Chapter

7

Data Preparation

The data, after collection, has to be prepared for analysis. The collected data is raw and it must be converted to the form that is suitable for the required analysis. The results of the analysis are affected a lot by the form of the data. So, proper data preparation is a must to get reliable results.

7.1 DATA PREPARATION PROCESS

The plan of data analysis is decided in advance before collecting the data. Data preparation process is guided by that plan of data analysis. Important steps of data preparation process are as follows:

(i) Questionnaire checking
(ii) Editing
(iii) Coding
(iv) Classification
(v) Tabulations
(vi) Graphical representation
(vii) Data cleaning
(viii) Data adjusting

We below describe each of the above processes.

7.1.1 Questionnaire Checking

When the data is collected through questionnaires, the first step of data preparation process is to check the questionnaires if they are acceptable or not. This involves the examination of all questionnaires for their completeness and interviewing quality. Usually this step is undertaken at the time of data collection. If questionnaires checking was not done at the time of collection, it should be done later. A questionnaire may not be acceptable if :

(i) It is incomplete partially or fully.
(ii) It is answered by a person who has inadequate knowledge or does not qualify for the participation

(iii) It is answered in such a way which gives the impression that the respondent could not understand the questions.

If sufficient number of questionnaires are not accepted the researcher may like to collect more data.

7.1.2 Editing

Editing of data is a process of examining the collected raw data (specially in surveys) to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and/or schedules. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as completed as possible and have been well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and central editing. *Field editing* consists in the review of the reporting forms by the investigator for completing (translating or rewriting) what the latter has written in abbreviated and/or in illegible form at the time of recording the respondents' responses. This type of editing is necessary in view of the fact that individual writing styles often can be difficult for others to decipher. This sort of editing should be done as soon as possible after the interview, preferably on the very day or on the next day. While doing field editing, the investigator must restrain himself and must not correct errors of omission by simply guessing what the informant would have said if the question had been asked.

Central editing should take place when all forms or schedules have been completed and returned to the office. This type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a large inquiry. Editor(s) may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule. At times, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final results, especially in the context of mail surveys.

Editors must keep in view several points while performing their work: (a) They should be familiar with instructions given to the interviewers and coders as well as with the editing instructions supplied to them for the purpose. (b) While crossing out an original entry for one reason or another, they should just draw a single line on it so that the same may remain legible. (c) They must make entries (if any) on the form in some distinctive colur and that too in a standardised form. (d) They should initial all answers which they change or supply. (e) Editor's initials and the date of editing should be placed on each completed form or schedule.

7.1.3 Coding

Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusively which means that

a specific answer can be placed in one and only one cell in a given category set. Another rule to be observed is that of unidimensionality by which is meant that every class is defined in terms of only one concept.

Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis. Coding decisions should usually be taken at the designing stage of the questionnaire. This makes it possible to precode the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch from the original questionnaires. But in case of hand coding some standard method may be used. One such standard method is to code in the margin with a coloured pencil. The other method can be to transcribe the data from the questionnaire to a coding sheet. Whatever method is adopted, one should see that coding errors are altogether eliminated or reduced to the minimum level.

7.1.4 Classification

Most research studies result in a large volume of raw data which must be reduced into homogeneous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes. Classification can be one of the following two types, depending upon the nature of the phenomenon involved:

(a) ***Classification according to attributes:*** As stated above, data are classified on the basis of common characteristics which can either be descriptive (such as literacy, sex, honesty, etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively; only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are known as *statistics of attributes* and their classification is said to be classification according to attributes.

Such classification can be simple classification or manifold classification. In simple classification we consider only one attribute and divide the universe into two classes—one class consisting of items possessing the given attribute and the other class consisting of items which do not possess the given attribute. But in manifold classification we consider two or more attributes simultaneously, and divide that data into a number of classes (total number of classes of final order is given by 2^n, where n = number of attributes considered)*. Whenever data are classified according to attributes, the researcher must see that the attributes are defined in such a manner that there is least possibility of any doubt/ambiguity concerning the said attributes.

