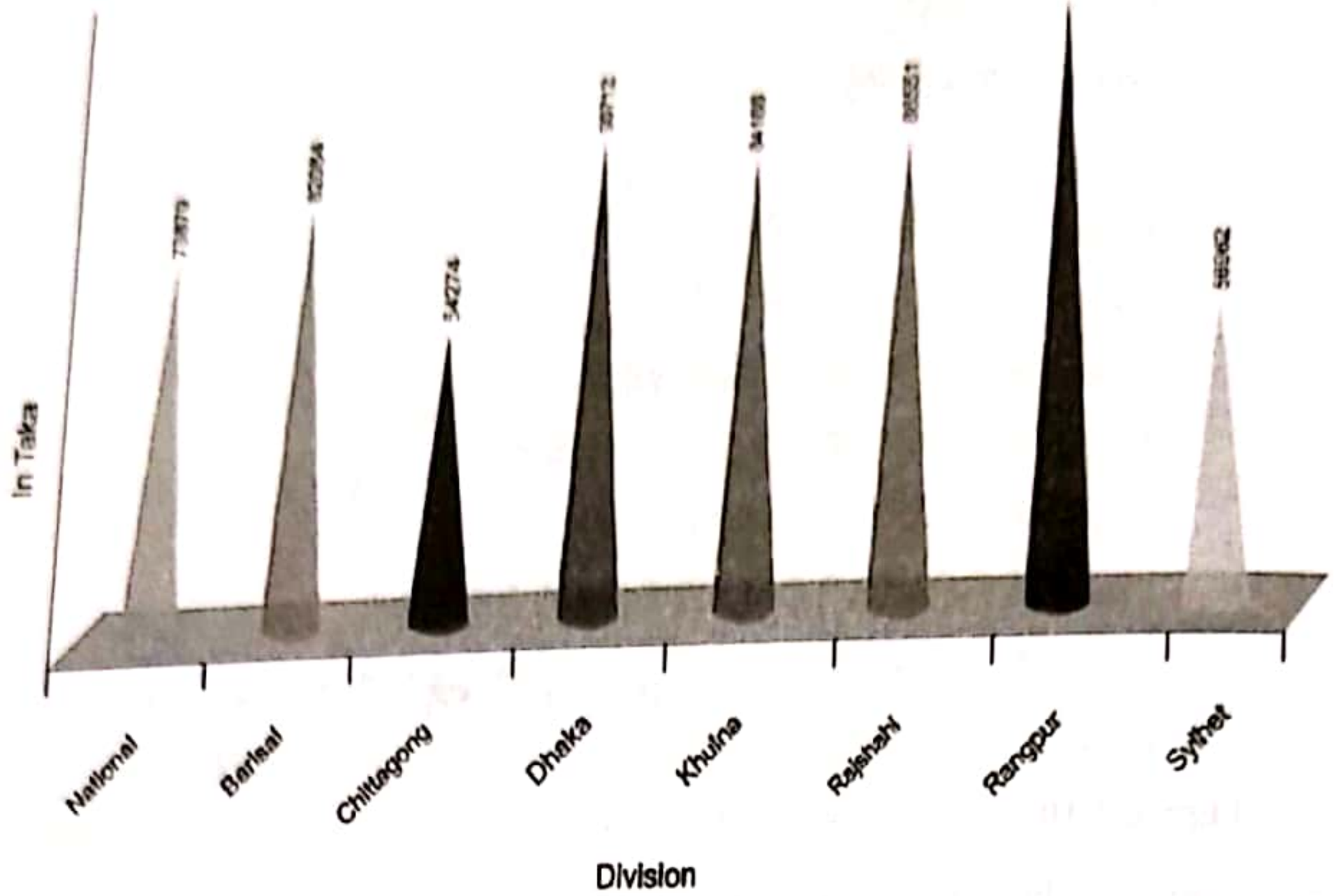


Figure 2.11: Cylindrical bar chart displaying remittance receiving households

(b) Stacked Bar Chart

A **stacked bar** chart, variously known as **component bar chart** or **sub-divided bar chart,** is a side by side chart that uses bars to show comparisons between categories of data, but with ability to break down and compare parts of a whole. Each bar in the chart represents a whole, and segments in the bar represent different parts or categories of that whole. Stacked bar chart is a good device to display categorical data. The component parts are variously colored or shaded to make them distinct and attractive.

As with the simple bar graph, the stacked bar graph uses rectangular boxes to represent categories of a variable. The variable located on the *x*-axis is known as the **stacked variable**. The staked bar graph differs from the simple bar graph in that each rectangular box on the *x*-axis is made up of smaller individual boxes that we call **segments**. The segments are stacked on top of one another and could also be called the *y*-axis variable. The height of each segment represents the number (or percentages) of cases in a category of the stacked variable and a category of the **segment variable.** The stacked bar graph depicts the relationship between the categories of the stacked and segment variables. The stacked bar graph provides a convenient way to discover and then visualize relationship between two discrete variables.

There are two types of stacked bar graphs: simple stacked bar and 100 percent stacked bar. Simple stacked bar graphs place each value for the segment after the previous one. The total value of the bar is all the segment values added together. It is ideal for comparing the total amounts across each group/segmented bar. The percent stacked bar graphs display the

percentage-of-the-whole of each group and are plotted by the percentage of each value to the total amount in each group. This makes it easier to see the relative differences between quantities in each group.

Stacked bar diagram is best suited for comparing two or more sets of data. One can draw such diagrams to compare two family budgets, stockholder's equity for two successive years, assets and liabilities of two companies and the like. We can plot the data by two types of stacked bar: simple and 100% stacked bar. In the construction of 100% stacked bar charts, we follow the following steps:

(a) Convert each component in the data set into percentage value of the corresponding total
(b) Draw one rectangle for each total, taking equal lengths of 100 units and breadths proportional to the totals
(c) Divide each rectangle so drawn into parts equal in number to the number of components
(d) Use shading or color to distinguish one component from the others.

The following example is designed to illustrate the construction of a stacked bar diagram of both types.

Example 2.18: The gross revenue expenditures of the government of Bangladesh in million BDT for the financial years 2011-12 and 2012-13 (budget estimate) were as follows:

Heads of gross revenue expenditure	**Expenditure in million taka**	
	FY 2011–12	**FY 2012–13**
Wages and salaries	225350	229400
Commodities and services	122220	130330
Transfer	346420	386270
Other services	335040	385330
Total	**1029030**	**1131330**

Represent the data in a simple stacked bar and 100% stacked bar diagram.

Solution: We first draw a simple stacked bar chart followed by a percentage stacked bar. In order to draw the percentage graph, we convert the tabular values into percentages and put them in the accompanying table.

Table 2.29: Conversion of Absolute Values into Percentages

Heads of revenue expenditure (gross)	**% Expenditure in Taka**	
	2011-12	**2012-13**
Wages and salaries	21.9	20.3
Commodities and services	11.9	11.5
Transfer	33.7	34.1
Other services	32.5	34.1
Total	**100.0**	**100.0**

The resulting stacked bar appears in Figure 2.12 below

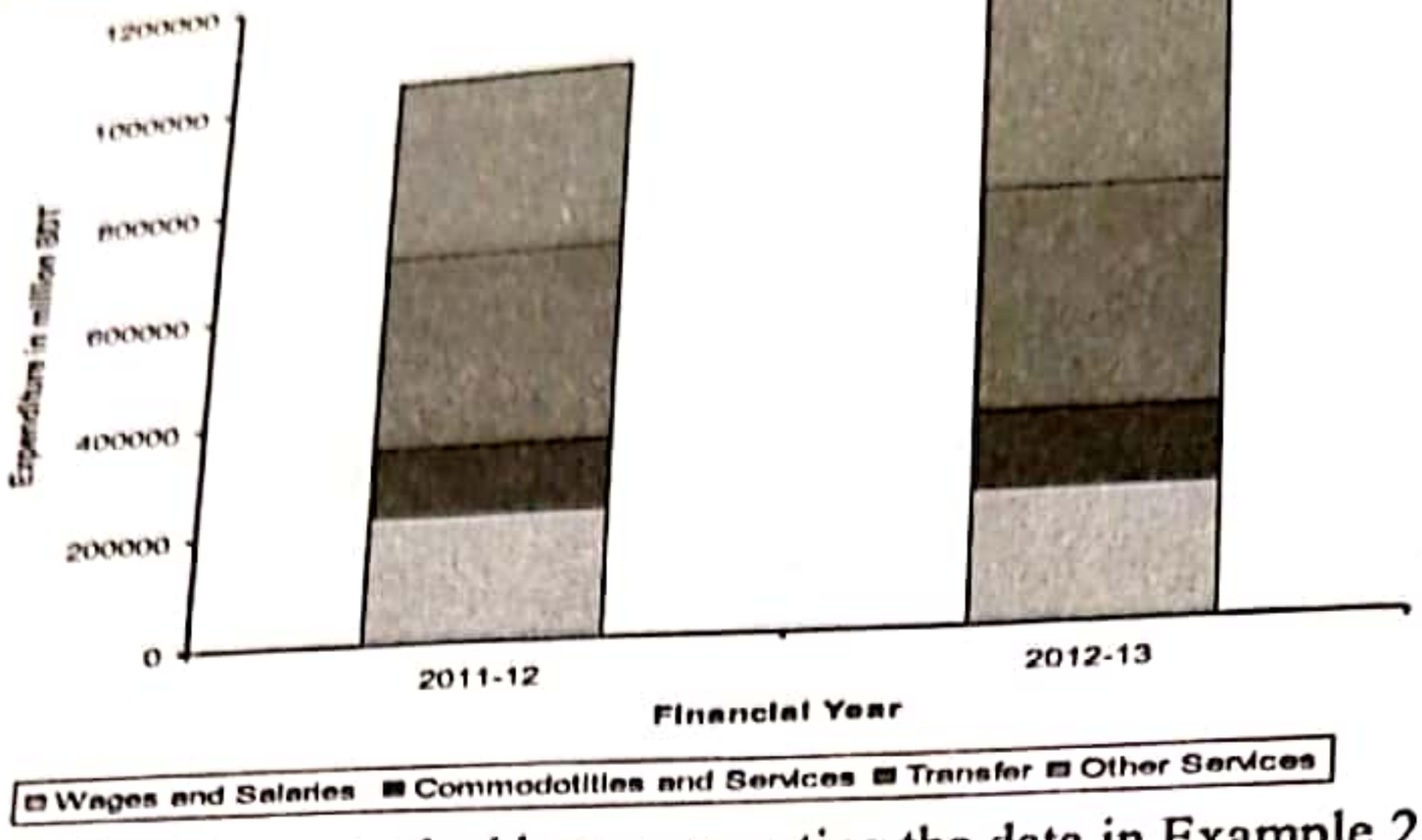


Figure 2.12: Simple stacked bar representing the data in Example 2.18

For instant comparison, you can join the respective segments of the bars by simple straight lines as shown in Figure 2.13

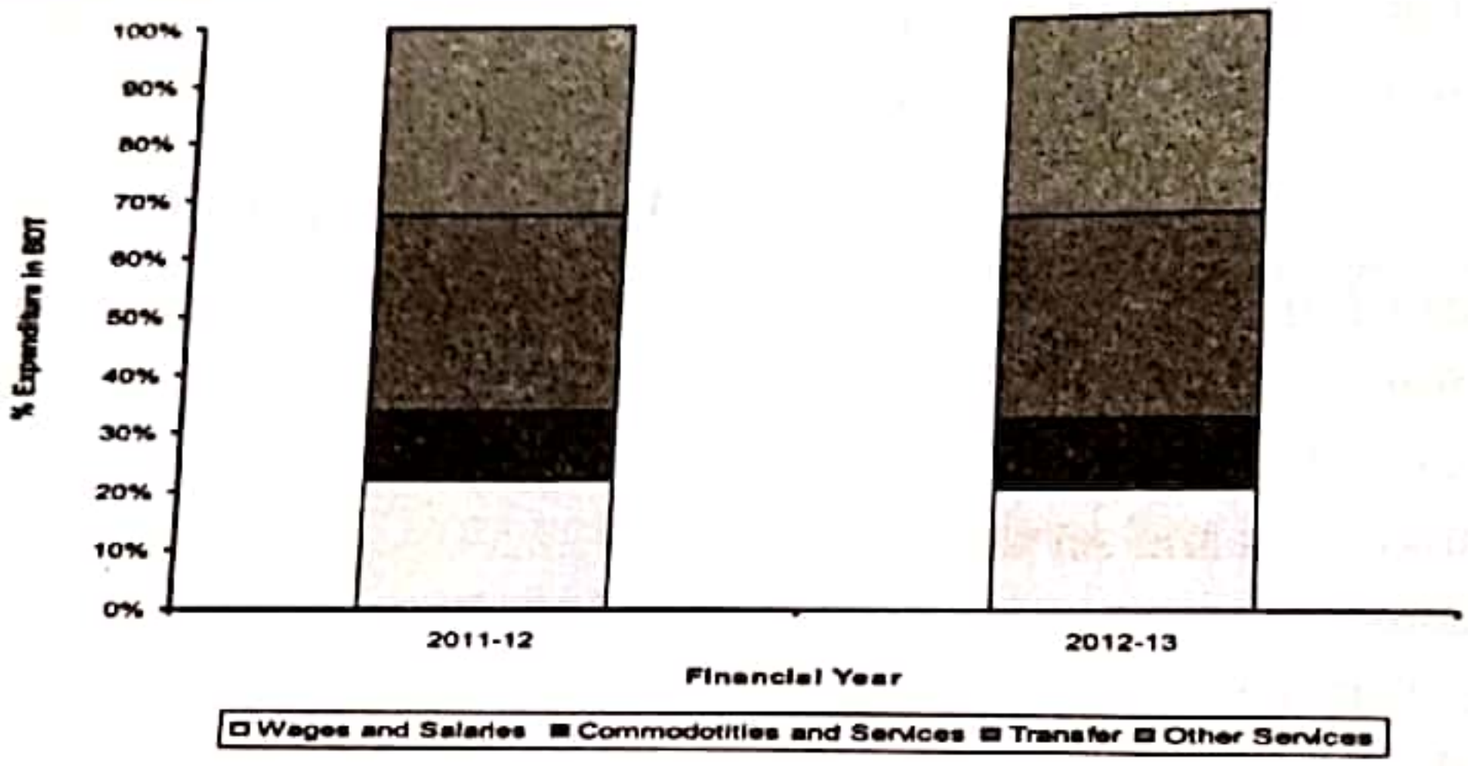


Figure 2.13: 100% stacked bar representing the data in Example 2.18

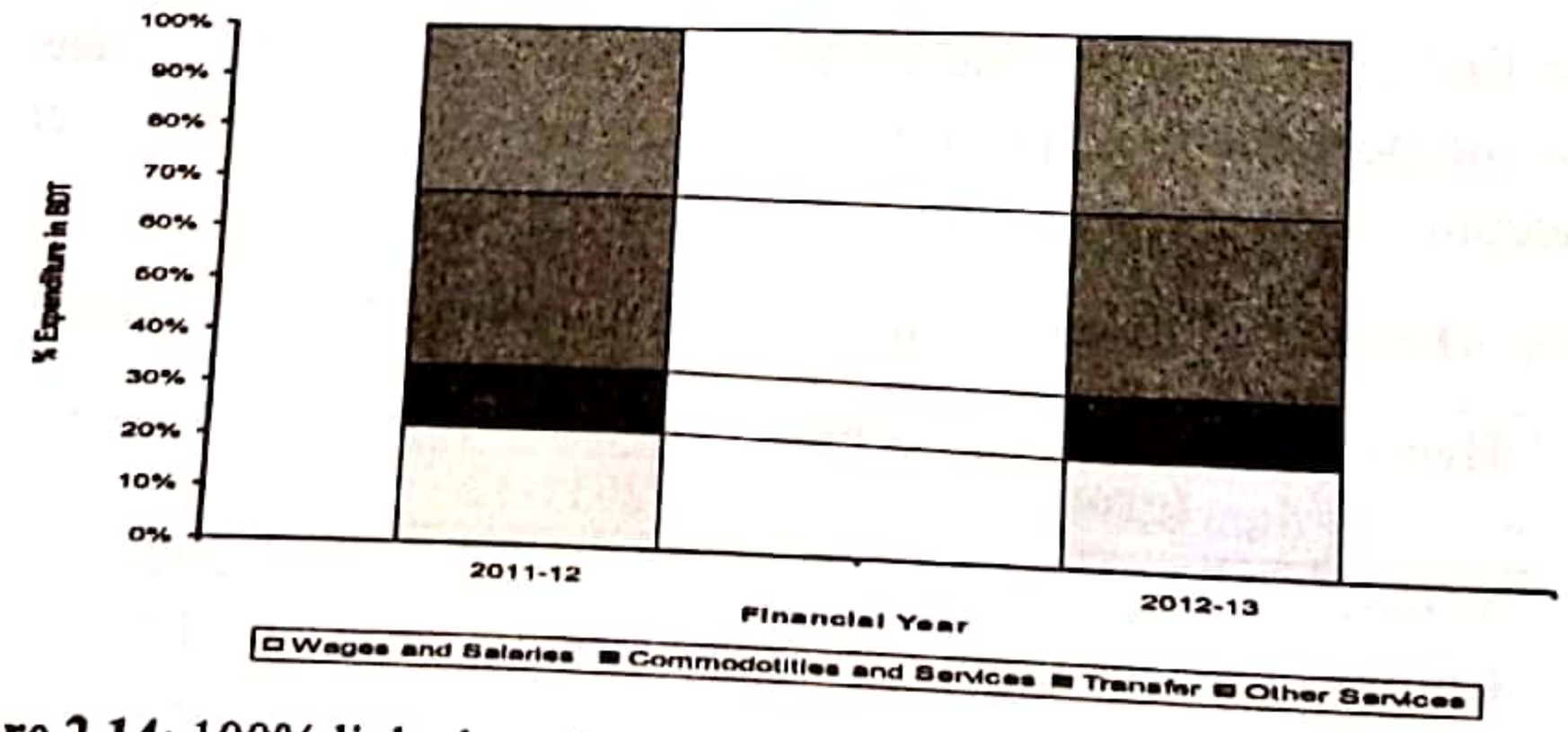


Figure 2.14: 100% linked stacked bar representing the data in Example 2.18

The intention of this current analysis based on the stacked bar chart is to compare the revenue expenditures of the government of Bangladesh by financial years. The stacked variable here is

financial years having two categories: 2012-13 and 2013-14 and segment variable is the revenue expenditures. The major query in this example we are trying to answer is whether there have been any changes in the revenue expenditures. Specifically, we may seek to answer the following questions:

- Have the percentage of wages and salaries remained the same over the financial years? This question may be repeated for all other categories of expenditures.
- Are the distributions of revenues for the financial years the same?
- Do the financial years differ in their proportions of the various heads of expenditure?

The stacked bar chart may also be used to represent the data arranged in a contingency table. Look at the example below illustrating a cross-table of family size and level of education of the BPC workers as shown in Table 2.5.

Example 2.19: Refer to Table 2.6 which shows the distribution of 50 workers by their level of education and family size. Represent the data by a stacked bar diagram.