(b) ***Classification according to class-intervals:*** Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to income, production, age, weight, etc. come under this

* Classes of the final order are those classes developed on the basis of 'n' attributes considered. For example, if attributes A and B are studied and their presence is denoted by A and B respectively and absence by a and b respectively, then we have four classes of final order viz., class AB, class Ab, class aB, and class ab.

category. Such data are known as *statistics of variables* and are classified on the basis of class intervals. For instance, persons whose incomes, say, are within Rs 201 to Rs 400 can form one group, those whose incomes are within Rs 401 to Rs 600 can form another group and so on. In this way the entire data may be divided into a number of groups or classes or what are usually called, 'class-intervals.' Each group of class-interval, thus, has an upper limit as well as a lower limit which are known as class limits. The difference between the two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution. Classification according to class intervals usually involves the following three main problems:

(i) How may classes should be there? What should be their magnitudes?

There can be no specific answer with regard to the number of classes. The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class-intervals should be of equal magnitudes, but in some cases unequal magnitudes may result in better classification. Hence the researcher's objective judgement plays an important part in this connection. Multiples of 2, 5 and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by H.A. Sturges, determining the size of class interval:

$$i = R/(1 + 3.3 \log N)$$

where

i = size of class interval;

R = Range (i.e., difference between the values of the largest item and smallest item among the given items);

N = Number of items to be grouped.

It should also be kept in mind that in case one or two or very few items have very high or very low values, one may use what are known as open-ended intervals in the overall frequency distribution. Such intervals may be expressed like under Rs. 500 or Rs. 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researcher must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are to be classified.

(ii) How to choose class limits?

While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by 2) of a class-interval and the actual average of items of that class interval should remain as close to each other as possible.

Consistent with this, the class limits should be located at multiples of 2, 5, 10, 20, 100 and such other figures. Class limits may generally be stated in any of the following forms:

Exclusive type class intervals: They are usually stated as follows:

10–20
20–30
30–40
40–50

The above intervals should be read as under:

10 and under 20
20 and under 30
30 and under 40
40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to the upper limit of a class are grouped in the next higher class. For example, an item whose value is exactly 30 would be put in 30–40 class interval and not in 20–30 class interval. In simple words, we can say that under exclusive type class intervals, the upper limit of a class interval is excluded and items with values less than the upper limit (but not less than the lower limit) are put in the given class interval.

Inclusive type class intervals: They are usually stated as follows:

11–20
21–30
31–40
41–50

In inclusive type class intervals the upper limit of a class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11–20 class interval. The stated upper limit of the class interval 11–20 is 20 but the real limit is 20.99999 and as such 11–20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be a discrete one (i.e., can be measured and stated only in integers), then we should adopt inclusive type classification. But when the phenomenon happens to be a continuous one capable of being measured in fractions as well, we can use exclusive type class intervals.*

(iii) How to determine the frequency of each class?

This can be done either by tally sheets or by mechanical aids. Under the technique of tally sheet, the class-groups are written on a sheet of paper (commonly known as the tally sheet) and for each item a stroke (usually a small vertical line) is marked against the class group in which it falls. The general practice is that after every four small vertical lines in a class group, the fifth line for the item falling in the same group, is

* The stated limits of class intervals are different than true limits. We should use true or real limits keeping in view the nature of the given phenomenon.

indicated as horizontal line through the said four lines and the resulting flower (𝍸) represents five items. All this facilitates the counting of items in each one of the class groups. An illustrative tally sheet can be shown as under:

Table 7.1: An Illustrative Tally Sheet for Determining the Number of 70 Families in Different Income Groups

Income groups (Rupees)	*Tally mark*	*Number of families or (Class frequency)*
Below 400	𝍸 𝍸 III	13
401–800	𝍸 𝍸 𝍸 𝍸	20
801–1200	𝍸 𝍸 II	12
1201–1600	𝍸 𝍸 𝍸 III	18
1601 and above	𝍸 II	7
Total		70

Alternatively, class frequencies can be determined, specially in case of large inquires and surveys, by mechanical aids i.e., with the help of machines viz., sorting machines that are available for the purpose. Some machines are hand operated, whereas other work with electricity. There are machines which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

7.1.5 Tabulation

When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarising raw data and displaying the same in compact form (i.e., in the form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows.