Solution: For convenience, we reproduce Table 2.6 below as Table 2.30:

Table 2.30: A Cross-table of Education and Family Size of BPC 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**

The resulting stacked bar chart is as shown in Figure 2.15 below:

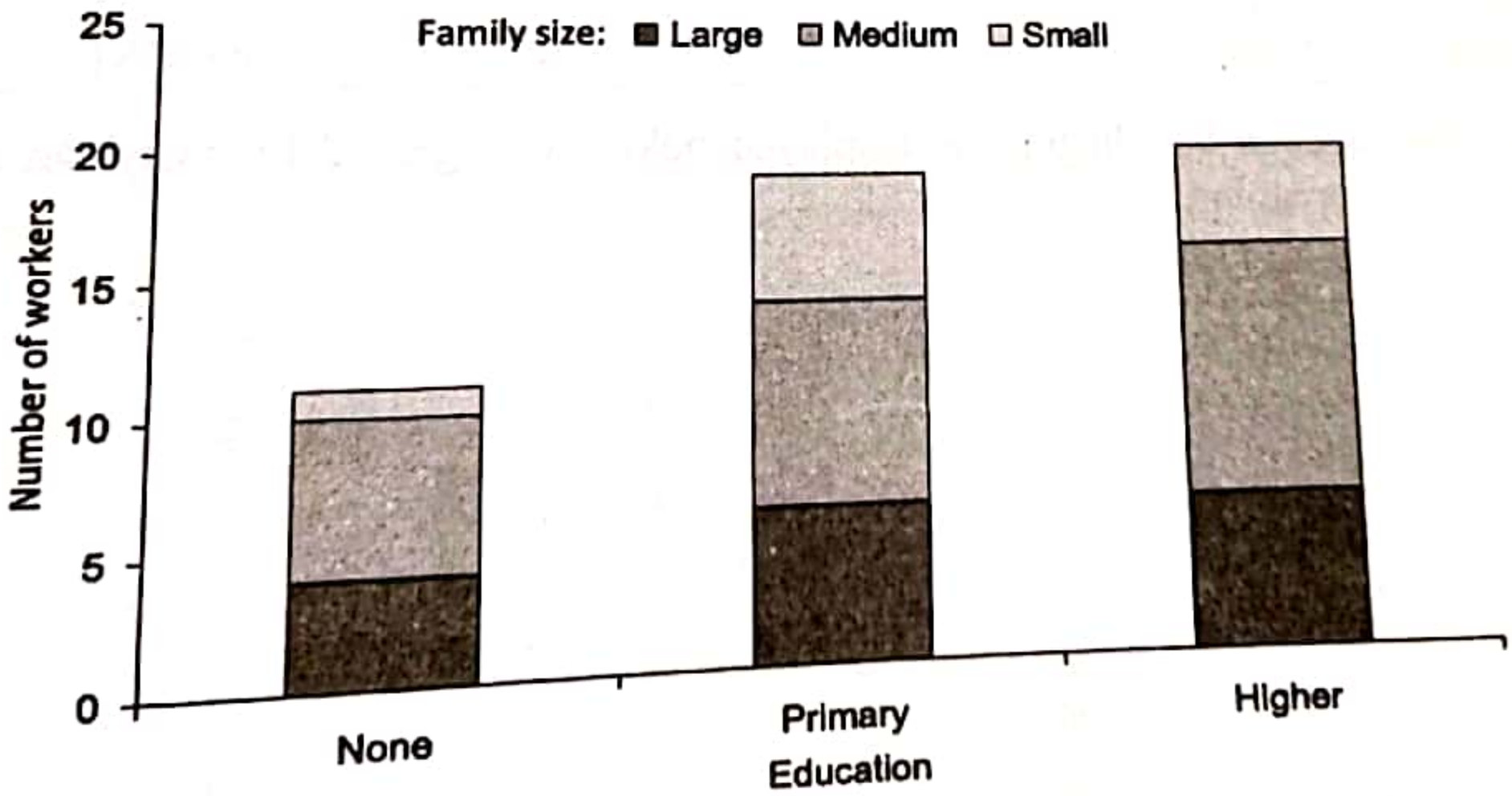


Figure 2.15: Stacked bar chart for family size and education level data

Cylindrical charts may also be used to represent data as stacked bars. Here is an example that illustrates the shape of such a chart.

Example 2.20: The following diagram shows the enumerated (for the years 1990 and 2010) and projected (for the years 2020 and 2030) international population in million in stacked cylindrical bars (Source: Internet).

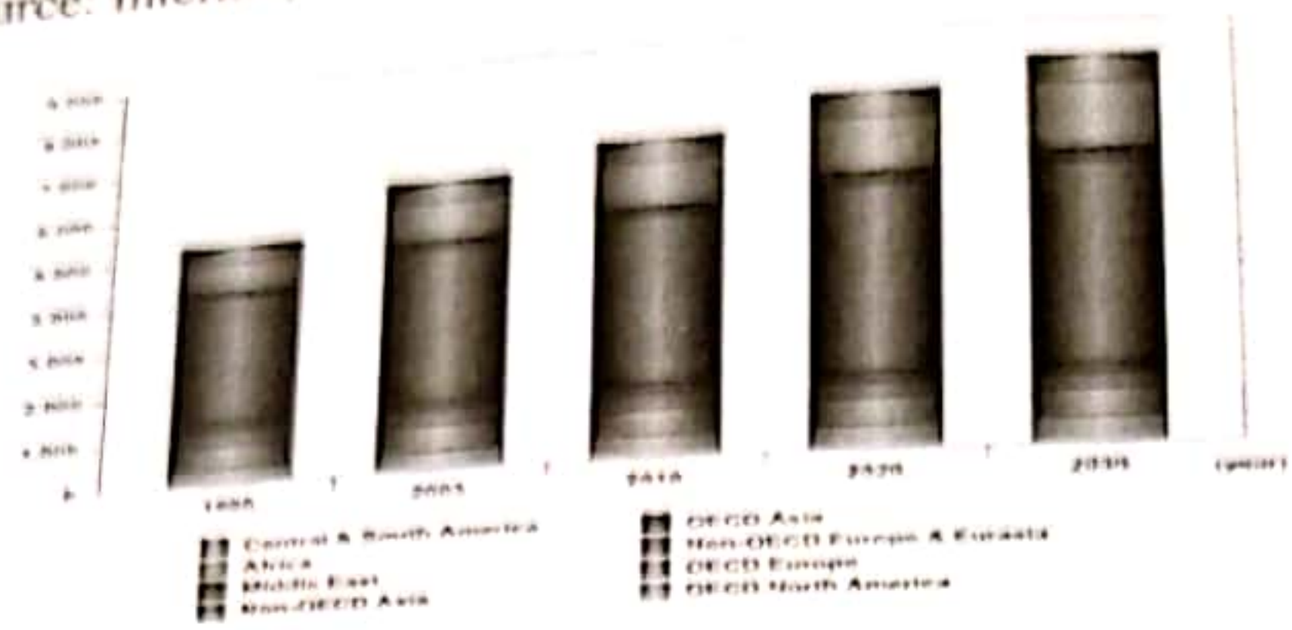


Figure 2.16: Stacked cylindrical bar chart showing the international population

Example 2.21: World University of Bangladesh awarded merit scholarship to deserving students of three faculties of the university based on their performance in the HSC examination during 2014–2018. Table below shows the details of the amount disbursed by faculties and years. Represent the data by a stacked bar chart

Year	Business	Arts and Humanities	Science and Engineering	Total
2014	968.29	524.54	939.43	2432.26
2015	383.19	311.83	2059.33	2754.35
2016	699.62	442.35	3087.74	4229.71
2017	334.16	320.82	2729.33	3384.31
2018	922.83	859.74	2986.24	4768.81

Solution: The stacked bar diagram is displayed below as Figure 2.17 based on the data presented in the foregoing table

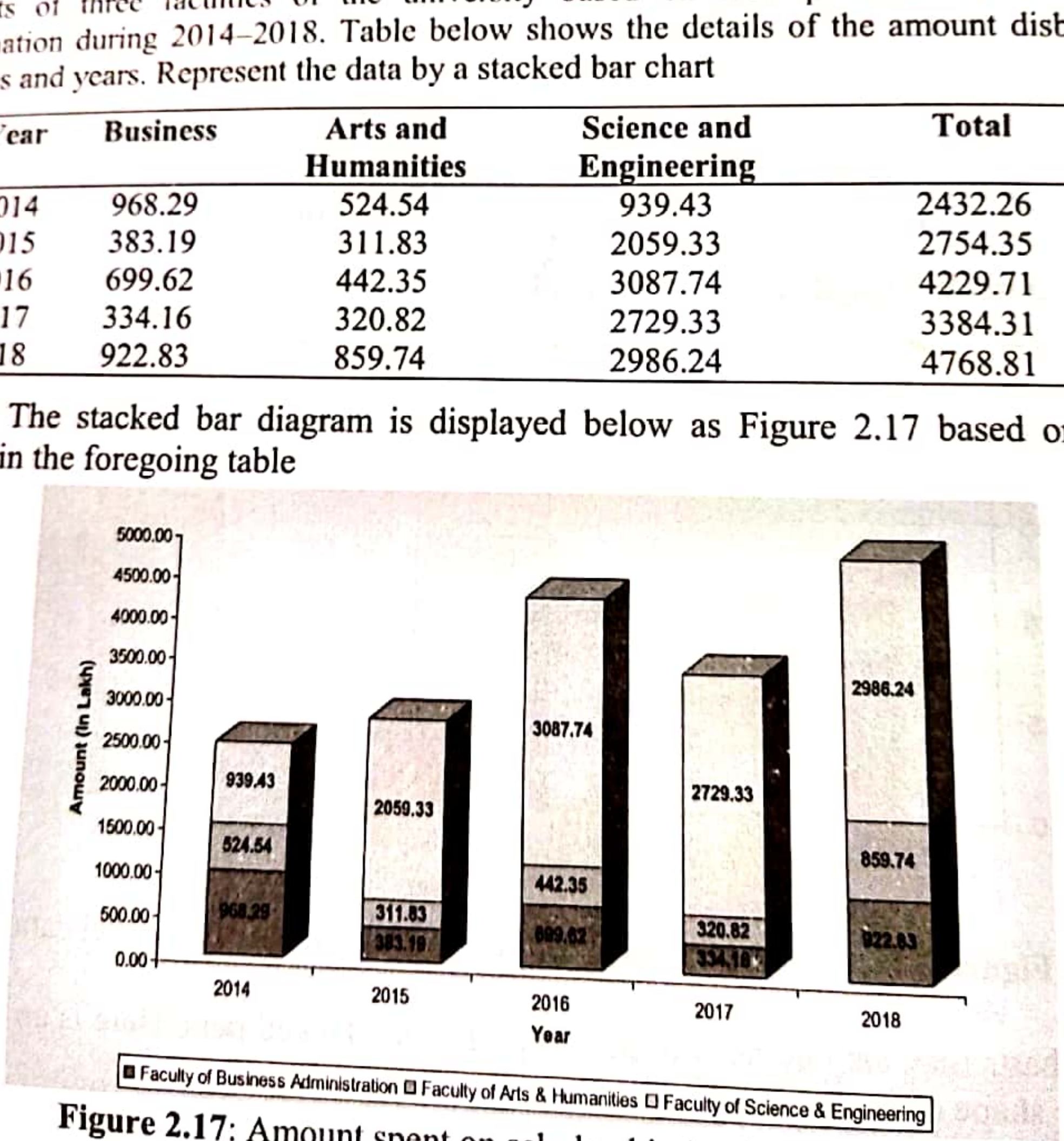


Figure 2.17: Amount spent on scholarship by faculty and year

Example 2.22: Five private banks Arab Bangladesh Bank (AB), One Bank (OB), Prime Bank (PB), Dutch-Bangla Bank (DB) and BRAC Bank (BB) reported the following remittance received in million US dollar during five financial years from 2015–2018 in their Annual Reports. The accompanying table shows these data. Represent the data by a stacked bar diagram.

Bank	Year				
	2015	2016	2017	2018	Total
AB	100	50	30	45	225
OB	96	52	52	75	275
PB	47	100	73	100	320
DB	50	90	35	75	250
BB	140	45	60	110	355

Solution: The four segments of each bank represent four years. The lower segment showing the amount in 2015 and the top most segment the amount in year 2018. The amount received in each year has been shown in each segment. The figures shown above the bars show the total of 4 years for each individual bank.

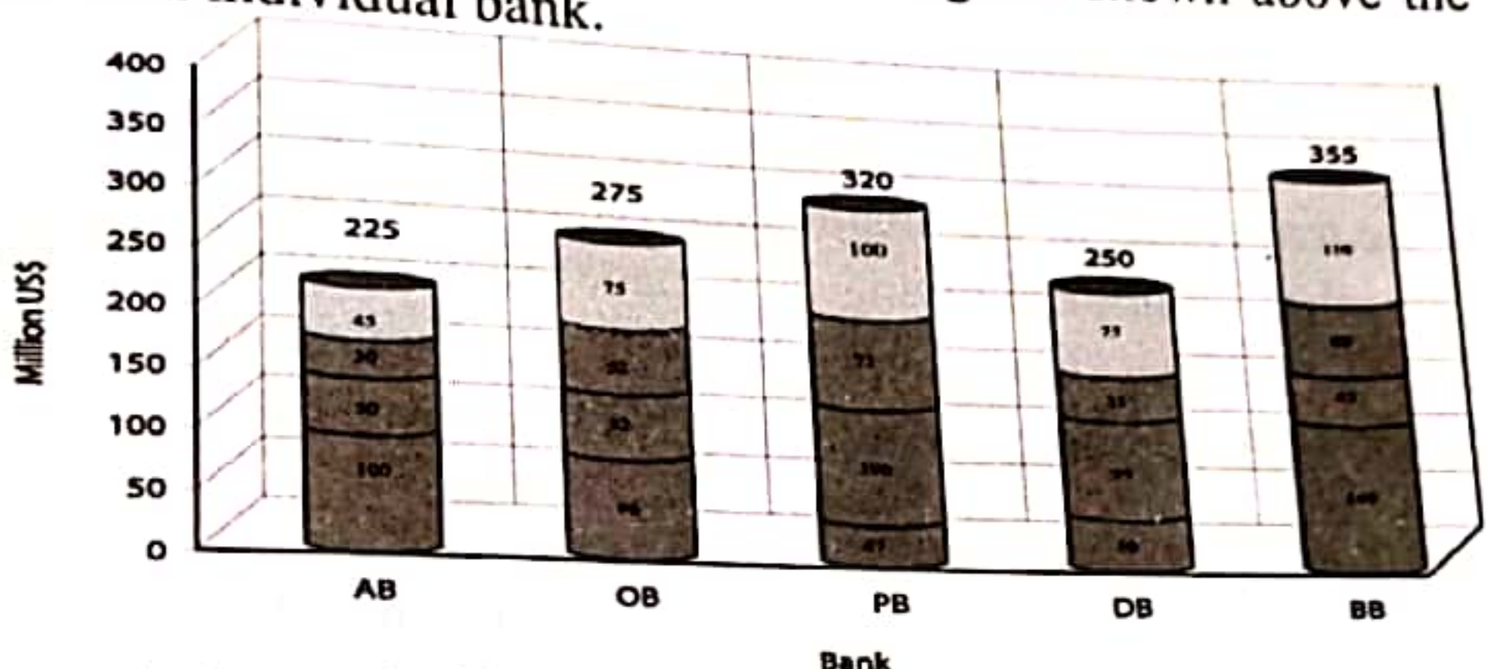


Figure 2.18: Stacked bar showing the remittances received by 5 banks

Cones may also be used to represent the data in the form of stacked bar chart. In a stacked cone chart, each series is vertically stacked one over the other. Here is an illustrative example.

Example 2.23: The consumption pattern of five different types of fruits by families of Afra, Ahnaf and Afreen in a particular month is shown in the accompanying table. Use a stacked cone chart displaying the data.

Fruit type	Family of			
	Afra	Ahnaf	Afreen	Total
Apples	5	3	2	10
Grapes	4	5	2	11
Bananas	4	4	3	11
Oranges	2	1	5	8
Melons	2	7	6	15
Total	**17**	**20**	**18**	**45**

Source: Internet (Hypothetical data)

Based on the above data, the cone chart is as follows:

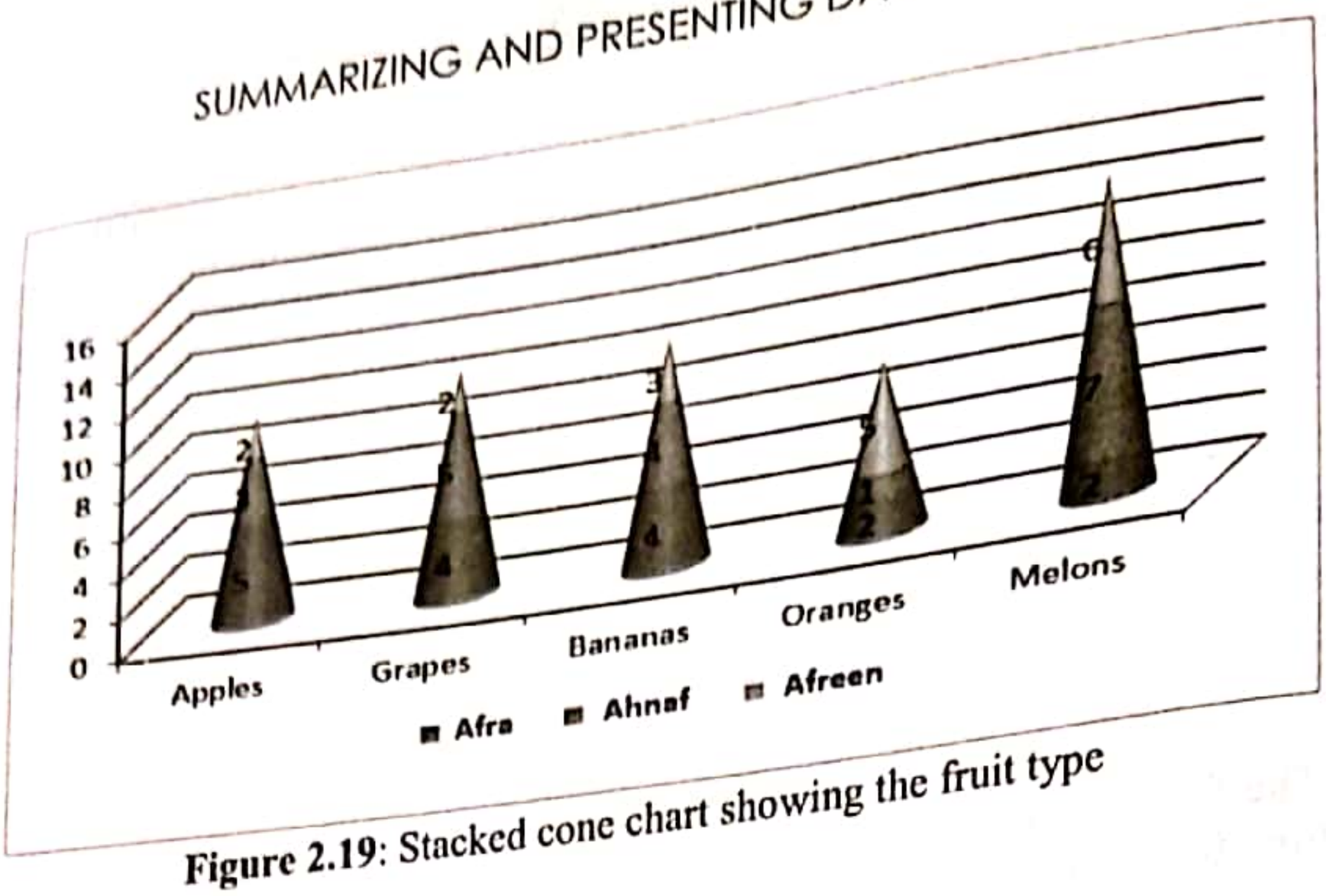


Figure 2.19: Stacked cone chart showing the fruit type

(c) Cluster Bar Chart

Another diagram, which is frequently used to present statistical data, is the **cluster** (or **multiple**) bar diagram. This is primarily used to compare two or more characteristics corresponding to a common variate value. Cluster bar charts are grouped bars, whose lengths are proportional to the magnitude of the characteristics. The bars of a cluster chart are usually put adjacent to each other without allowing any space between them. Different shading or color can be used to distinguish one group of bars from other groups. Data for which the cluster bar chart is appropriate include, among others, the following:

- Population values for different regions
- Literacy rates by sex, volume of exports by type of production
- Price earning ratios,
- Earning and dividend yields by companies in the share market, etc.

Example 2.24: 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 cluster (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)

The associated cluster bar chart is presented below in Figure 2.20

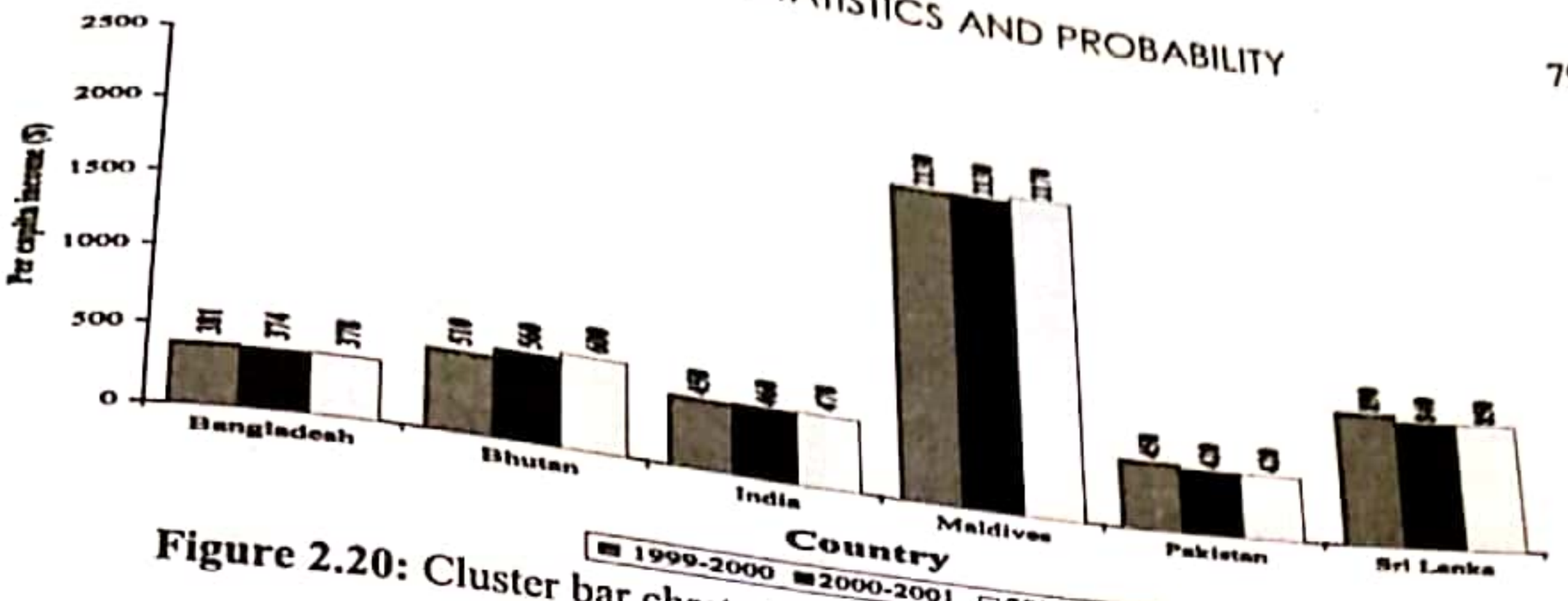


Figure 2.20: Cluster bar chart representing the data in Example 2.24

The diagram depicts that Bangladesh is lagging behind in per capita income compared to all other countries listed in the table. Maldives tops the list with an increasing trend over the study period followed by Sri Lanka.