Tabulation is essential because of the following reasons.

1. It conserves space and reduces explanatory and descriptive statement to a minimum.
2. It facilitates the process of comparison.
3. It facilitates the summation of items and the detection of errors and omissions.
4. It provides a basis for various statistical computations.

Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulating machines or computers. In relatively large inquiries, we may use mechanical or computer tabulation if other factors are favourable and necessary facilities are available. Hand tabulation is usually preferred in case of small inquiries where the number of questionnaires is small and they are of relatively short

length. Hand tabulation may be done using the direct tally, the list and tally or the card sort and count methods. When there are simple codes, it is feasible to tally directly from the questionnaire. Under this method, the codes are written on a sheet of paper, called tally sheet, and for each response a stroke is marked against the code in which it falls. Usually after every four strokes against a particular code, the fifth response is indicated by drawing a diagonal or horizontal line through the strokes. These groups of five are easy to count and the data are sorted against each code conveniently. In the listing method, the code responses may be transcribed onto a large work-sheet, allowing a line for each questionnaire. This way a large number of questionnaires can be listed on one work sheet. Tallies are then made for each question. The card sorting method is the most flexible hand tabulation. In this method the data are recorded on special cards of convenient size and shape with a series of holes. Each hole stands for a code and when cards are stacked, a needle passes through particular hole representing a particular code. These cards are then separated and counted. In this way frequencies of various codes can be found out by the repetition of this technique. We can as well use the mechanical devices or the computer facility for tabulation purpose in case we want quick results, our budget permits their use and we have a large volume of straight forward tabulation involving a number of cross-breaks.

Tabulation may also be classified as simple and complex tabulation. The former type of tabulation gives information about one or more groups of independent questions, whereas the latter type of tabulation shows the division of data in two or more categories and as such is deigned to give information concerning one or more sets of inter-related questions. Simple tabulation generally results in one-way tables which supply answers to questions about one characteristic of data only. As against this, complex tabulation usually results in two-way tables (which give information about two inter-related characteristics of data), three-way tables (giving information about three interrelated characteristics of data) or still higher order tables, also known as manifold tables, which supply information about several interrelated characteristics of data. Two-way tables, three-way tables or manifold tables are all examples of what is sometimes described as cross tabulation.

Generally accepted principles of tabulation: Such principles of tabulation, particularly of constructing statistical tables, can be briefly states as follows:*

1. Every table should have a clear, concise and adequate title so as to make the table intelligible without reference to the text and this title should always be placed just above the body of the table.
2. Every table should be given a distinct number to facilitate easy reference.
3. The column headings (captions) and the row headings (stubs) of the table should be clear and brief.
4. The units of measurement under each heading or sub-heading must always be indicated.
5. Explanatory footnotes, if any, concerning the table should be placed directly beneath the table, along with the reference symbols used in the table.
6. Source or sources from where the data in the table have been obtained must be indicated just below the table.
7. Usually the columns are separated from one another by lines which make the table more readable and attractive. Lines are always drawn at the top and bottom of the table and below the captions.