The data in Example 2.19 used in constructing Figure 2.15 may also be represented by a cluster bar char as shown below in Figure 2.20:

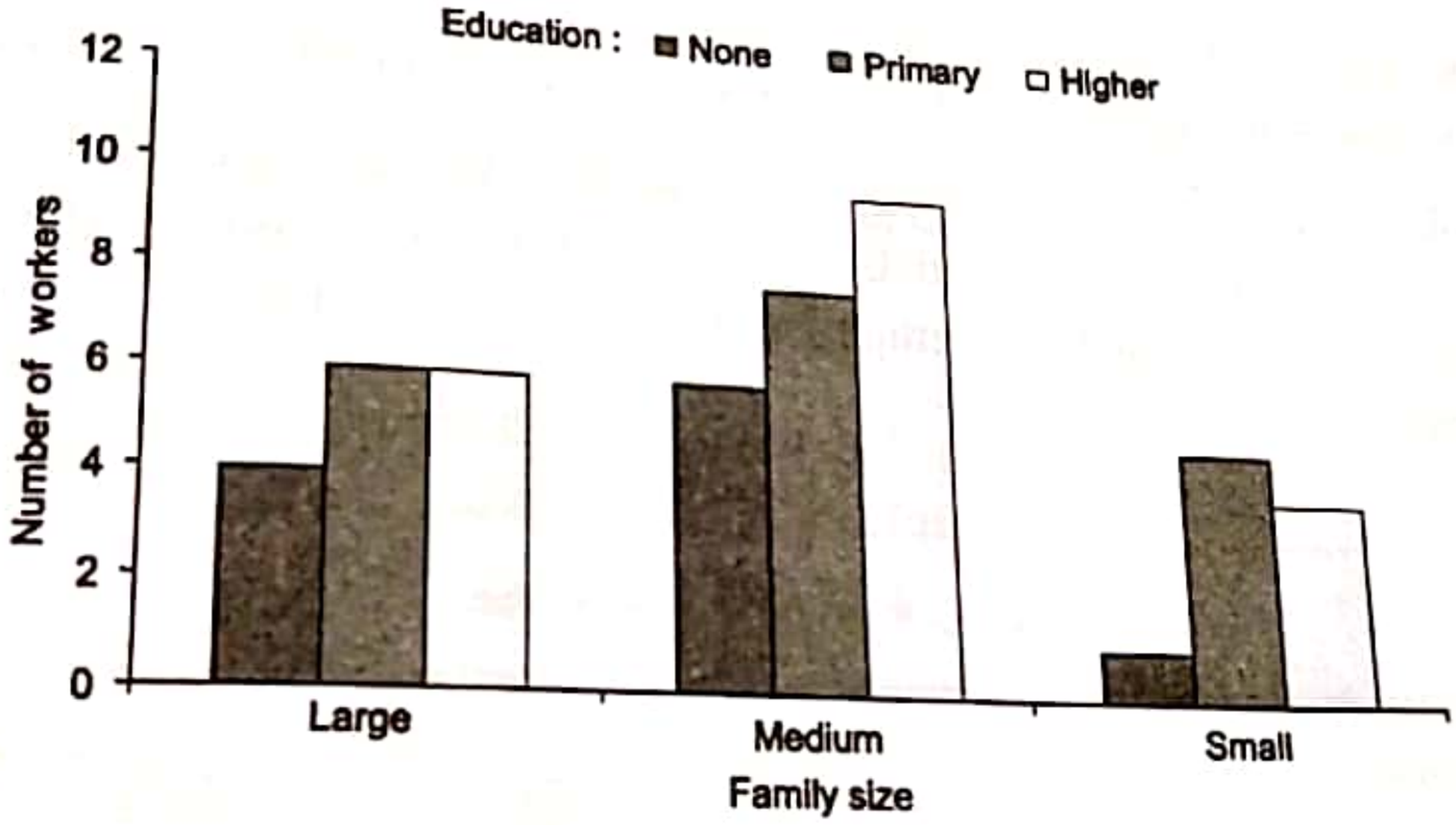


Figure 2.21: Cluster bar chart for family size and education level

Example 2.25: The accompanying table shows the government borrowing from sources during FY 1998–FY 2007–08. Display the situation on foreign aid flow in a cluster bar chart.

	Disbursement of loans and grants		
Financial year	**Grant**	**Loan**	**Total**
2002–03	510	1075	1585
2003–04	338	695	1033
2004–05	234	1257	1491
2005–06	501	1067	1568
2006–07	590	1040	1630
2007–08[R]	702	1337	2039

Solution: The given data have been displayed in the accompanying figure (Figure 2.22)

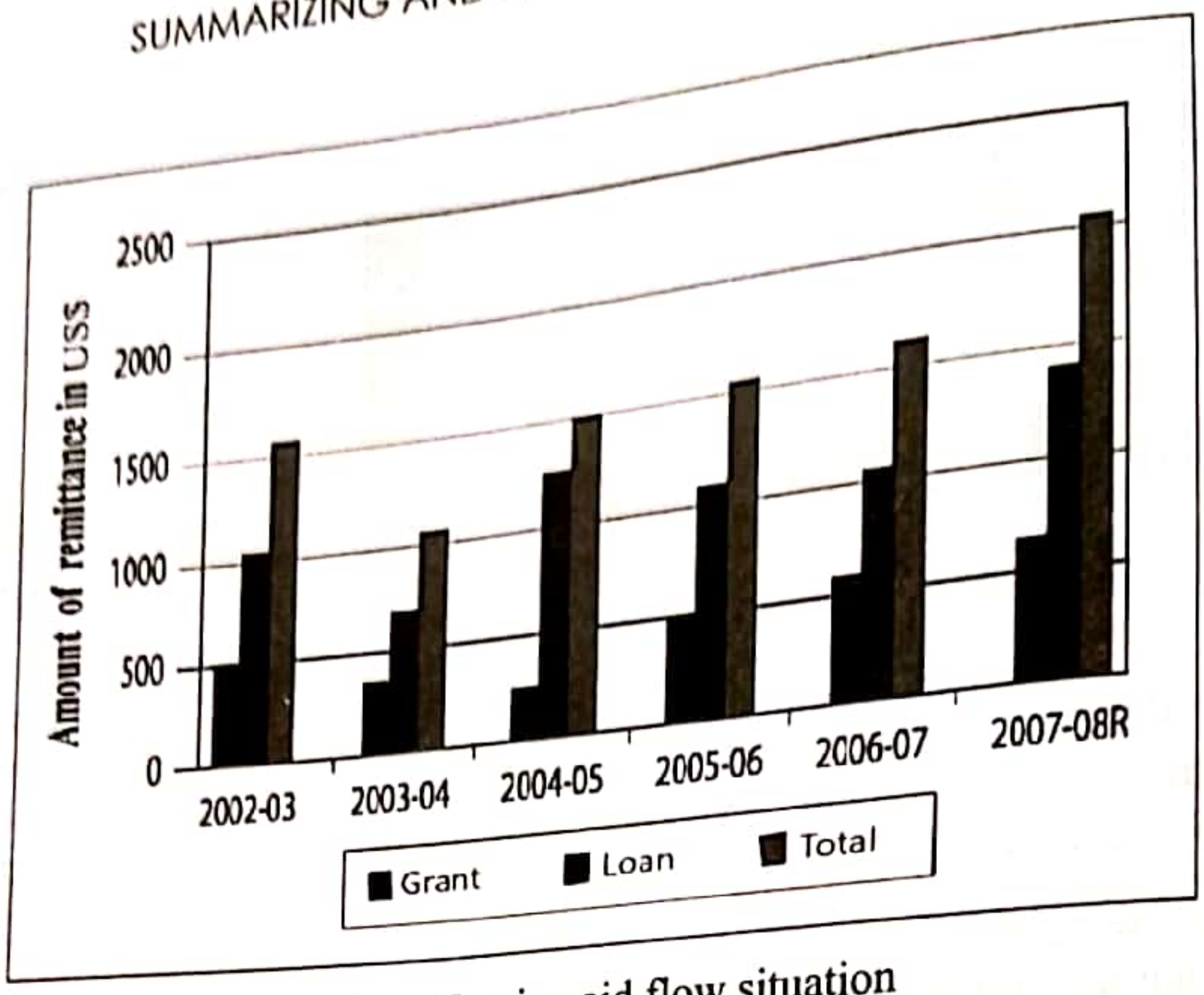


Figure 2.22: Foreign aid flow situation

Example 2.26: The trends in average labor productivity, 1995–96 prices have been displayed by a cluster bar below in Figure 2.23. Examine the figure and interpret the results.

Solution: Looking at the sectoral level, we find that the average labor productivity in manufacturing is the highest followed by services. Average productivity is the lowest in agriculture, which pulls down the average productivity. The diagram below describes the trend more succinctly.

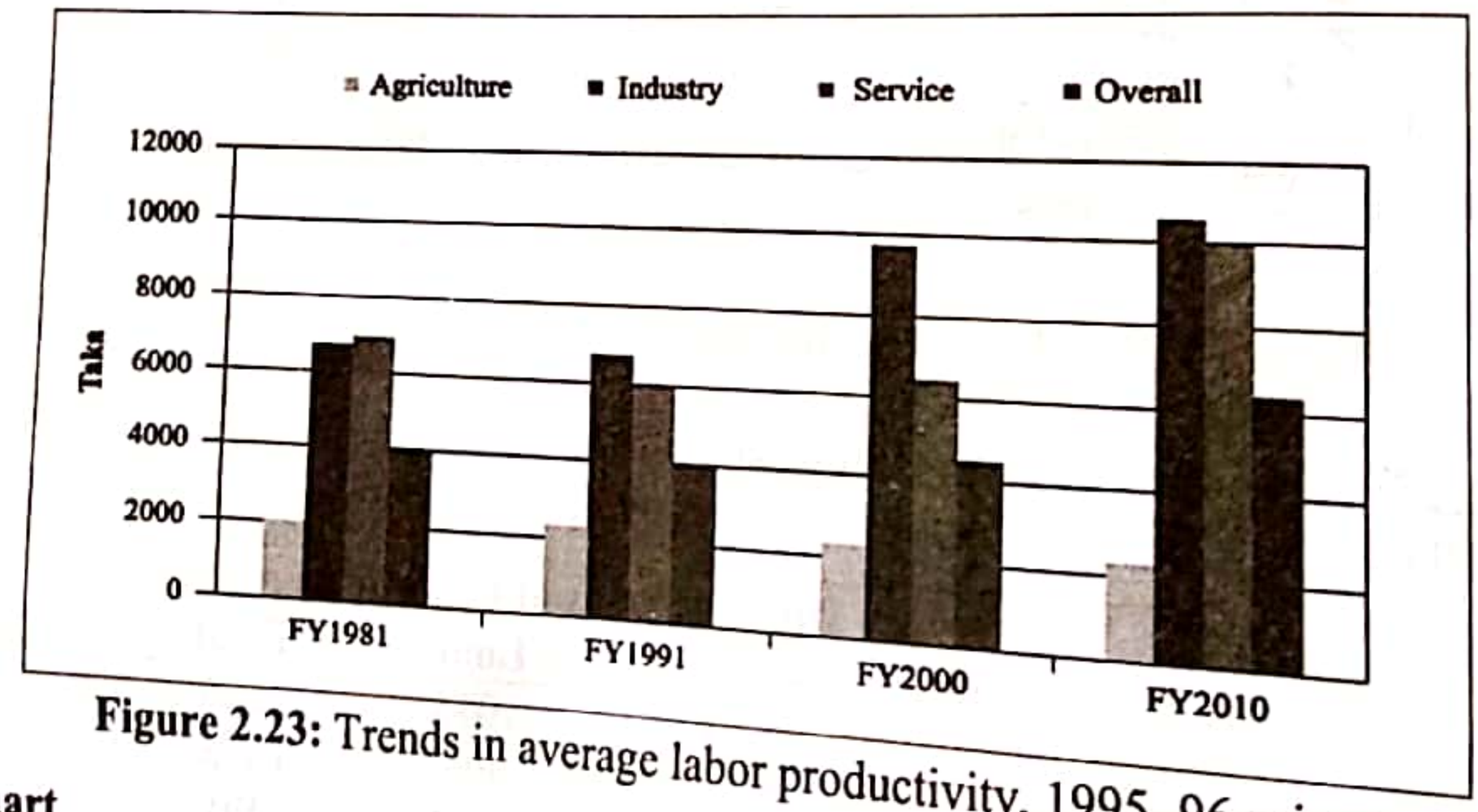


Figure 2.23: Trends in average labor productivity, 1995–96 prices

(d) Pie Chart

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.

Note: The various parts of the pie chart drawn may be identified either by angles in degrees or in percentage values.

Example 2.27: The table below shows the job opportunities in different sectors of Bangladesh as reported in Bangladesh Protidin, a daily newspaper in its recent issue. Draw a pie chart to represent the data.

Sectors	**Percent of jobs**	**Angles in degrees**
Govt.	4	14.4
Non-govt.	14	50.4
Private	61	219.6
Others	21	75.6
Total	**100**	**360.0**

Solution: To draw pie diagram, we recall that the given data, whether it is given in the form of frequency or percentages, must first be converted into a degrees to form the angles. Here we are given percentages of the sectors. We convert them as follows:

$$\text{Angle} = \frac{\text{Percent value}}{100} \times 360^0$$

Thus, for non-government sector, for example, $\text{Angle} = \frac{14}{100} \times 360 = 50.4^0$

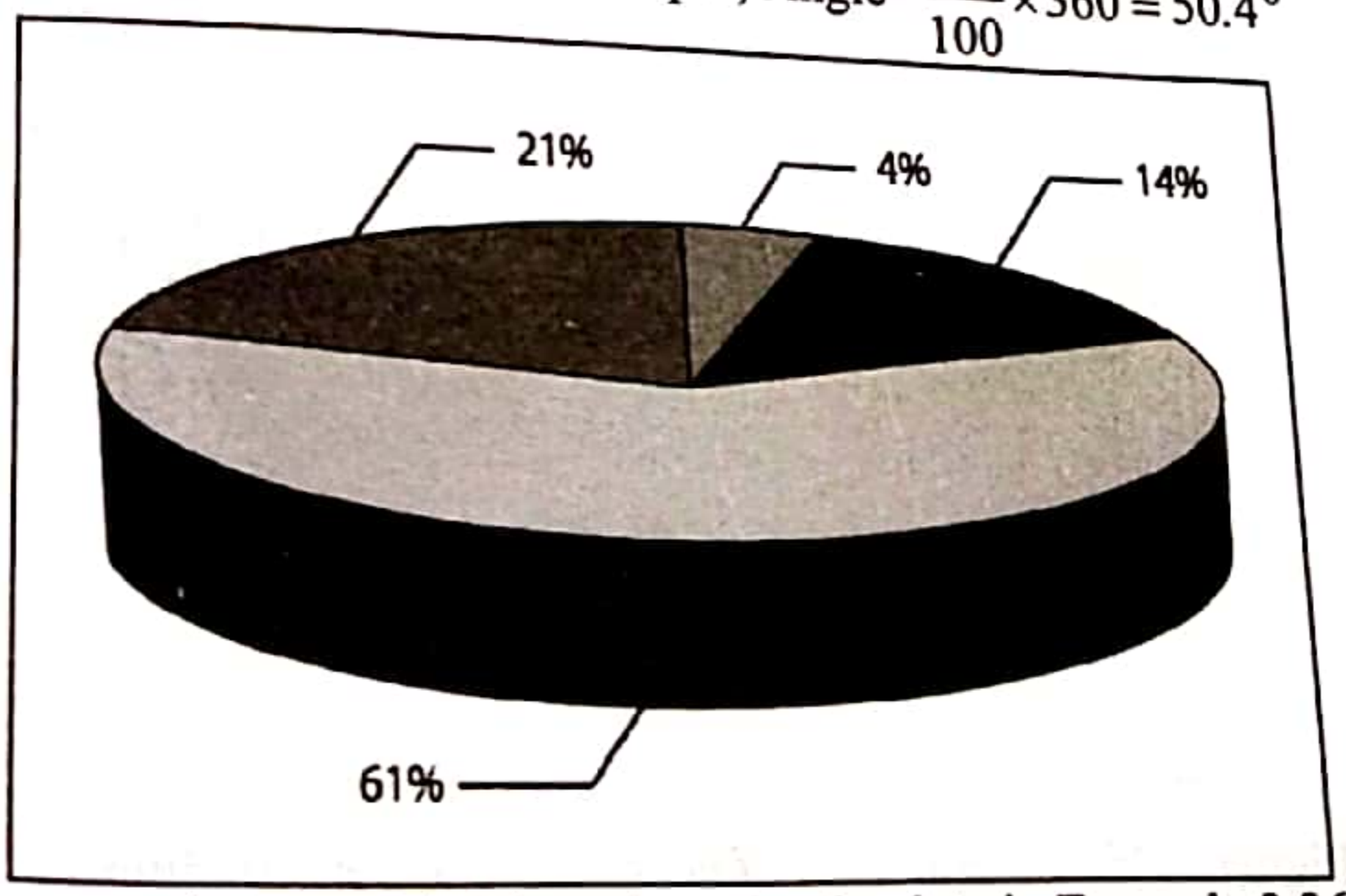


Figure 2.24: Pie chart for the data in Example 2.26

In the diagram, you might opt for values of the angles as shown in the last column of the table instead of percentage values. For non-govt. sector, for example, you may replace 14% by 50.4^0.

Interpretation: The pie chart shows that among the four sectors listed in the aforesaid table, private sector is providing the highest proportion of jobs, while the govt. sector the least.

Example 2.28: The following estimates of different materials and components of constructing an apartment were available from the Annual Report of Asset Developments and Holdings Limited.

Components	Estimated costs in percentage	Angles in degrees
Cement	20.0	72.0
Timber	10.0	36.0
Steel	15.0	54.0
Bricks	15.0	54.0
Labor	25.0	90.0
Supervision	15.0	54.0
Total	**100.0**	**360.0**

Represent the data by a pie chart.

Solution: Here we convert the percentage values into angles in degrees. As before, the angles are determined simply dividing the each percent value by 100 and multiplying the resulting value by 360 and shown in the last column of the above table. For cement, for example, it is found as follows:

$$\frac{20}{100} \times 360^0 = 72^0$$

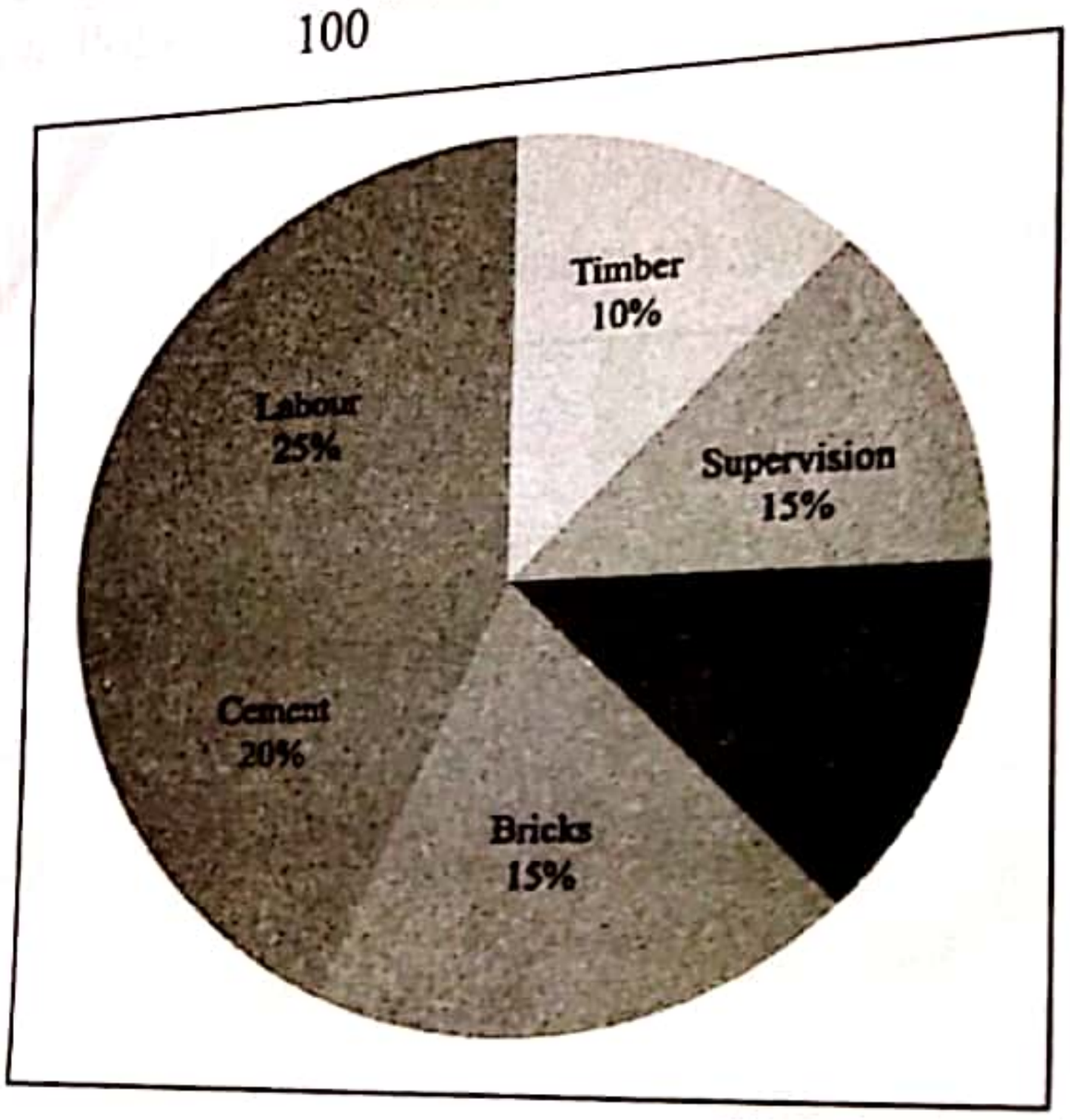


Figure 2.25: Components of cost of house construction

Example 2.29: Table below shows the percentage distribution of principal remittance receivers in Bangladesh on behalf of expatriates as reported in a survey conducted by Bangladesh Bureau of Statistics (BBS) in 2016. Represent the data by a pie chart. Calculate the angles in degrees in the blank column shown in the accompanying table.

Receiver type	Percentage of receivers	Angles in degrees
Spouse	41.78	150.4
Offspring	2.22	8.0
Parents	44.18	159.0
Siblings	9.87	35.5
Others	1.95	7.0
Total	**100.0**	**360**

Solution: As before, we convert the percentage values into angles in degrees and construct the desired chart as shown in Figure 2.26 below.

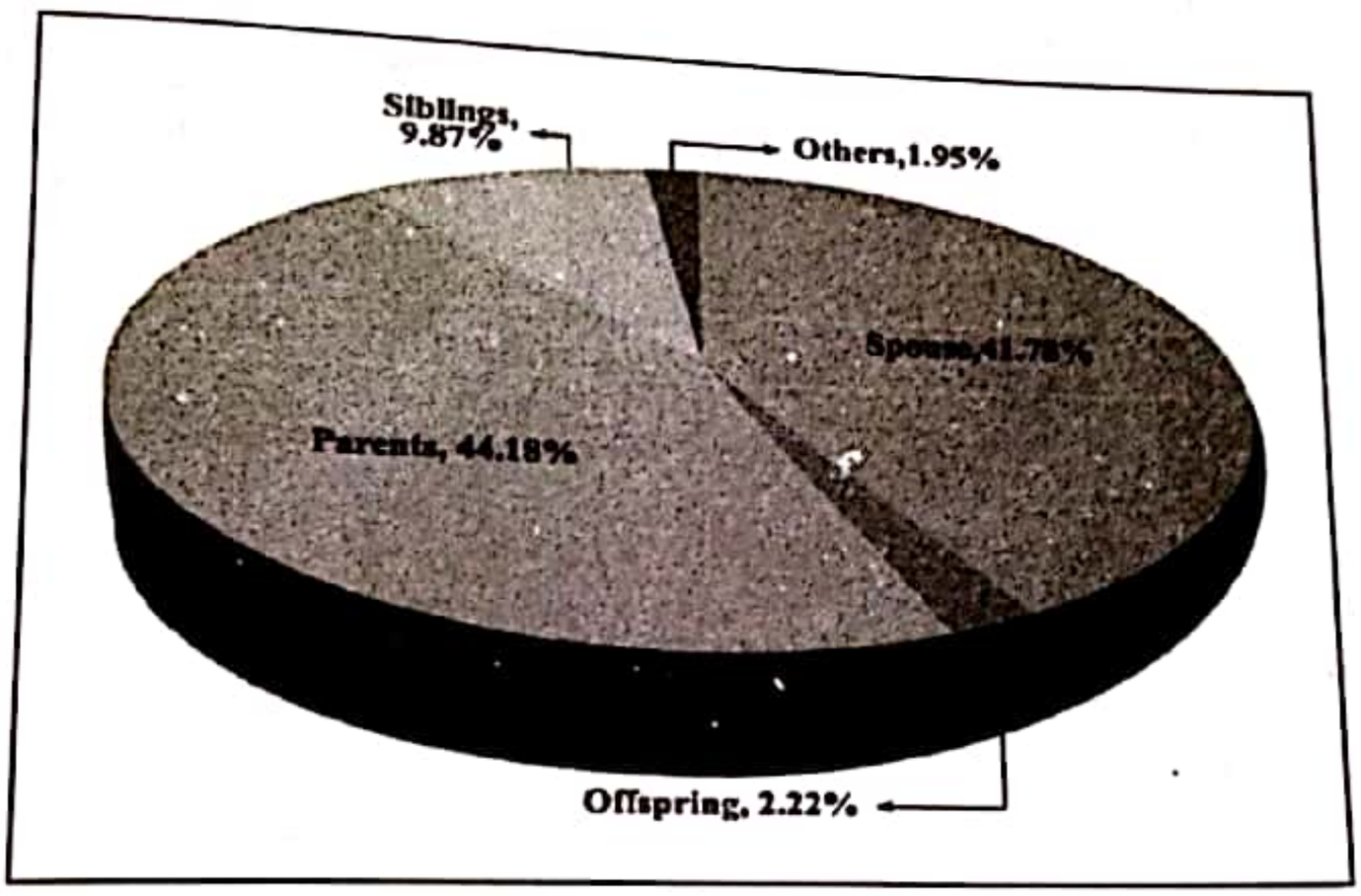


Figure 2.26: Pie chart showing the remittance receiver

An alternative way of presenting the diagram is shown Figure 2.27 below:

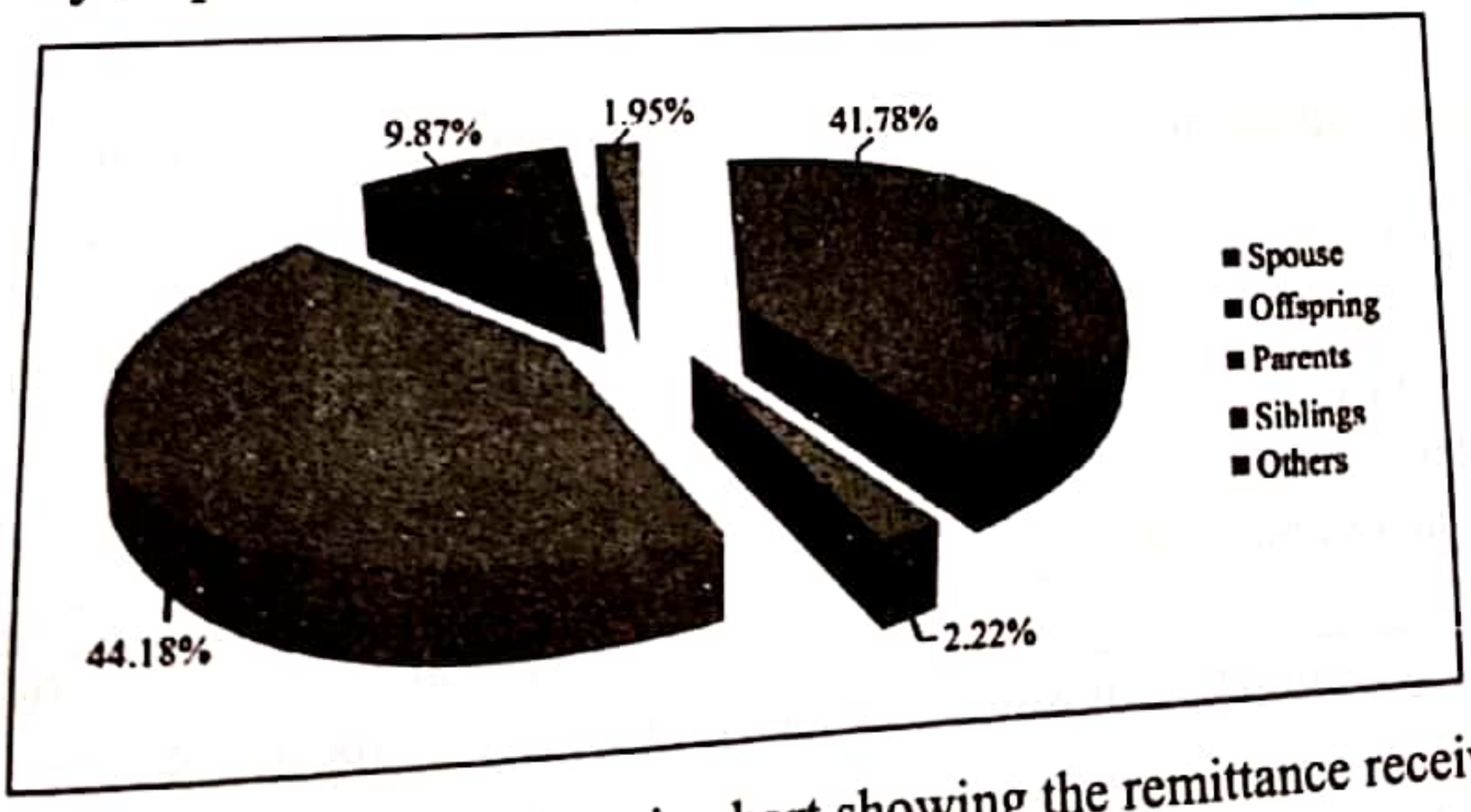


Figure 2.27: An alternative pie chart showing the remittance receiver

Example 2.30: The accompanying table shows the percentages of Bangladeshi expatriates in different countries by categories in 2018. Display the data by a pie chart

Category	Expatriates	Percent	Angle
Professional	2673	0.4	1.4
Skilled	317528	43.2	155.5
Semi-skilled	117734	16.0	57.6
Less-skilled	283002	38.5	138.7
Others	13244	1.9	6.8
Total	**734181**	**100.0**	**360**

Source Bangladesh Economic Review, 2018, pp: 36

Solution: First convert the percentages into angles (see the last column) and proceed in usual way. The resulting pie chart is shown in Figure 2.28.

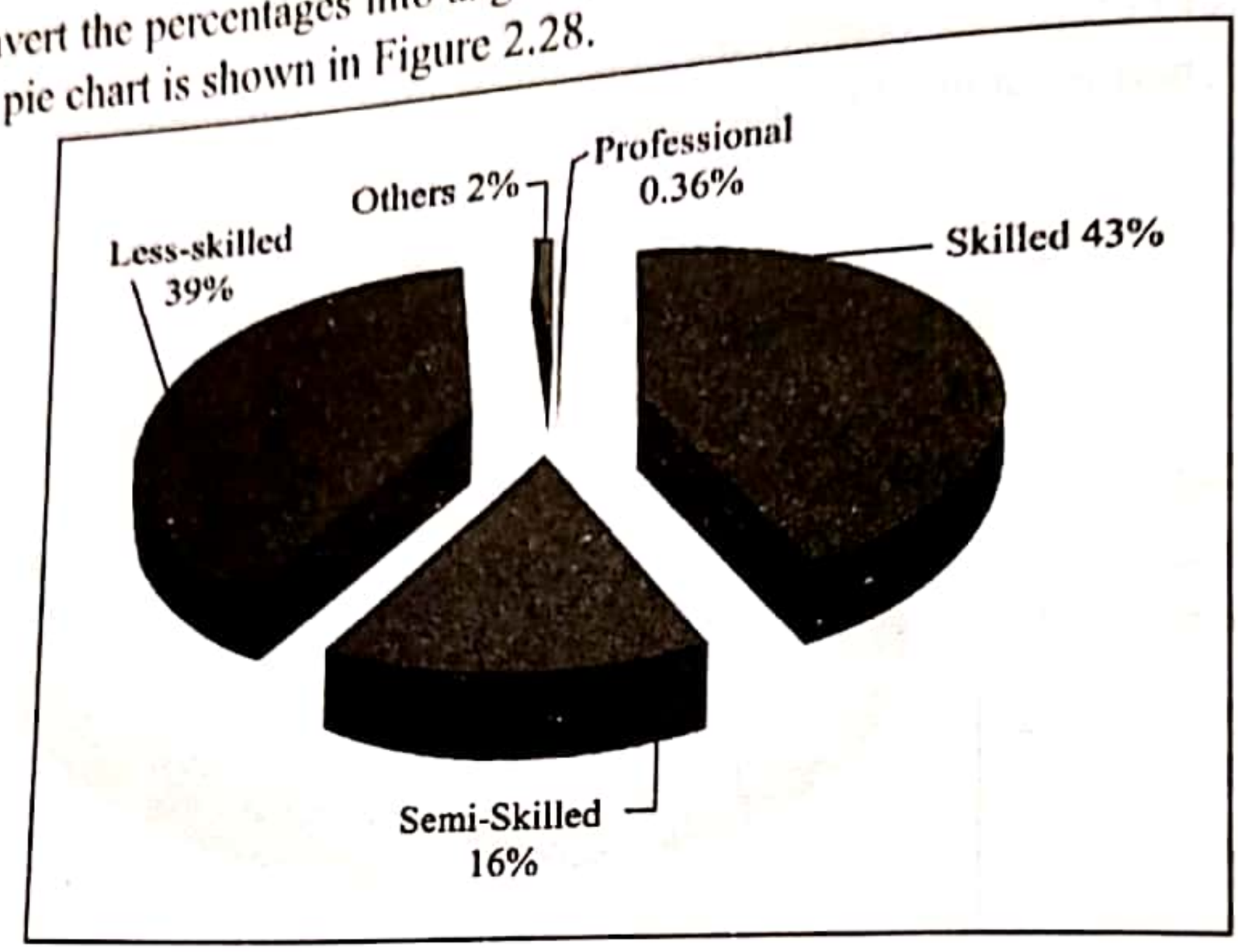


Figure 2.28: Pie chart showing the data on Bangladeshi expatriates

Example 2.31: Item-wise revenue collection in Bangladesh (in crore taka) for FY 2007–08 is shown in the accompanying table. Construct a pie chart to represent the data

Items of collection	FY 2007–08	Percent	Angles
VAT	17657.8	37.34	134.42
Import duty	9618.6	20.34	73.22
Income tax	11595.3	24.52	88.38
Excise duty	212.8	0.45	1.62
Supplementary duty	7755.4	16.40	59.04
Other taxes and duties	449.2	0.95	3.42
Total	**47288.1**	**100.0**	**360**

Source: Bangladesh Economic Review, 2008, pp: 42

Solution: As you see, there are different variants of pie char. We employ a new one to represent the data as you can see in Figure 2.29 below.

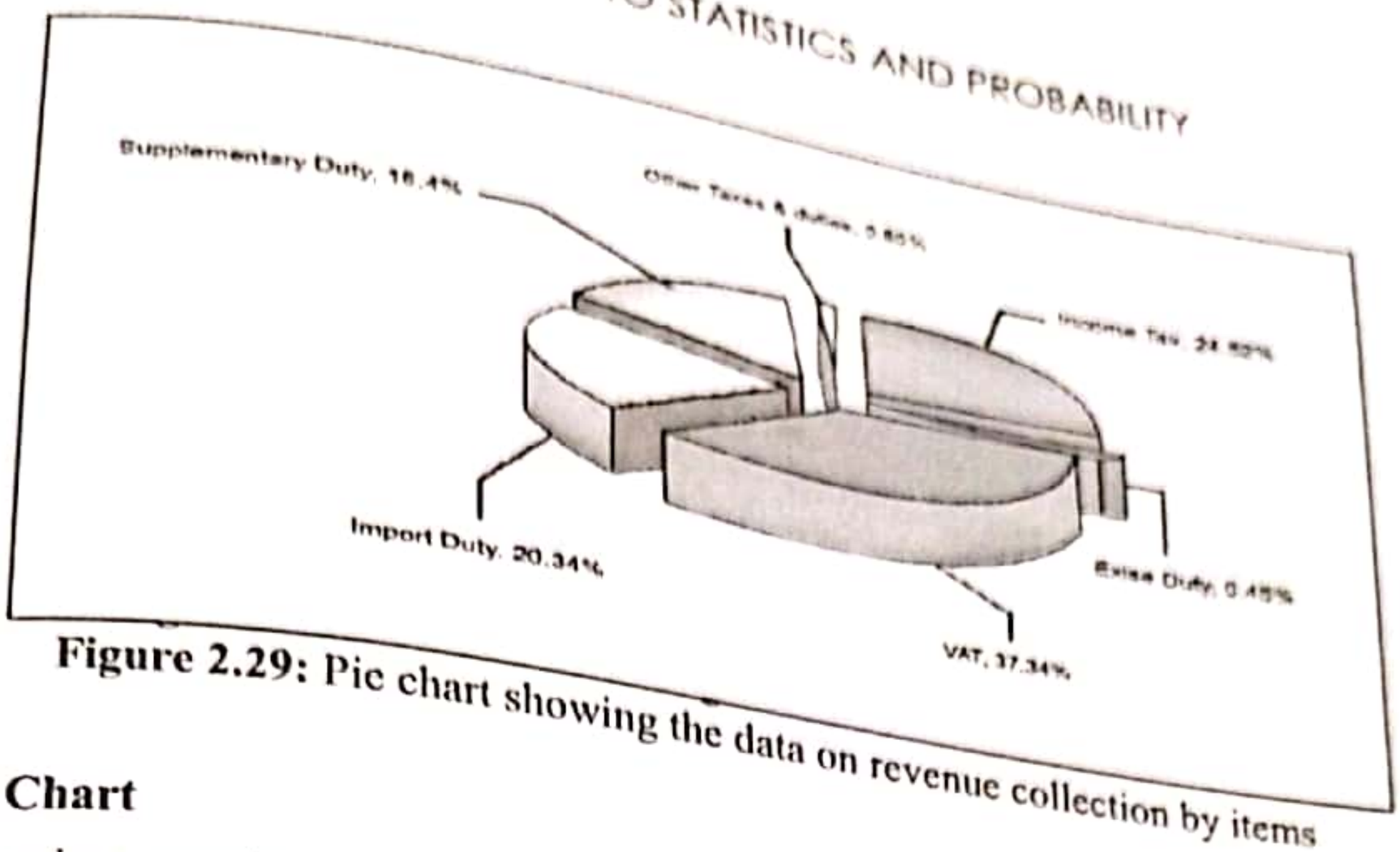


Figure 2.29: Pie chart showing the data on revenue collection by items

(e) Pareto Chart

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.

The Pareto chart is one of the seven basic tools of quality control. The independent variables on the chart are shown on the horizontal axis and the dependent variables are portrayed as the heights of bars. A point-to-point graph, which shows the cumulative relative frequency, may be superimposed on the bar graph. Because the values of the statistical variables are placed in order of relative frequency, the graph clearly reveals which factors have the greatest impact and where attention is likely to yield the greatest benefit.

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. Vilfredo Pareto found that roughly 80% of the wealth was held by 20% of the population. Therefore, this principle is known as the 80/20 principle. The chart achieves its greatest utility when the categorical variable of interest contains many categories.

Let us illustrate the importance of Pareto chart by an example. Suppose a clothing store was observing a steady decline in business. Before the manager did a customer survey, he assumed the decline was due to customer dissatisfaction with the clothing line he was selling and he blamed his supply chain for his problems. After charting the frequency of the answers in his customer survey, however, it was very clear that the real reasons for the decline of his business had nothing to do with his supply chain. By collecting data and displaying the same

in a Pareto chart, the manager could see which variables were having the most influence. In this example, parking difficulties, rude sales people and poor lighting were hurting his business most. Following the Pareto principle, those were the areas where he should focus his attention to build his business back up.