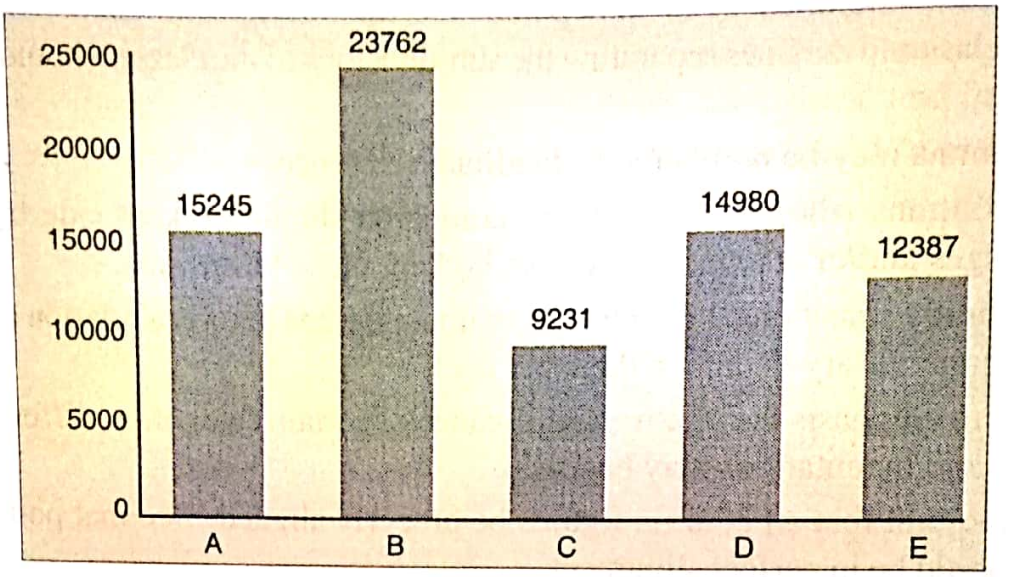
* All these points constitute the characteristics of a good table.

8. There should be thick lines to separate the data under one class from the data under another class and the lines separating the sub-divisions of the classes should be comparatively thin lines.
9. The columns may be numbered to facilitate reference.
10. Those columns whose data are to be compared should be kept side by side. Similarly, percentages and/or averages must also be kept close to the data.
11. It is generally considered better to approximate figures before tabulation as the same would reduce unnecessary details in the table itself.
12. In order to emphasis the relative significance of certain categories, different kinds of type, spacing and indentations may be used.
13. It is important that all column figures be properly aligned. Decimal points and (+) or (–) signs should be in perfect alignment.
14. Abbreviations should be avoided to the extent possible and ditto marks should not be used in the table.
15. Miscellaneous and exceptional items, if any, should be usually placed in the last row of the table.
16. Table should be made as logical, clear, accurate and simple as possible. If the data happen to be very large, they should not be crowded in a single table for that would make the table unwieldy and inconvenient.
17. Total of rows should normally be placed in the extreme right column and that of columns should be placed at the bottom.
18. The arrangement of the categories in a table may be chronological, geographical, alphabetical or according to magnitude to facilitate comparison. Above all, the table must suit the needs and requirements of an investigation.