We can apply Pareto's principle to almost anything:

- If you have many products, 80% 0f your sales come from 20% of your product;
- 80% of customer problems relate to 20% of the issues;
- 80% of the defects are due to 20% of the problems;
- 80% of the complaints are due to 20% of the defects;
- 20% of the defects cause 80% of the problems;
- 80% of customer complaints arise from 20% of the products and services;
- 20% of the product and services account for 80% of the profit;
- 20% of the sales forces produces 80% of the company revenues;
- 20% of a system defects cause 80% of its problems. .

In general, this phenomenon can be interpreted as follows: Roughly 80% of the problems will be due to 20% of the causes, or the majority of the issues will be due to a small number of causes.

Pareto chart helps project managers to identify the causes of most of the problems the process is facing. It also helps management to prioritize tasks and activities. Being a variant of a bar chart, it is simple to draw, use and communicate problems to stakeholders

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) Create a vertical bar chart with causes on the x-axis and count (frequency) on the y-axis.
(d) Arrange the bar chart in descending order of cause importance, that is, cause with the highest frequency first.
(e) Calculate the cumulative frequency (count) for each cause in descending order.
(f) Calculate the cumulative count (frequency) percentage for each cause in descending order.
(g) Create a second y-axis with percentages descending in increments of 10 from 100% to 0%.
(h) Plot the cumulative count percentage of each cause on the x-axis.
(i) Join the points to form a curve.
(j) Draw a line at 80% on the y-axis running parallel to the x-axis. Then drop the line at the point of intersection with the curve on the x-axis. This point on the x-axis separates the important causes on the left (vital few) from the less important causes on the right (trivial many)

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 2.32: 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.

Types of defect	Frequency
Printing error	35
Crooked label	80
Wrinkled label	20
Smudged label	15
Loose label	25
Missing label	45
Total	**220**

Solution: To develop a Pareto chart, the following table is constructed: The resulting Pareto chart appears in Figure 2.30

Table 2.31: Table for Constructing Pareto Chart

Labeling defect	Frequency	%	Cumulative frequency	Cumulative %
Crooked label	80	36.36	80	36.36
Missing label	45	20.45	125	56.81
Printing error	35	15.91	1160	72.72
Loose label	25	11.36	185	84.08
Wrinkled label	20	9.09	205	93.17
Smudged label	15	6.83	220	100.0
Total	**220**	**100.0**	–	–

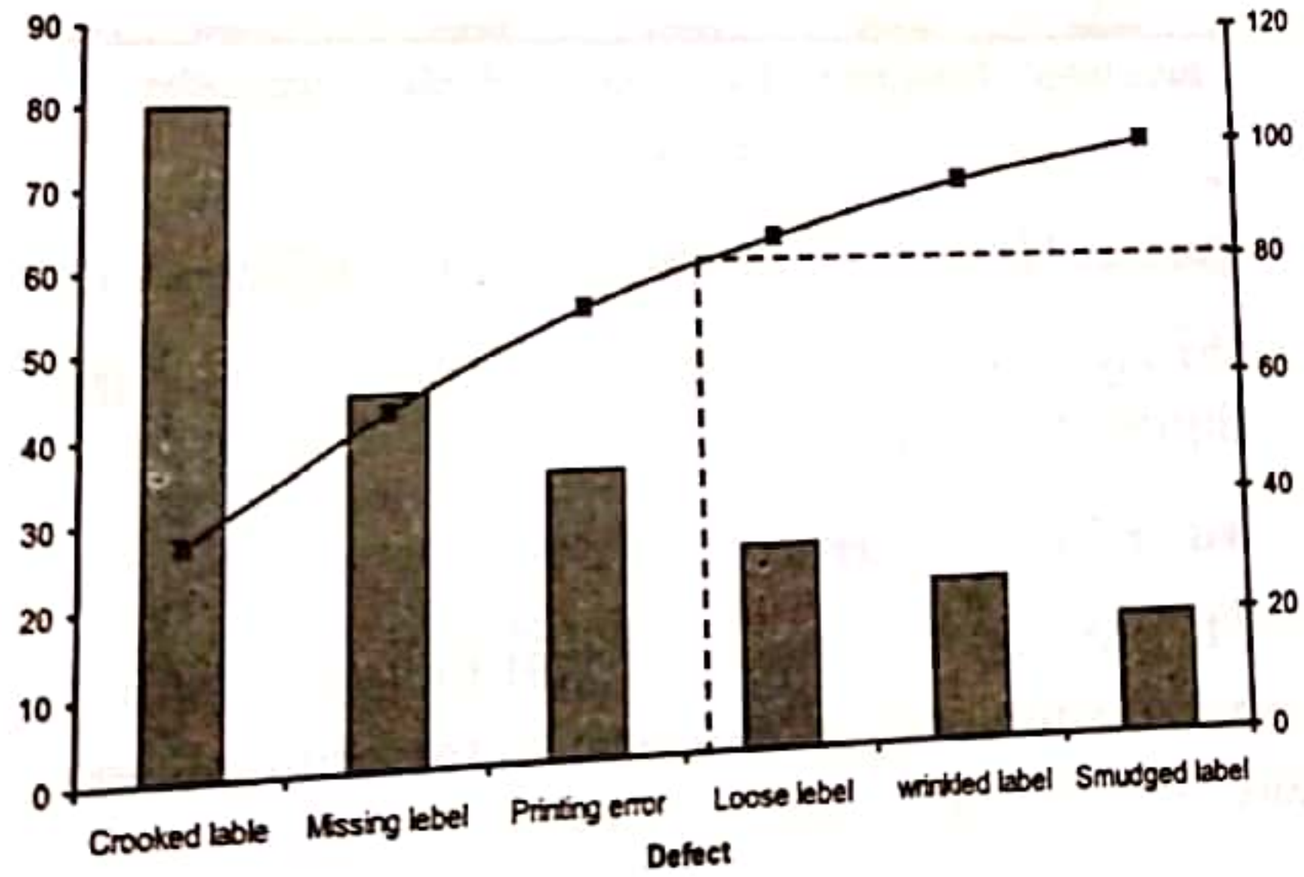


Figure 2.30: Pareto chart for labeling data

Interpretation: 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 (72.7%) of the labeling defects.

Example 2..33: Upon scrutinizing, the following types of defects were detected in 600 ready made shirts prepared by a local readymade garments manufacturing industry in Tongi. Use a Pareto chart to represent the data and interpret your results.

Type of defects	Frequency	Percent	Cumulative frequency	% cumulative frequency
Button defects	230	38.3	230	38.3
Pocket defects	160	26.7	390	65.0
Color defects	100	16.7	490	81.7
Cuff defects	70	11.7	560	93.4
Sleeve defects	40	6.6	600	100.0

Solution: The resulting chart appears below in Figure 2.31.

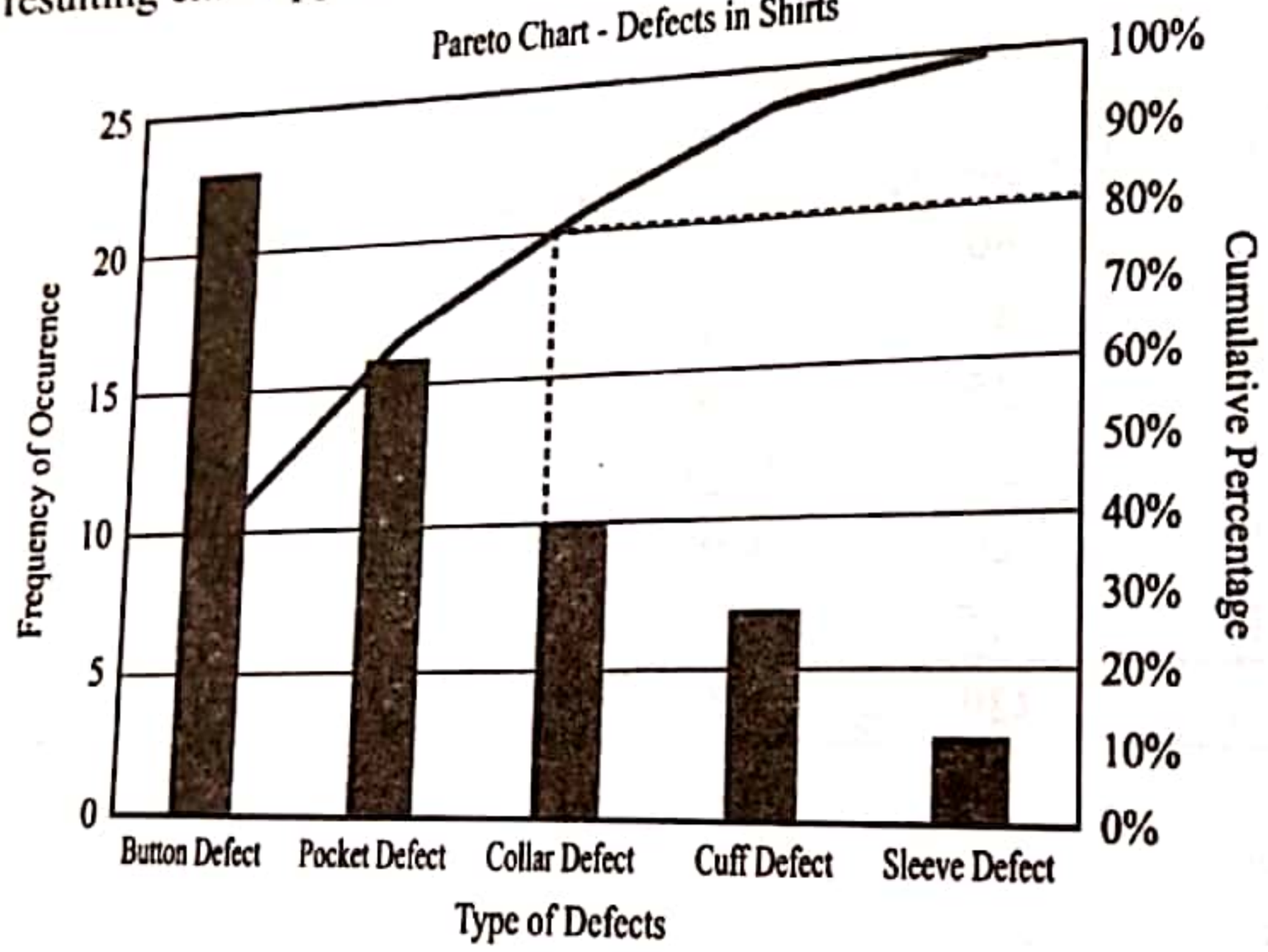


Figure 2..31: Pareto chart displaying the defects in shirt

The Pareto chart tends to say that button defects and pocket defects are the 'vital few', which account for about two thirds of the total defects.

(f) Combination of 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 partitioned into two or more components. Here is an example of this type:

Example 2.34: The percentages of child laborers in five administrative divisions of Bangladesh as reported in 1996 Labor Force Survey were as follows:

Administrative Divisions	Child laborers (%)
Barishal	10.0
Chattogram	25.1
Dhaka	28.2
Khulna	11.5
Rajshahi	25.2
Total	**100.0**

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. Of the 28.2% laborers in Dhaka division, 10.1% were seasonal, 82.7% temporary and the remaining 7.2% permanent. Present the data by a suitable diagram.

We use Figure 2.32 to represent the above data.

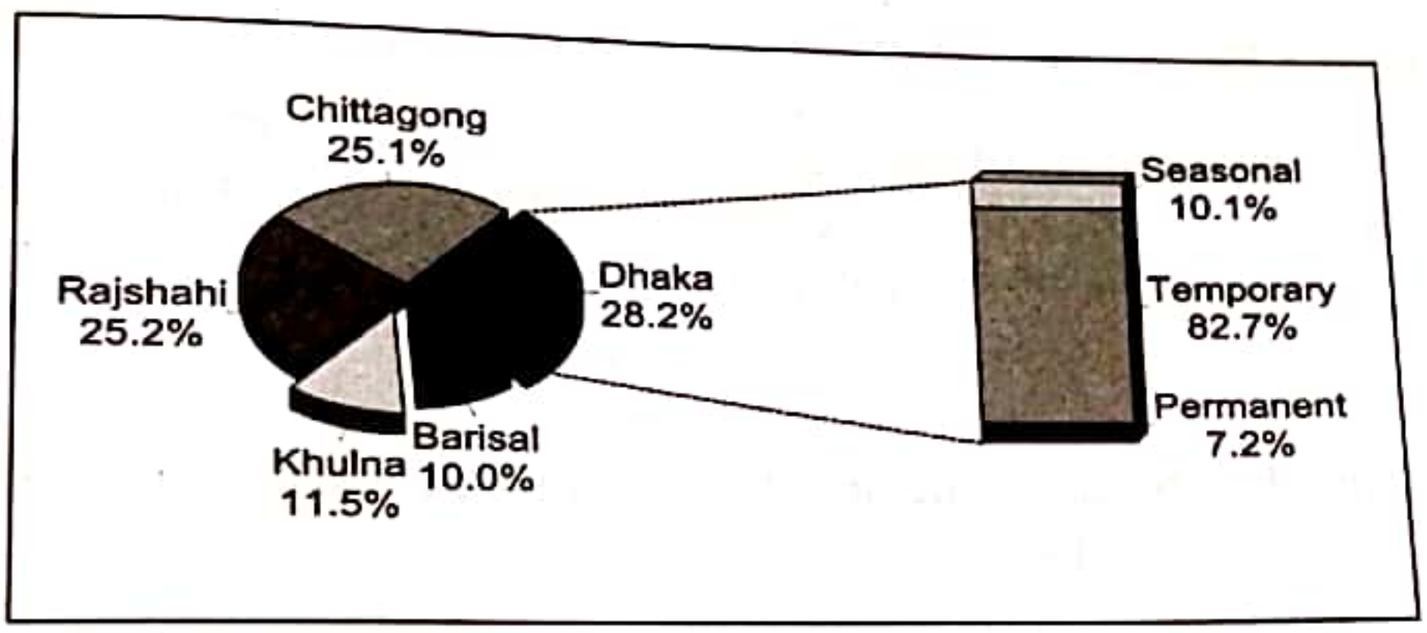


Figure 2.32: Pie-cum-stacked bar chart for data in Example 3.32

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.

(g) Dot Plots

When sample sizes are small, it is difficult (if not impossible) to construct meaningful histograms. However, a stem and leaf plot can sometimes be informative even though the sample is small. Alternatively, another graphical device that is particularly useful for small data set is the so called **dot plot**.

A **dot plot** is a type of graphic display used to compare frequency counts within categories or groups. As you might guess, a dot plot is made up of dots plotted on a graph. Here is how to interpret a dot plot.

- Each dot can represent a single observation from a set of data, or a specified number of observations from a set of data.
- The dots are stacked in a column over a category, so that the height of the column represents the relative or absolute frequency of observations in the category.

Here is an example to show what a dot plot looks like and how to interpret it.

Example 2.35: Suppose 30 first graders were asked to pick their favorite color. Their choices were as shown in the accompanying table. Summarize the data using a dot plot.

Solution: The plot is shown in Figure 2.33 below:

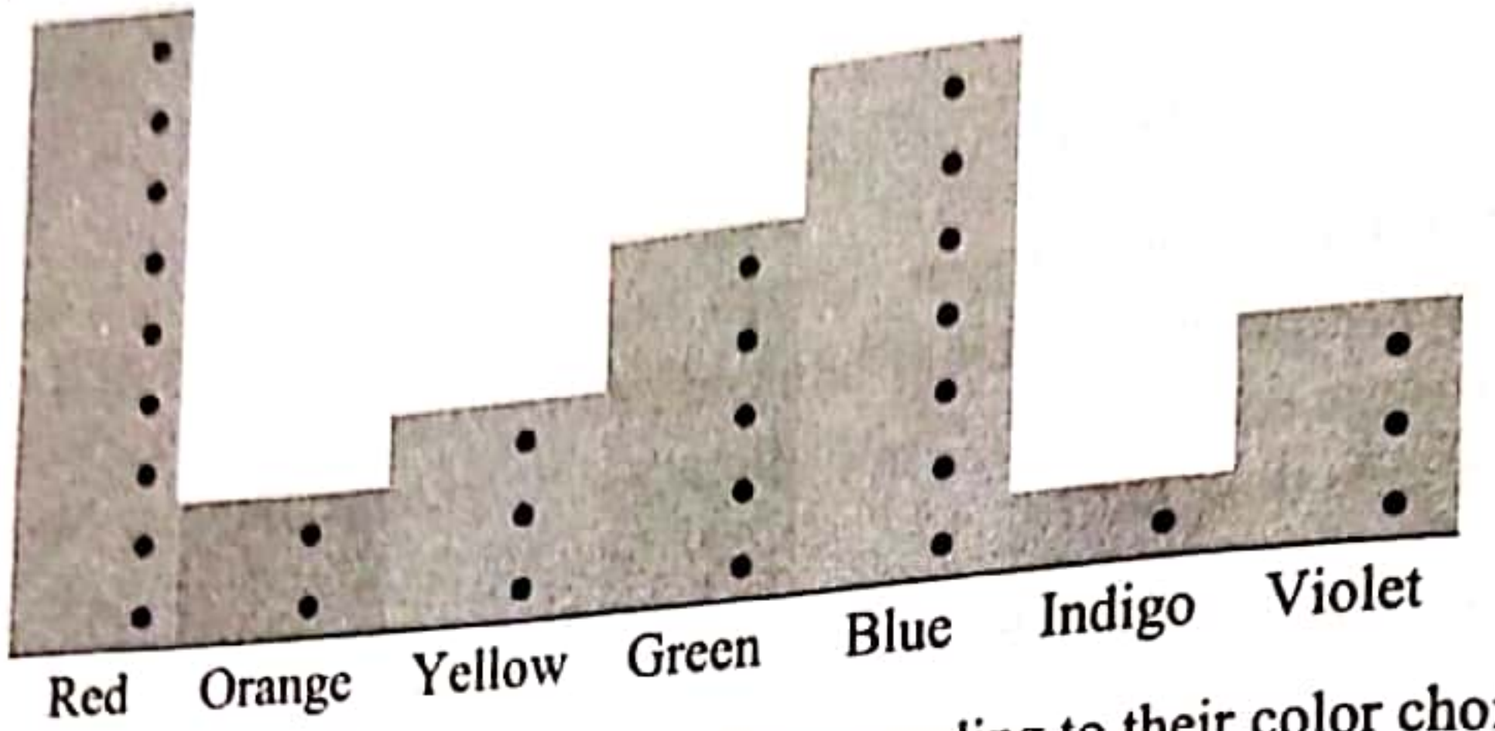


Figure 2.33: Dot plots of 30 first graders according to their color choice

Each dot represents one student, and the number of dots in a column represents the number of first graders who selected the color associated with that column. For example, red was the most popular color (selected by 9 students), followed by blue (selected by 7 students). Being selected by only 1 student, indigo was the least popular color.

Example 2.36: Table below shows the distribution of world's 102 countries by percentage of their population having excess to electricity facilities.

Access to Electricity (% of population, nearest 10%)	Number of Countries
10	5
20	6
30	12
40	5
50	4
60	5
70	6
80	10
90	15
100	34

The table says that, for example, there were 5 countries where only 10% of the people had access to electricity, 6 countries where 20% of the people had access to electricity, etc. Display the data by a dot plot.

Solution: The diagram is plotted in Figure 2.34. Look at the graph carefully. Each vertical plot has been represented by dots. You can verify that each dot represent one country in each column containing the dots.

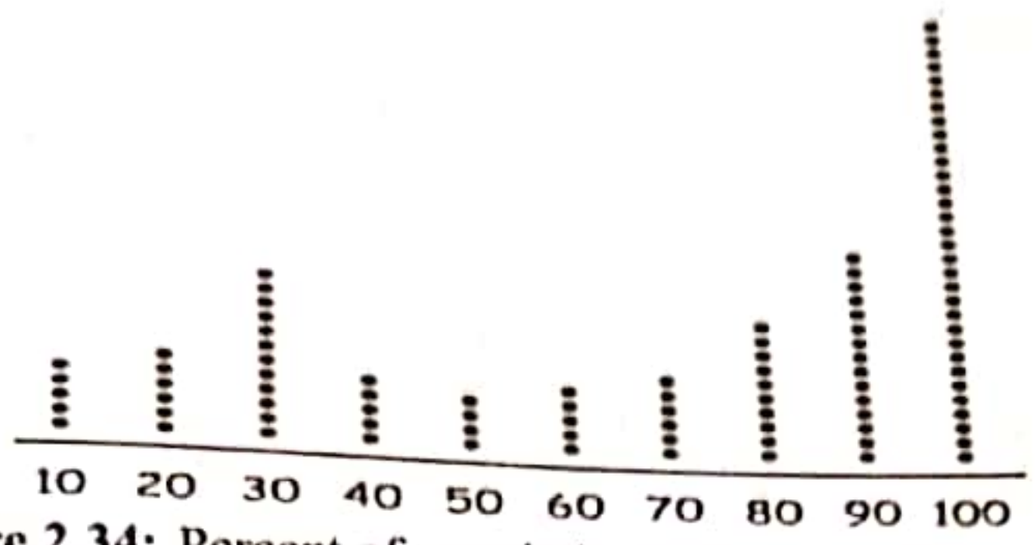


Figure 2.34: Percent of population with access to electricity

2.13 GRAPHICAL 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.

2.13.1 Presentation of Discrete Ungrouped Distribution

Discrete data may be presented by either dotted or continuous lines. We illustrate these presentations by some examples below.