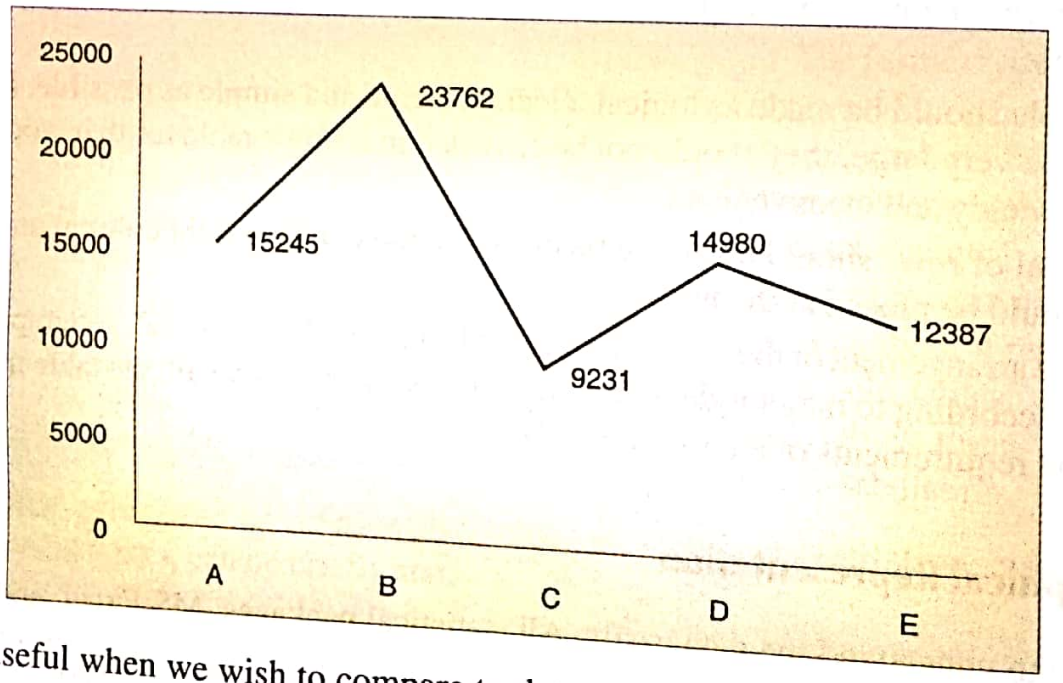
7.1.6 Graphical Representation

Graphs help to understand the data easily. All statistical packages, MS Excel, and OpenOffice.org offer a wide range of graphs. In case of qualitative data (or categorized data), most common graphs are bar charts and pie charts.

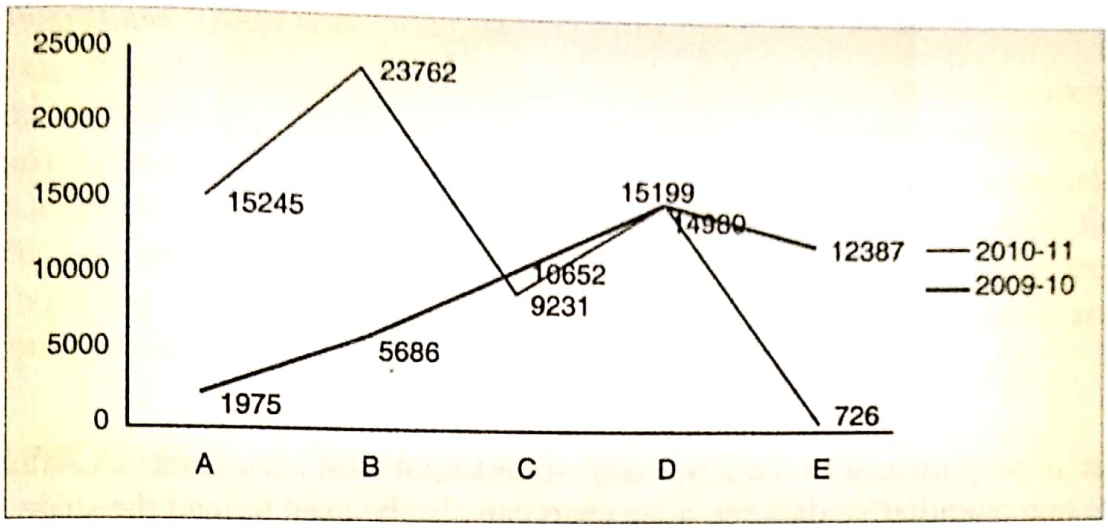
Bar chart: A bar chart consists of a series of rectangles (or bars). The height of each rectangle is determined by the frequency of that category. Suppose that the sales of a popular soft drink in the year 2010-11, in five geographical regions, denoted as A, B, C, D, and E, are 15245, 23762, 9231, 14980, and 12387, respectively, measured in 10,000 USD. A bar chart of this data is as below.



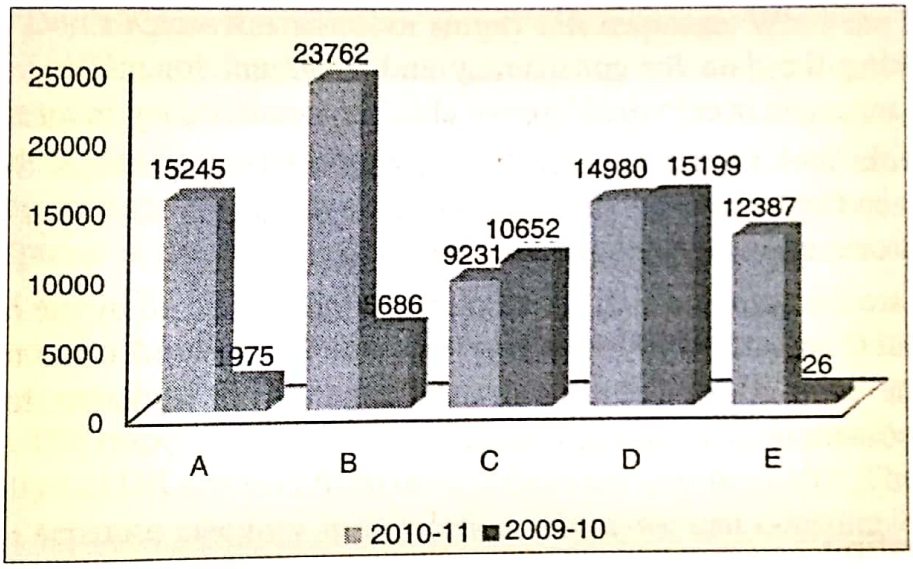
A line chart can also be plotted in this data by connecting the midpoints of each rectangle as below:



Line charts are useful when we wish to compare to data sets as we can overlap to line charts. For example, the sales data of the same soft drink in the same geographical regions in the year 2009-10 were 1975, 5686, 10652, 15199 and 726, respectively, measured in 10,000 USD. The line chart showing the data for both the years is

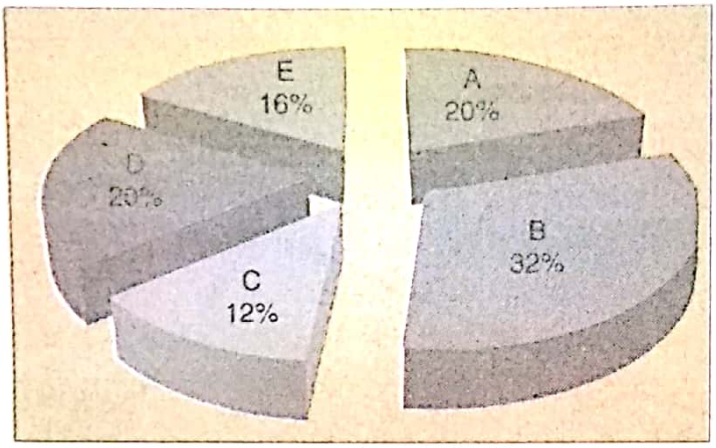


Another graphical representation of the same data is using multiple bars as below:



Pie chart: A pie chart is used to emphasize relative proportion or shares of each category. It's a circular chart divided into sectors, illustrating relative frequencies. The relative frequency in each category or sector is proportional to the arc length of that sector or the area of that sector or the central angle of that sector. Suppose in the previous example, if the soft drink has its market only in the five geographical regions, denoted as A, B, C, D, and E. For the year 2010-11, the sales data in these regions are 15245, 23762, 9231, 14980, and 12387, respectively, measured in 10,000 USD. A total sale of the soft drink is 75605 times 10,000 USD. A pie chart can be plotted to have the idea of the shares of different markets.