Example 2.37: Suppose we have a survey data on number of members (family size) in 22 families as shown below:

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

The data in the above table may be represented either by dot plots or by continuous lines as shown in the accompanying graphs.

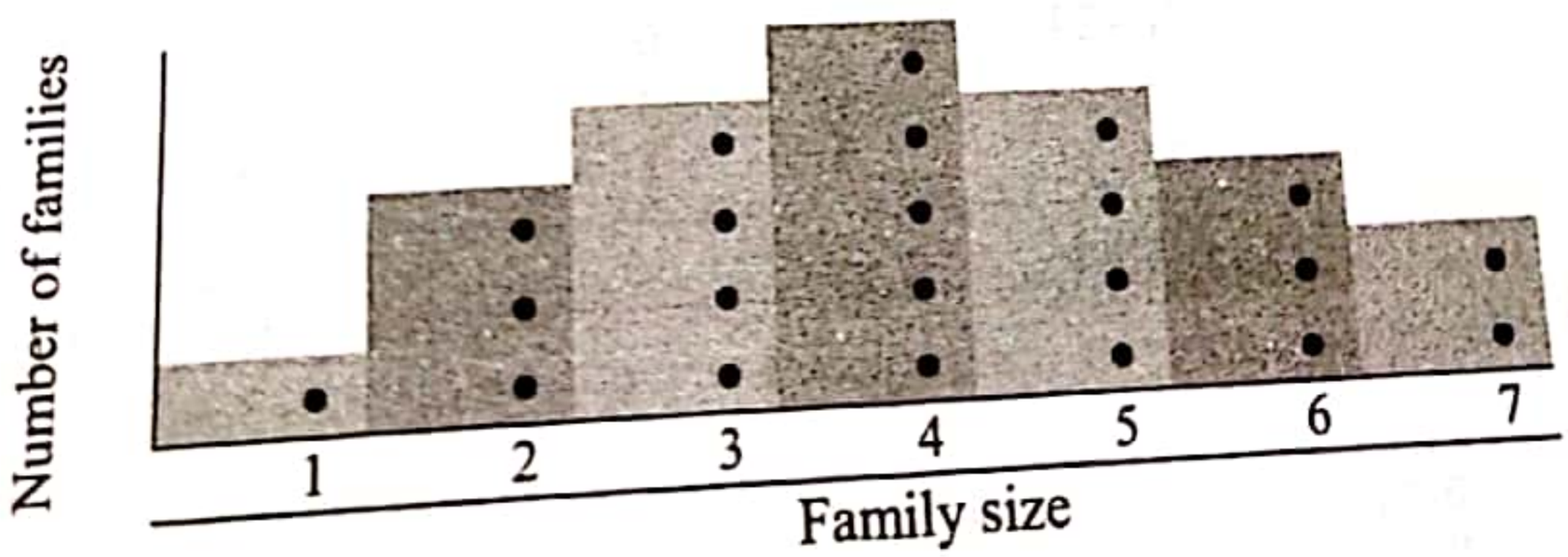


Figure 2.35: Dot diagram representing the family size data

The diagram is known as **dot diagram**. The total number of members in the 22 families is 91, as you can check as follows:

$$\text{Total members} = \text{Family size} \times \text{Number of dots}$$
$$= (1\times1) \times ((2\times3)+ (3\times4)+ (4\times5)+ (5\times4)+ (6\times3)+ (7\times2)=91$$

There are thus 91 members in the 22 families.

You can display the same data by continuous lines also. Such a diagram is called **line diagram**. Here is a line diagram as shown in Figure 2.36 for the above data:

Figure2.36: Line diagram for family size data

2.13.2 Presentation of Discrete Frequency Distribution

For a discrete frequency distribution, bar diagram can conveniently be employed keeping reasonable gap between successive bars owing to its discrete nature.

Example 2.38: Use the data in Example 2.4 that show the distribution of 50 workers by the number of days they were on medical leave due to their sickness. Represent the data by a bar chart.

Solution: We have here a discrete distribution, which may well be represented by a bar diagram. For ready reference, we reproduce the data here.

Days absent	Number of workers
5–9	4
10–14	15
15–19	21
20–24	6
25–29	3
30–34	1
Total	50

The diagram is as follows.

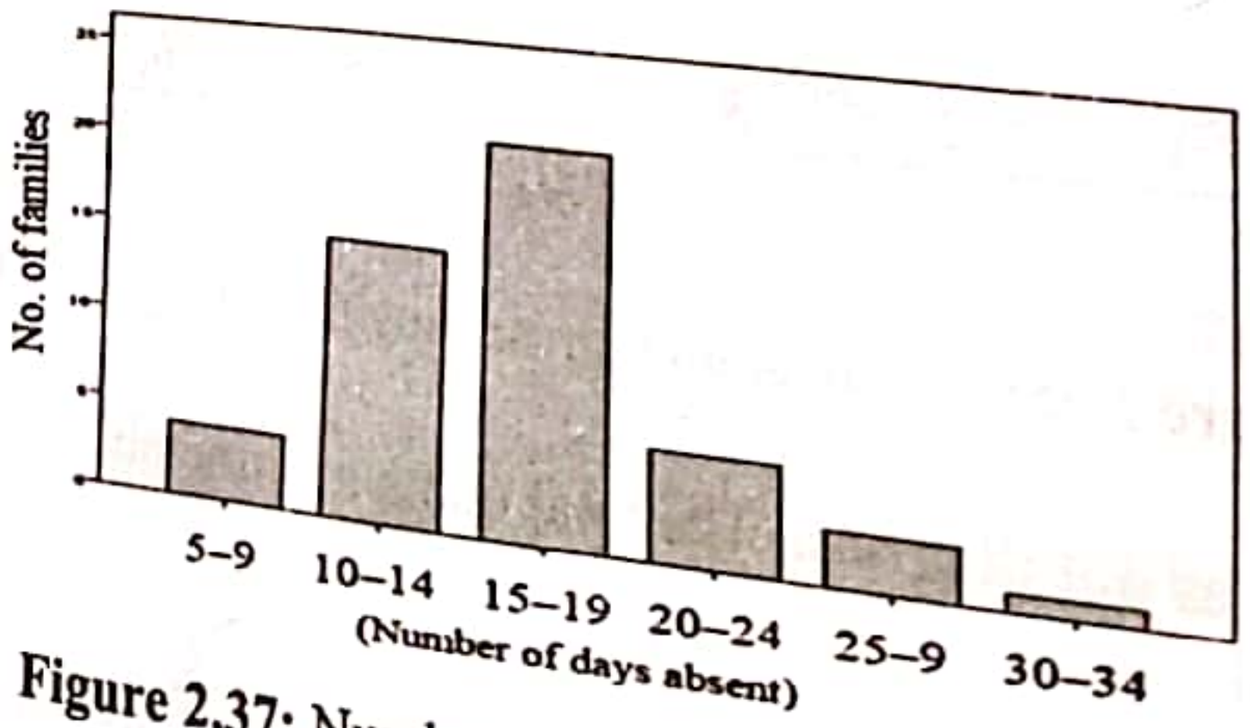


Figure 2.37: Number of employees by days absent

2.14 PRESENTATIO OF CONTINUOUS DATA

Once we develop frequency distributions, we are in a position to graph this information. The appropriate graphs to be considered here are the following:

(a) Histogram
(b) Frequency polygon
(c) Ogive

2.14.1 Histogram

The most common form of graphical presentation of a frequency distribution is the **frequency histogram**. A frequency 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. In such case, 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. These proportional frequencies are often called **frequency densities**.

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 6×5=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.

We now illustrate below how a histogram is constructed when we have equal as well as unequal class widths. We illustrate below first the case of equal class width.

Example 2.39: In a recent household income and expenditure survey among 80 urban families in Dhaka city revealed the following data on house rent paid by the families as percentage of their total income.

Represent the data by a histogram.

House rent (% of income in taka)	Class width	Frequency (no. of families	Heights of rectangles
4.5–9.5	5	8	8
9.5–14.5	5	29	29
14.5–19.5	5	27	27
19.5–24.5	5	12	12
24.5–29.5	5	4	4
Total	–	**80**	–

Solution: Here the class widths are of equal length. The vertical scale shows the number families. The histogram when constructed with the above data will look like as in Figure 2.

Figure 2.38: Histogram with equal class widths

We also illustrate the construction of histogram with unequal class boundaries with the same data as shown in Example 2.39 above.

Example 2.40: Construct a histogram by reconstructing the class widths of equal size into unequal size using the data in Example 2.39.

Class boundary	Class frequency	Class width	Height of rectangles
4.5–14.5	37	10	$37 \div 10 = 3.7$
14.5–19.5	27	5	$27 \div 5 = 5.4$
19.5–29.5	16	10	$16 \div 10 = 1.6$
Total	**80**	–	–

In the above table, the widths of class intervals are made unequal by combining the first and the second and then the third and the fourth intervals so that the widths in these two cases become 10. The heights of the rectangles in the last column are arrived at by dividing the class frequencies in the second column by the widths in the third column. The resulting histogram appears in Figure 2.39.

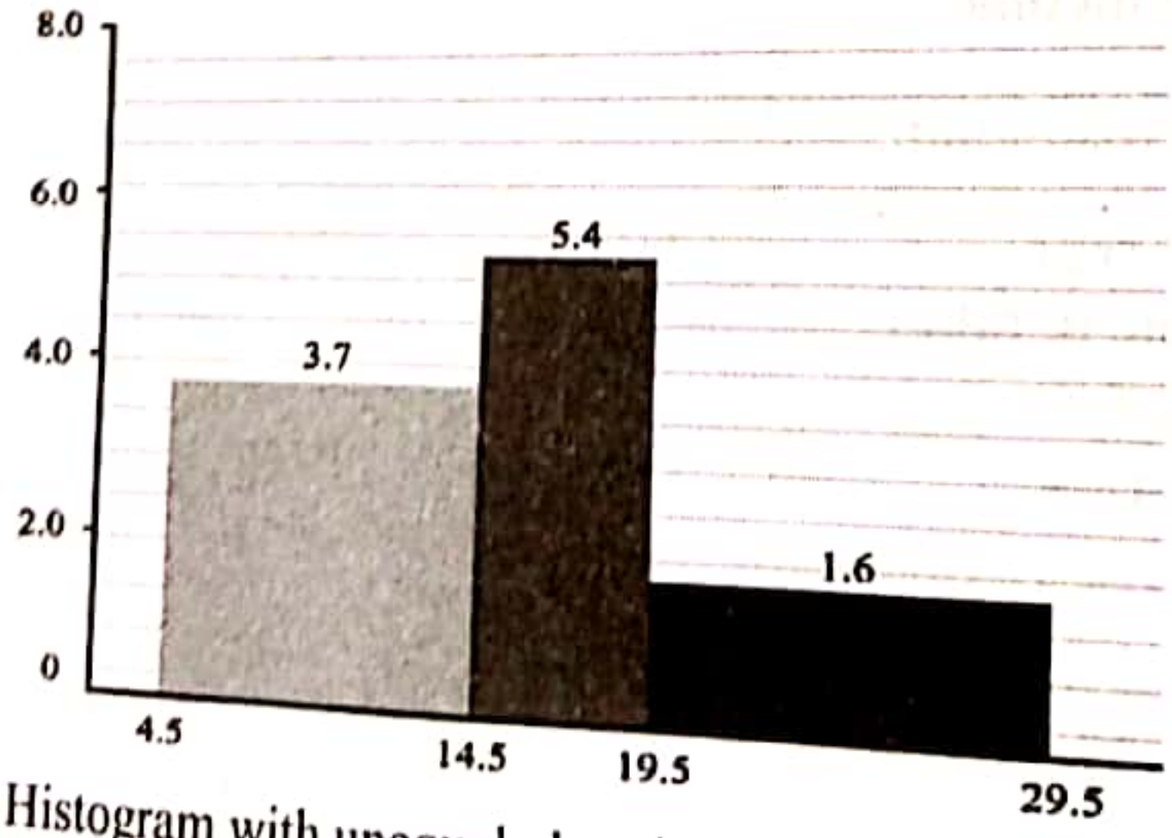


Figure 2.39: Histogram with unequal class intervals for data in Example 2.39

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) A histogram is constructed with ratio and interval level data, while a bar diagram is constructed for nominal and ordinal level data.

(b) The height of the rectangle in a histogram is proportional to the frequency of the class it represents, while the height of the bar in a bar diagram is equal to the the frequency or proportional to the magnitude it represents.

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

(d) The width of a bar diagram bears no significance, while in a histogram, the class width and the height relative to the class width represent the class frequency.

(b) 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 frequencies 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 closed figure. The frequency polygon is then formed by connecting the points with straight lines.

Example 2.41: Construct a frequency polygon with the data in Example 2.39 below on percentage of income paid for house rent by 80 families.

Table 2.32: Data for Constructing Frequency Polygon

Class boundary	Mid-values	Frequency
–0.5–4.5	2	0
4.5–9.5	7	8
9.5–14.5	12	29
14.5–19.5	17	27
19.5–24.5	22	12
24.5–29.5	27	4
29.5–34.5	32	0
Total	–	**80**

Solution: To construct a frequency polygon with the data in Example 2.39, we reconstruct the table under reference with classes of zero frequencies at each end of the table so constructed. Note that the lower limit of the first class has been made –0.5 to make the interval of equal length, although a negative value in the lower class limit (here) has no valid meaning in the present instance (see table above). The frequency polygon for the frequency distribution shown in Table 3.32 is displayed in Figure 2.40. It may be emphasized that you can draw a frequency polygon just by joining the mid-points of the rectangles of the histogram thereby removing the histogram. You will then be left with only the frequency polygon.

The frequency polygon drawn based on the data in Table 2.32 appears below in Figure 3.40

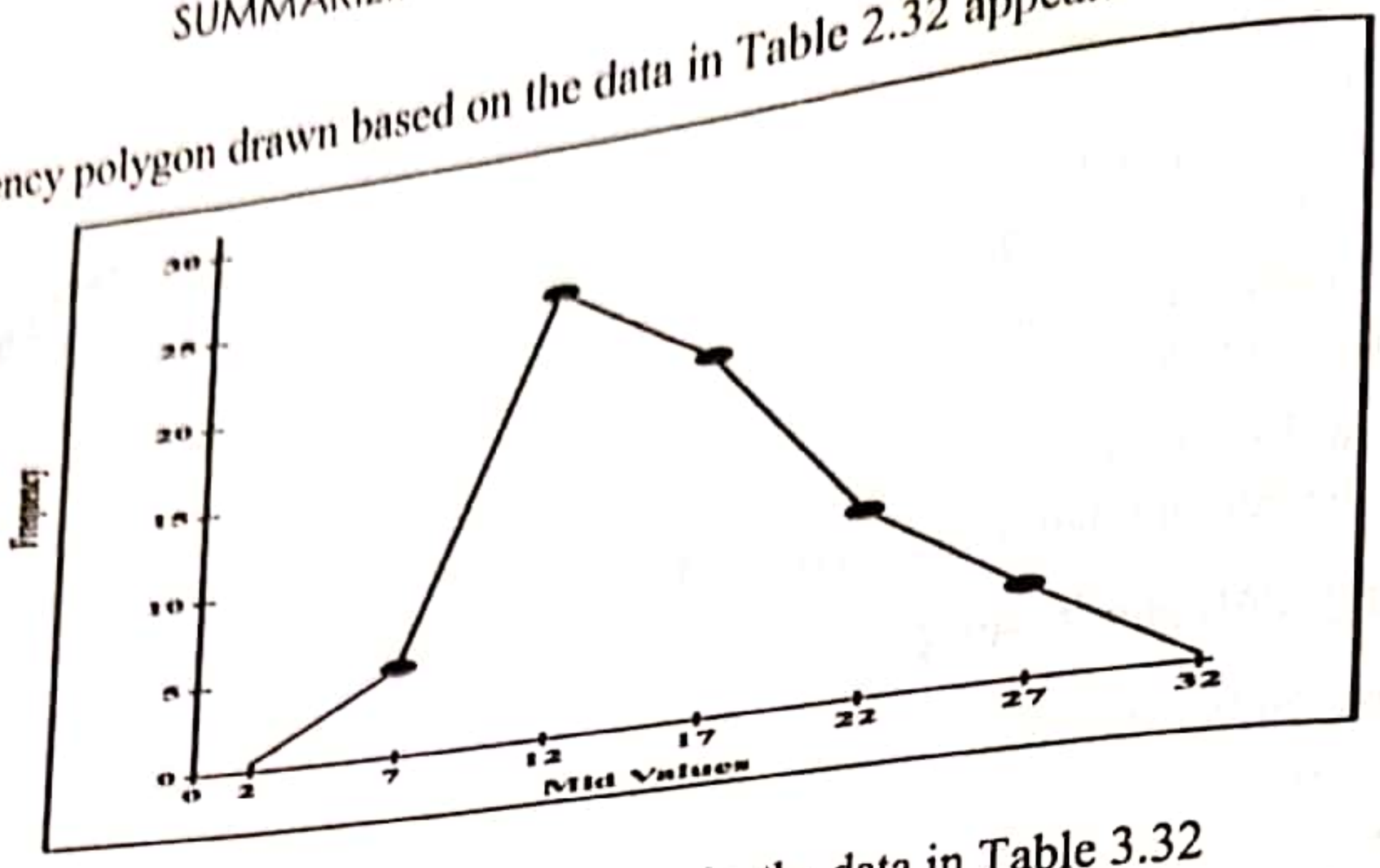


Figure 2.40: Frequency polygon for the data in Table 3.32

Example 2.42: The following data refer to the GDP growth rate (real) in Bangladesh for a period of 40 years from 1980 to 2019. Present data in a frequency distribution and display the resulting distribution by a frequency polygon.

Year	1980	1981	1982	1983	1984	1985	1986	1987
GDP (%)	3.1	5.6	3.2	4.6	4.2	3.7	4.0	2.9
Year	1988	1989	1990	1991	1992	1993	1994	1995
GDP (%)	2.4	4.3	4.6	4.2	4.8	4.3	4.5	4.8
Year	1996	1997	1998	1999	2000	2001	2002	2003
GDP (%)	5.0	5.3	5.0	5.4	5.6	4.8	4.8	5.8
Year	2004	2005	2006	2007	2008	2009	2010	2011
GDP (%)	6.1	6.3	6.9	6.5	5.5	5.3	6.0	6.5
Year	2012	2013	2014	2015	2016	2017	2018	2019
GDP (%)	6.3	6.0	6.3	6.8	7.2	7.6	7.9	8.1

Source: World Bank Report 2020: (internet)

You can check by actually tallying the values that the resulting frequency distribution is as follows:

Table 2.33: GDP Growth Rate (real)

GDP growth rate (%)	Mid-values	No. of years
1.90–2.90	2.4	1
2.90–3.90	3.4	4
3.90–4.90	4.4	12
4.90–5.90	5.4	9
5.90–6.90	6.4	9
6.90–7.90	7.4	3
7.90–8.90	8.4	2
Total	–	**40**

The resulting frequency polygon with the data in Table 2.33 is as follows:

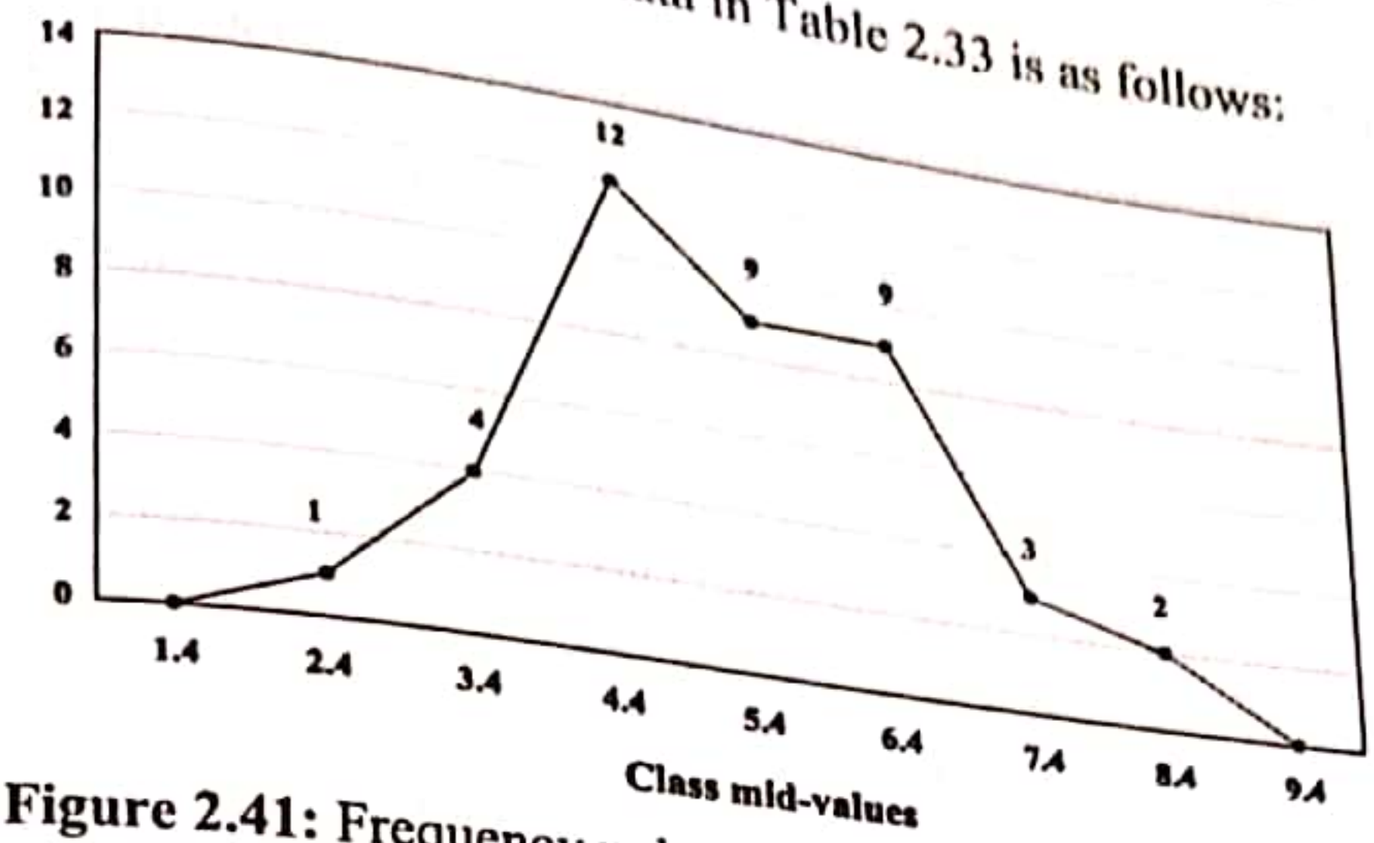


Figure 2.41: Frequency polygon for the GDP growth rate data

While the histogram and frequency polygon in our illustrated examples were based on the absolute frequency distribution, they could have been based on 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. The 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. Such a curve is sketched as an illustration with the frequency polygon as shown in Figure 3.42 below.

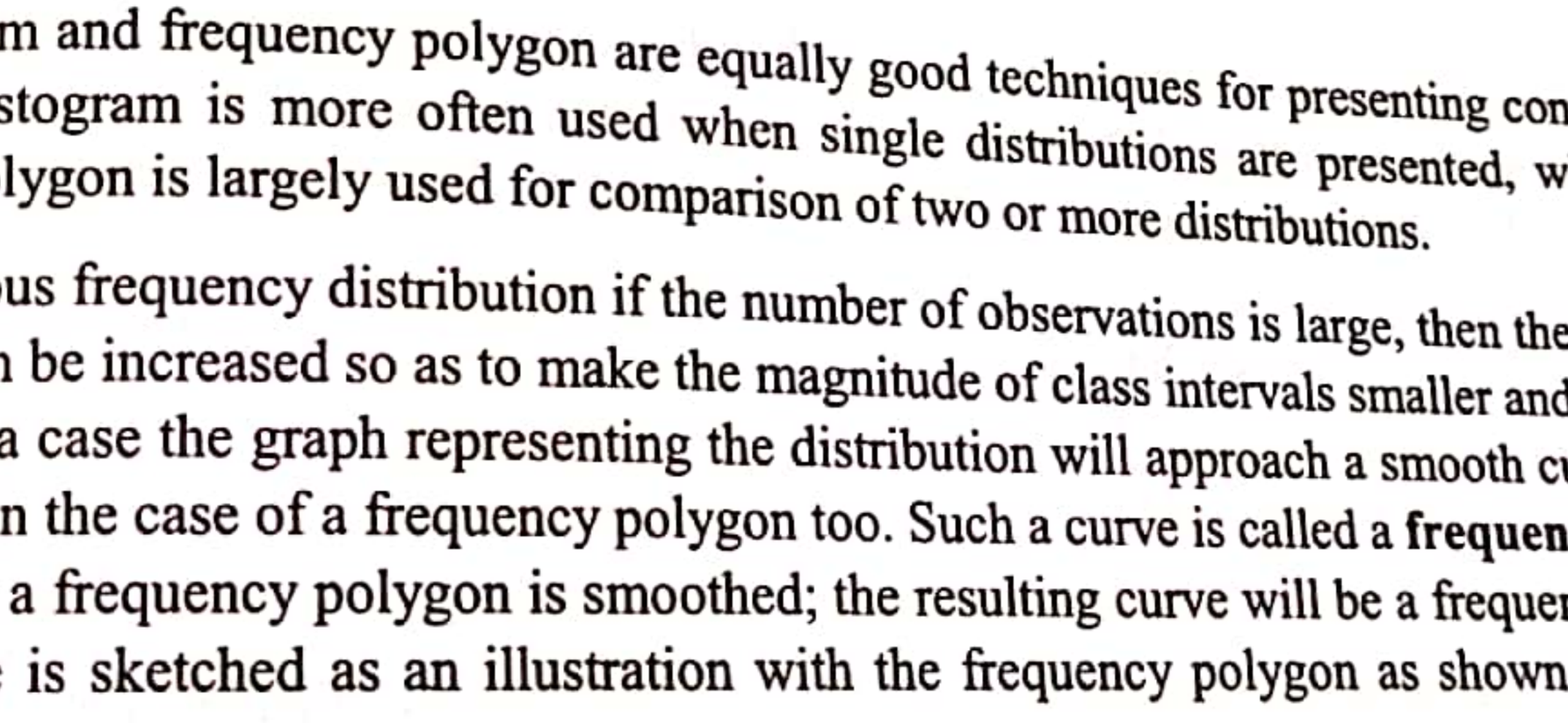


Figure 2.42: Frequency curve of Figure2.39

(c) Cumulative frequency polygon

A graph of the cumulative frequency distribution or cumulative relative frequency distribution is called an **ogive**. Two types of ogive can be constructed: more than type and less than type. To construct a **less than type ogive**, follow the steps below:

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

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

To construct a **more than type ogive**, the steps are as follows:

- Put the lower class limits on the horizontal axis.
- Plot a point against each lower class limit at a height corresponding to the cumulative frequency at that lower class limit.
- Plot an additional point above the upper class limit for the terminal class at a height of zero frequency.

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

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

We illustrate the construction of these ogives by an example.

Example 2.43: Refer to Example 2.7 (Table 2.22). The data display the longevity of 80 electric bulbs. Use the data to construct an ogive of both more than type and less than type. For instant citing, we reproduce the table below.

Longevity (in month)	Number of bulbs
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**

Solution: We employ the above data and construct Table 2.34 below to draw the required *ogives*. The resulting ogives are sketched in Figures 2.43 and 2.44.

Table 2.34: Cumulative Frequency Distributions for Less Than and More than Type Ogives Based on Data on Longevity of Electric Bulbs

Less than type		More than type	
Inflation rate (%)	Cumulative Frequency	Inflation rate (%)	Cumulative Frequency
Less than 1.45	0	1.45 or more	80
Less than 1.95	4	1.95 or more	76
Less than 2.45	6	2.45 or more	74
Less than 2.95	14	2.95 or more	66
Less than 3.45	44	3.45 or more	36
Less than 3.95	64	3.95 or more	16
Less than 4.45	74	4.45 or more	6
Less than 4.95	80	4.95 or more	0

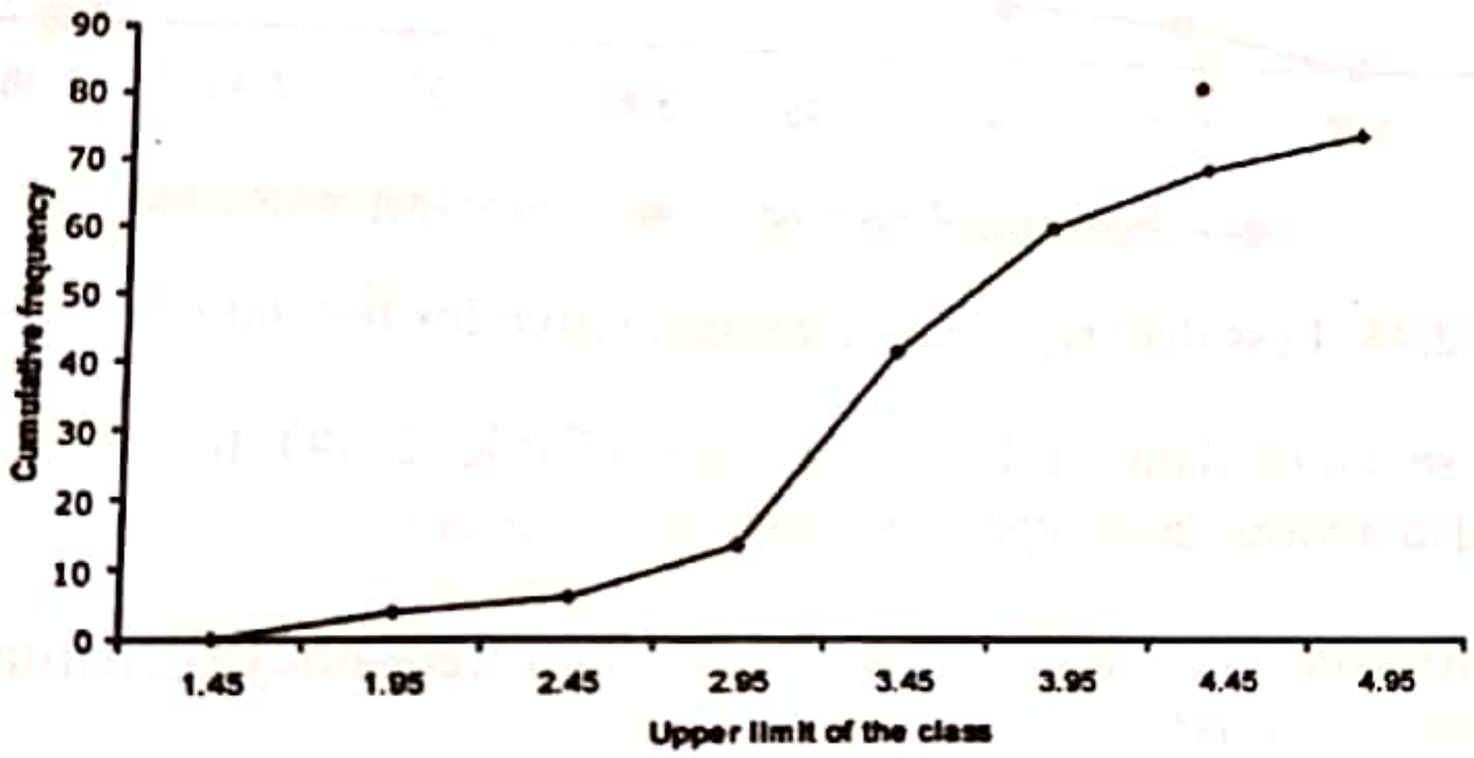
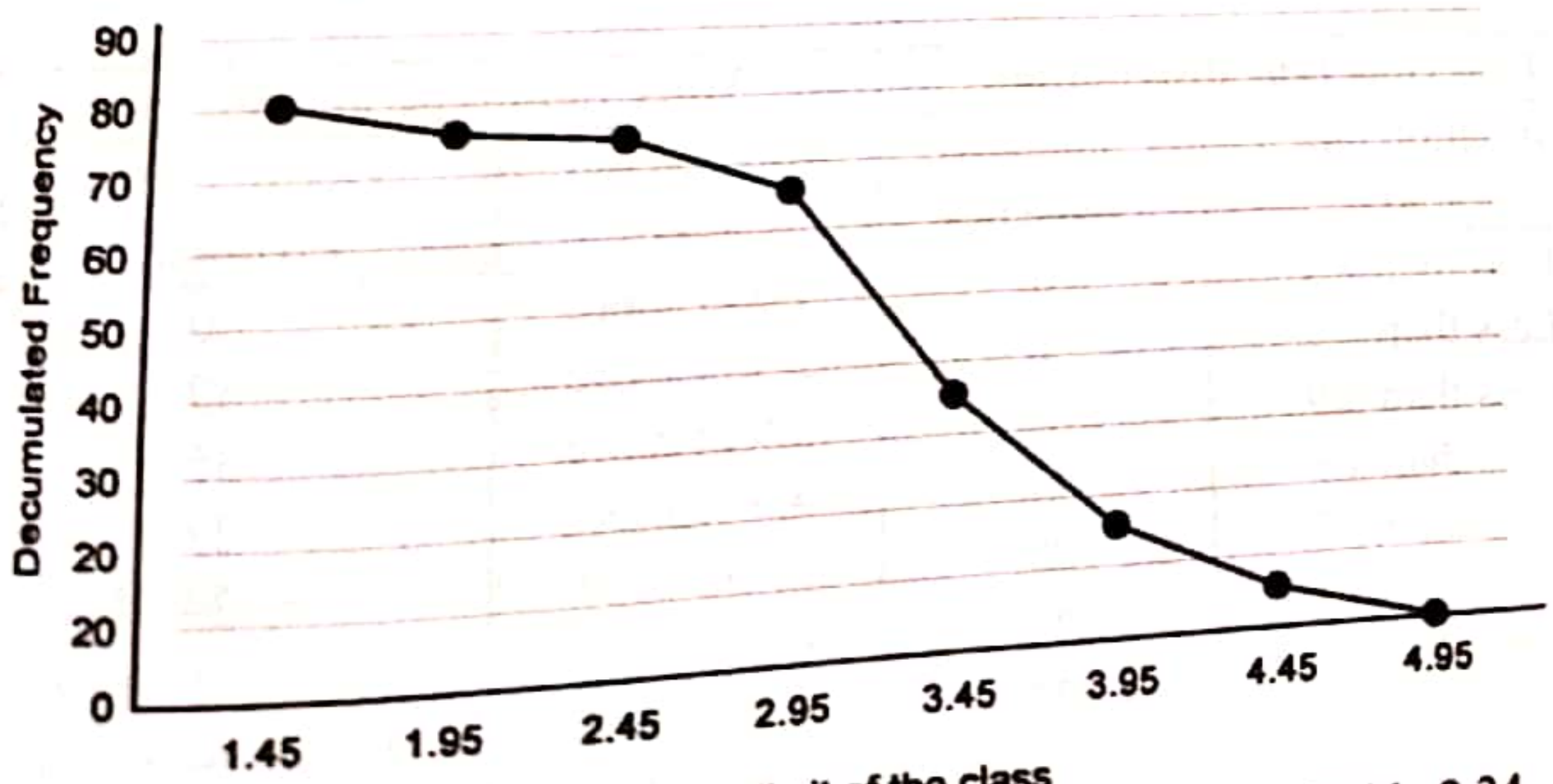


Figure 2.43: Less than type ogive with longevity data in Table 2.34



ogive with longevity data in Table 2.34

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 particularly the median of a distribution. Here is such a graph (Figure 2.45) based on the less than type and more than cumulative frequency shown in Table 2.34.

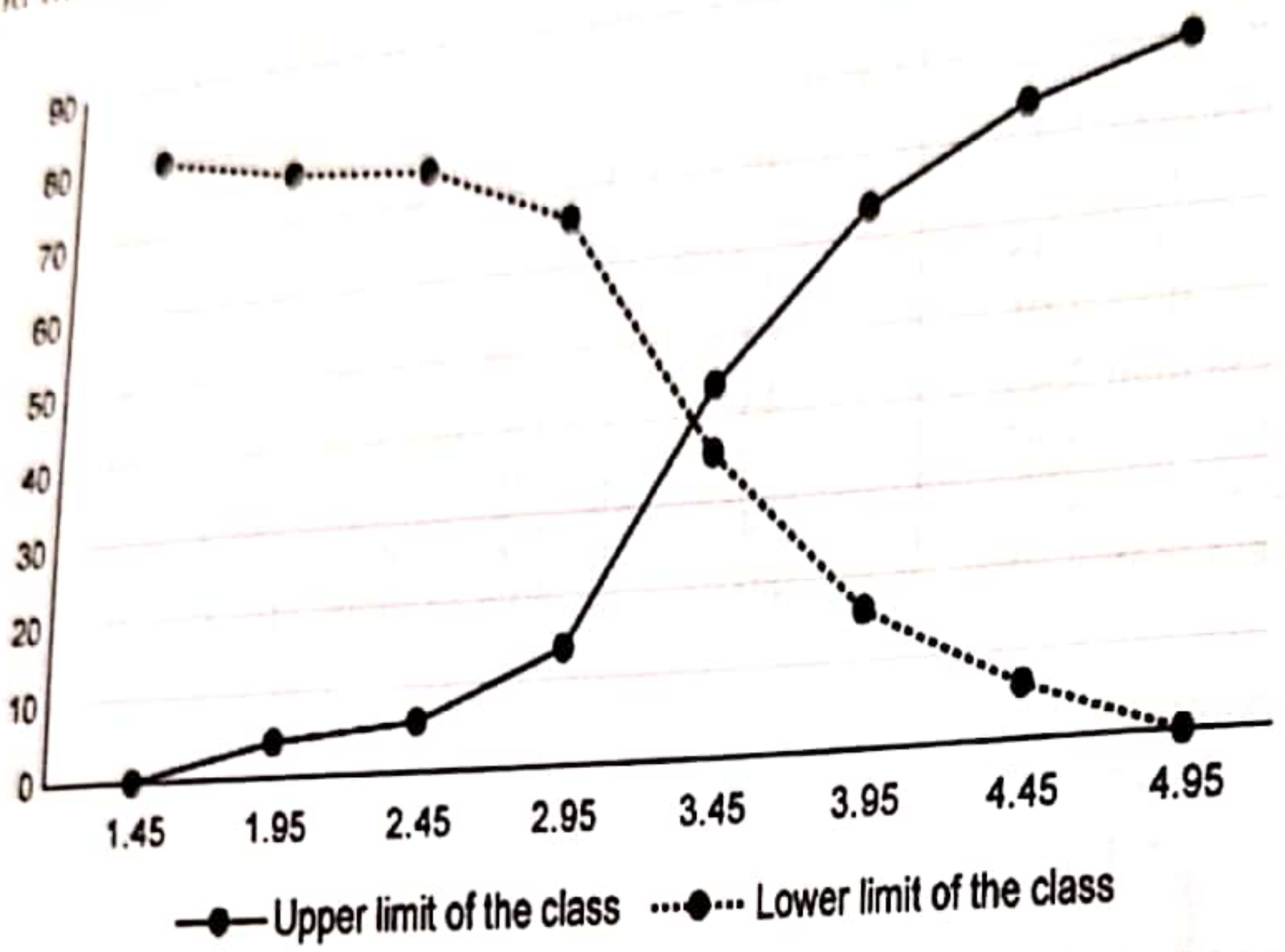


Figure 2.45: Less than type and more than ogive for the data in Table 2.34

Example 2.44: Use GDP data in Example 2.43 (Table 2.34) to draw a less than type (cumulative) and also a more than type (decumulative) ogives.

Solution: Table below forms the less than and more than frequency distributions to illustrate the construction of the required ogives.

Table 2.35: Cumulative Frequency Distributions for Less than and More than Type Ogives Based on the Data on GDP Growth Rate (real)

Less than type distribution		More than type distribution	
Inflation rate (%)	Cumulative Frequency	Inflation rate (%)	Decumulative Frequency
Less than 1.9	0	1.9 or more	40
Less than 2.9	1	2.9 or more	39
Less than 3.9	5	3.9 or more	35
Less than 4.9	17	4.9 or more	23
Less than 5.9	26	5.9 or more	14
Less than 6.9	35	6.9 or more	5
Less than 7.9	38	7.9 or more	2
Less than 8.9	40	8.9 or more	0

The less than type ogive based on the GDP data is displayed in Figure 2.

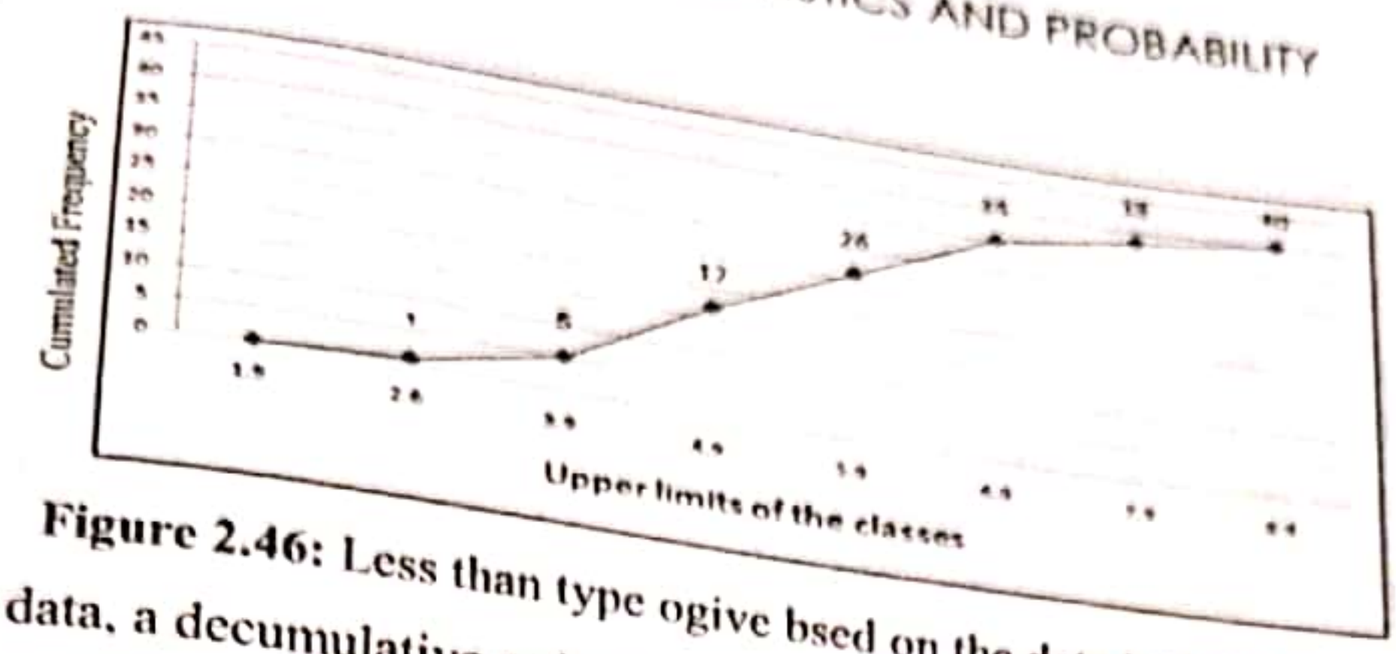


Figure 2.46: Less than type ogive bsed on the data in Table 2.35

With the same data, a decumulative ogive is drawn below in Figure 2.47.

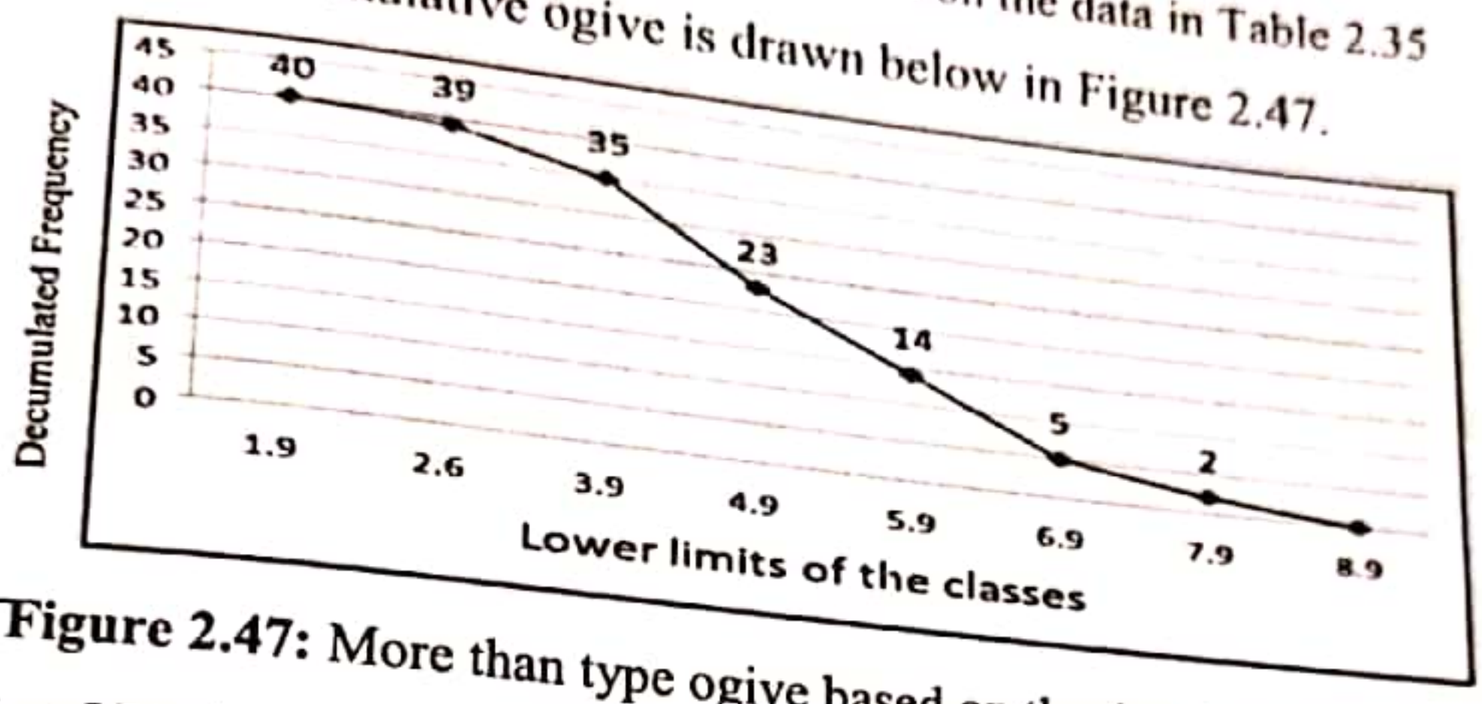


Figure 2.47: More than type ogive based on the data in Table 3.35

(d) Stem and Leaf Plot

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 in 1970 as an aid to understanding and exploring data through statistical analysis, called **Exploratory Data Analysis (EDA)**.

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

The stem and leaf plot is a clever simple graphical 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 of score, 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. We now illustrate the technique by an example.

Example 2.45: The operating profit ratios (OPR) of Janata Bank Limited for 20 consecutive years from 2001 to 2020 were reported to be as follows in the annual report of the bank under reference.

38	84	17	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 OPR as the stem and the *trailing* (units) digits as the leaf. For example, for the ratio 38, the leading digit is 3, and the trailing digit is 4; for 84, the leading digit is 8 and the trailing digit is 4 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 ratio are then written in the appropriate row. In this way each ratio 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). The construction of the plot then can be accomplished in three steps. The steps are identified as (a), (b) and (c) in the table below. Part (a) of the table shows the placing of the stems for the ratios; while part (b) shows how the first ratio '38' is placed. Each new score added to the stem and leaf plot results in a new leaf. The completed stem and leaf plot for the scores is shown in part (c) of the table.

(a) Stem only	(b) Stem with first observation (38)		(c) Complete stem and leaf plot	
1	1		1	7 9
2	2		2	2
3	3	8	3	8 3 9
4	4		4	5 7
5	5		5	3 4 5 4 1 4
6	6		6	6 5
7	7		7	6 5 2
8	8		8	4

Key: 2|2 represents 22, 3|8 represents 38

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 table. The final figure is as shown below:

Stem	Leaf
1	7 9
2	2
3	3 8 9
4	5 7
5	1 3 4 4 4 5
6	5 6
7	2 5 6
8	4

Key: 1|7 represents 17, 7|2 represents 72

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.

To read the score from the above figure, start at the first row and read the ratios 17 and 19. These ratios 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 ratios: 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 ratio (84) and the smallest ratio (17) can be readily located. In addition, an entire picture of how the ratios are distributed (or scattered) emerges. For example, it is readily apparent that there are more ratios in the fifties than any other group; 8 ratios are less than 50, and only 4 ratios are above 70. Additionally, some of the numbers on the stem may have no corresponding leaves. That is in the figure, the stem position "2" would have no corresponding leave if the observation "2" were removed from the data set.

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 the frequencies, concentrations of ratios 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.

Example 2.46: The book values per share in million taka, as shown in the 2018 Annual Report of Premier Bank for 16 consecutive years, were as follows:

6, 8, 12, 14, 14, 15, 15, 16, 18, 19, 19, 23, 23, 24, 26, 26

Display the data by stem and leaf diagram.

Stem		Leaf
5		1 3
10		2 4 4
15		0 0 1 3 4 4
20		3 3 4
25		1 1

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

Example 2.47: The Consumer Price Indexes (CPI) with 2001–02 as base year as reported in Bangladesh Economic Review (2019) were as follows:

161, 163, 163, 166, 168, 168, 169, 169, 170, 170, 171, 173, 173, 174, 175, 177, 179, 180

Display the data by stem and leaf diagram.

Solution: Here is the desired plot:

Stem		Leaf
16		1 3 3
16		6 8 8 9 9
17		0 0 1 3 3 4
17		5 7 9
18		0

Key: 16|8 means168, 17|0 means170

Example 2.48: Once again, we employ the data on longevity of 80 electric bulbs as shown in Example 3.5. Display the data by a stem and leaf plot. For instant reference, we reproduce the data below:

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 having the whole numbers 1, 2, 3 etc, as the stems. These values are placed, as usual, on the left side of the display as in the table below. The second (tenth place) digits are the leaves of the display. These are placed in rows corresponding to the appropriate stems. The stem and leaf display in the unordered and ordered form appears below:

Unordered display:

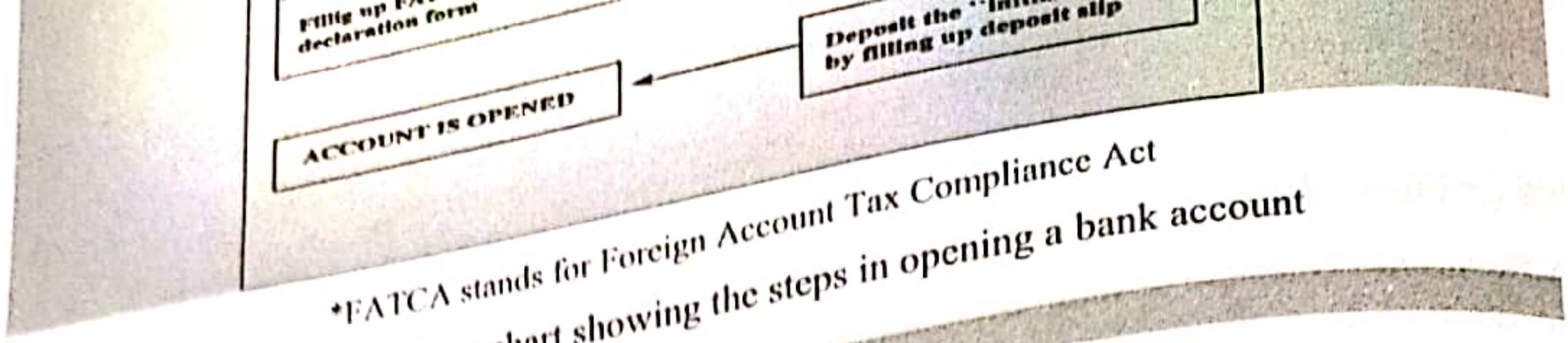
Stem		Leaf
1		5967
2		2659625689
3		527003413817224639114317244238290530942577987682138
4		1573417210433678

Ordered display:

Stem		Leaf
1		5679
2		2255666899
3		000011111122222223333333444444555667777778888899999
4		011123334456778

Key: 1|5 means1.5, 2|2 means 2.2. 4|0 means 4.0

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 the frequency 10. The other two classes are formed in a similar manner and these two classes contain the frequencies 50 and 16.



*FATCA stands for Foreign Account Tax Compliance Act

Figure 2.61: Flow chart showing the steps in opening a bank account

EXERCISES 2

1. Define statistical data. How are they collected? What are the various types of data available for statistical analysis you are familiar with?
2. What is a variable? How does it differ from a constant? How do you distinguish a dependent variable from an independent variable, a discrete variable from a continuous variable, a numeric variable from a categorical variable? What is an attribute? Give example in each case.
3. Compare and contrast various levels of measurement with illustrative examples on business related data.
4. 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.
5. 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.
6. What do you mean by presentation of statistical data? What are the various methods of presenting statistical data?
7. How do you present the statistical data when the data are (i) categorical and (ii) numerical?
8. 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.
9. What is frequency table? Distinguish between a univariate frequency table and bivariate frequency table. What purposes do they serve?
10. 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?
11. 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?

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

13. Define the following in connection with a frequency table:
(a) Class interval (b) Class mark or class mid-point (c) Class limits (d) Class boundaries and (e) Class frequency.

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

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

16. The following figures refer to the employment of the Bangladesh nationals 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

17. Describe how an ogive can be constructed for a frequency distribution. How does a less than type ogive differ from a more than type ogive? Illustrate with an example.

18. The number of Initial Public offerings (IPOs) in Dhaka Stock Exchange for 100 consecutive months since 2005–06 to 2014-15 were as recorded below:

18,10, 13, 17, 23, 19, 15, 15, 13, 16, 11, 9, 11, 10, 19, 10, 14, 18, 23, 19, 15, 15, 13, 20, 11, 9, 12, 10, 15, 13, 16, 11, 9, 11, 10, 1 3, 19, 15, 15, 13, 9, 18, 17, 25, 16, 21, 22, 14, 15, 19, 23, 25, 10, 10, 19, 10, 14, 18, 14, 15, 20, 10, 15, 13, 16, 11, 9, 11, 10, 1, 3, 19, 15, 15, 11, 10, 1, 3, 19, 15, 15, 21, 24, 25, 10, 10, 19, 10, 14, 18, 14, 15, 20, 10, 13, 17, 23, 19, 15, 15, 13, 16, 18

(a) Form an array in ascending order.

(b) Organize the data in 10 classes such as of reasonable widths and complete the frequency distribution.
(c) Using the above data, draw a suitable diagram.
(d) Draw a stem and leaf diagram to represent the data.

19. 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

20 The following table shows the percent distribution of country-wise remittance inflow in FY 2017–18,

Country	Remittance (%)
KSA	17
UAE	16
USA	13
Kuwait	8
UK	7
Malaysia	7
Qatar	6
Oman	6
Other countries	20

Construct a pie chart of this distribution.

21. The data below show the category-wise expatriates in 2009 in percentages.

Category	Percent
Skilled	27.7
Semi-skilled	16.0
Professional	0.3.0
Less skilled	54.0
Others	2.0

Use pie chart and bar diagram to represent the data

22. A study of air pollution in a city yielded the following daily readings of 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
0.08	0.15	0.05	0.09	0.10	0.13	0.07	0.14	0.09	0.17
0.15	0.09	0.10	0.16	0.12	0.27	0.10	0.12	0.05	0.06
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	0.17	0.11	0.18	0.20	0.13	0.17	0.07	0.15	0.15
0.09	0.14	0.14	0.19	0.18	0.17	.016	0.12	0.19	0.18

(a) Group these data into a frequency table having the classes 0.00–0.04, 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 than" percentage distribution and hence an ogive

(c) Construct a histogram of the distribution obtained in (a)

23. The **Bhorer Kagoj** (A daily in Bangladesh) reported the following figures on number of suicides that were committed in Comilla and Jhenaidah 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 cluster bar chart and comment on the trend in suicides of these two districts. What other alternative diagrams could be used to represent the data?

24. 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 and hence draw a histogram of the data.
(b) Set up a relative frequency distribution, 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?

25. 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?

26. 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.

27 The following values represent the numbers of Bangladeshi workers for a period of 10 calendar years in Saudi Arabia: Draw a horizontal bar diagram to illustrate the result of this study.

. Calendar year	Number of expatriates
2010	7069
2011	15039
2012	21232
2013	12654
2014	10657
2015	58270
2016	143913
2017	551308
2018	257317
2019	86219

28. 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+									

29. 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?

30. 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

31 The accompanying table shows the data on trends in inflation rates of Bangladesh at the national level since 1995–1996 derived on the basis of CPI computed with 1985–86 as the base year and the measure of changes in producer prices based on GDP deflators. Compare the data by a runs plot.

Year	Rate of inflation (%) based on CPI	Rate of inflation (%) based on GDP deflator
1995–1996	6.65	4.23
1996–1997	2.52	3.09
1997–1998	6.99	5.28
1998–1999	8.91	4.65
1999–2200	3.41	1.86
2000–2001	1.58	1.59
2001–2002	2.36	3.20
2002–2003	4.50	4.37

Source: National Accounts Statistics, BBS, and July 2004. P: 27

32. The following data were reported in Global Econy.com showing the Inflation rate of Bangladesh based on the CPI for the period 1987-2018.

Year	Rate of inflation (%) based on CPI	Year	Rate of inflation (%) based on CPI
1987	9.9	2003	5.7
1988	7.4	2004	7.6
1989	6.0	2005	7.0
1990	6.1	2006	6.8
1991	6.4	2007	9.1
1992	3.4	2008	8.9
1993	3.0	2009	5.4
1994	5.3	2010	8.1
1995	10.3	2011	10.7
1996	2.4	2012	6.2
1997	5.3	2013	4.5
1998	8.4	2014	7.0
1999	6.1	2015	6.2
2000	2.2	2016	5.5
2001	2.0	2017	5.7
2002	3.3	2018	6.0

of merchandise for the period 1996-97 to 2014-15

in crore taka in Bangladesh. Display the data by a runs plot.

Year	Exports	Imports	Year	Export	Imports
1996–97	16564	30540	2006-07	78931	118478
1997–98	20393	34183	2007-08	87022	148370
1998–99	20851	38480	2008-09	97498	154823
1999–00	24923	42131	2009-10	102148	164241
2000–01	32419	50371	2010-11	144431	240028
2001–02	30934	49049	2011-12	180313	280963
2002–03	33242	55918	2012-13	189437	272328
2003–04	40581	64257	2013-14	213374	316172
2004–05	50853	80898	2014-15	226486	315185
2005-06	62601	99130	2015-16	2634668	3869349

Source: Statistical Yearbook, Bangladesh 2016: 249

34. In many countries, major consulting firms employ statistical analysts to assess the effectiveness of the systems they develop. This case concerns a consulting firm that developed a new computer-based electronic billing system for Navana Company that was experiencing problems with its billing and account receivable process. The Navana's former billing system employed a computer to generate invoices, which were mailed to the customers. However, customers were taking too long time to make payments. The standard payment time for most companies is 30 days, where the payment time is measured from the date on the invoice to the date payment is received. A number of trucking company's customers were not meeting the 30-day standard. In fact, typical payments times were 39 days or even more. To reduce payments, the consulting firm installed a new billing system. Specifically, the consulting firm decides to randomly select 65 of the `invoices produced by the new system during its 12 weeks of operation. The following data refer to these payment times.

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

35. Construct a runs plot and determine if the process appears to be in statistical control.

The following data relate to the current liabilities (in million dollars) derived from the Annual Report of McDonald's Corporation for 1995. Represent the data by a pie chart.

Current liabilities	Amount
Notes payable	413.0
Accounts payable	564.3
Income and other taxes	182.5
Accrued interest	117.4
Other accrued liabilities	352.5
Current maturities of long term debt	
Total	

36. The gross revenue expenditures of the government of Bangladesh in million BDT for the financial year 2001–2002 (budget estimate) were as follows:

Expenditure heads	Expenditure in million taka
Wages and salaries	66782
Commodities and services	42872
Transfer	101219
Other services	9506
Total	**220,379**

Source: Statistical Yearbook of Bangladesh, p. 371

Display the data by a pie and a bar diagram.

37 Harley Manufacturing Company is in the process of preparing its statement of cash flows for the year ending in December 31, 2004. As a part of that, the company prepared a statement of the operating expenses, which appears below:

Expenditure heads	Amount (in US $)
Materials and supplies	250,000
Direct labor	400,000
Manufacturing overhead	181, 500
Depreciation	93,5000
Selling expenses	245,000
Interest expenses	7,500
Income tax expenses	6,500
Total	**1,184,000**

Represent the data by a frequency bar diagram.

38. The following data display the percent distribution of Bangladesh non-development and development budget in 6 different heads for the year 2017–2018. Present the data by a pie chart.

Heads	Percent
Foreign grants	1.4
Foreign loan	11.6
Domestic loan	15.1
Non-tax revenue	7.8
Tax revenue (Non-NBR)	2.1
Tax revenue (NBR)	62.1
Total	**100.0**

Source: Internet

(Total Budget: 4002.66 Billion)