In case of quantitative data, one important chart is histogram which is a generalization of bar chart. The data is first summarized in terms of class intervals and each bar represents a class interval. The width of the bar is proportional to the width of corresponding class interval. The



area of the bar is proportional to the frequency of corresponding class interval. After making the class intervals in a quantitative data set, a pie chart can also be used to read the share of each class interval.

7.1.7 Data Cleaning

This includes checking the data for consistency and treatment for missing value. Preliminary consistency checks are made in editing. Here we check the consistency in an extensive manner.

Consistency checks look for the data which are not consistent or outlines. Such data may either be discarded or replaced by the mean value. However, the researcher should be careful while doing this. Extreme values or outlines are not always erroneous.

Missing values are the values which are unknown or not answered by the respondent. In place of such missing values, some neutral value may be used. This neutral value may be the mean of available values. The other option could be to use the pattern of responses to other questions to calculate a suitable substitute to the missing values.

7.1.8 Data Adjusting

Data adjusting is not always necessary but it may improve the quality of analysis sometimes. This consists of following methods.

(i) **Weight-assigning:** Each respondent or case is assigned a weight to reflect its importance relative to other respondents or cases.

Using this method, the collected sample can be made a stronger representative of a target population on specific characteristics. For example, the cases of educated people could be assigned higher weights and of uneducated people could be assigned lower weights in some survey. The value 1.0 means unweightage case.

(ii) **Variable Respecification:** This involves creating new variables or modifying existing variables. For example, if the usefulness of a certain product is measured on 10 point scale, it may be reduced on a 4 point scale — 'very useful', 'useful', 'neutral', 'not useful'. Ratio of two variables may also be taken to create a new variable.

Method of dummy variables for respecifying categorical variables is also very popular. Dummy variable is a variable which usually takes numerical values based on the corresponding category in the original variable. For example, a group of people is divided

into smokers and non-smokers. We can define a dummy variable taking the value '1' for smokers and '0' for non-smokers.

(iii) **Scale Transformation:** Scale transformation is done to ensure the comparability with other scales or to make the data suitable for analysis. Different type of characteristics are measured on different scales. For example, attitude variables are measured on continuous scale, life style variables are usually measured on a 5 point Likert scale. So the variables which are measured on different scales, cannot be compared. A common transformation is subtracting all the values of a characteristic by corresponding mean and dividing by corresponding standard deviation.

7.2 SOME PROBLEMS IN PREPARATION PROCESS

We can take up the following two problems of processing the data for analytical purposes:

(a) *The problem concerning "Don't know" (or DK) responses:* While processing the data, the researcher often comes across some responses that are difficult to handle. One category of such responses may be 'Don't Know Response' or simply DK response. When the DK response group is small, it is of little significance. But when it is relatively big, it becomes a matter of major concern in which case the question arises: Is the question which elicited DK response useless? The answer depends on two points viz., the respondent actually may not know the answer or the researcher may fail in obtaining the appropriate information.. In the first case the concerned question is said to be alright and DK response is taken as legitimate DK response. But in the second case, DK response is more likely to be a failure of the questioning process.

How DK responses are to be dealt with by researchers? The best way is to design better type of questions. Good rapport of interviewers with respondents will result in minimising DK responses. But what about the DK responses that have already taken place? One way to tackle this issue is to estimate the allocation of DK answers from other data in the questionnaire. The other way is to keep DK responses as a separate category in tabulation where we can consider it as a separate reply category if DK responses happen to be legitimate, otherwise we should let the reader make his own decision. Yet another way is to assume that DK responses occur more or less randomly and as such we may distribute them among the other answers in the ratio in which the latter have occurred. Similar results will be achieved if all DK replies are excluded from tabulation and that too without inflating the actual number of other responses.

(b) *Use of percentages:* Percentages are often used in data presentation for they simplify numbers, reducing all of them to a 0 to 100 range. Through the use of percentages, the data are reduced in the standard form with base equal to 100 which fact facilitates relative comparisons. While using percentages, the following rules should be kept in view by researchers:

1. Two or more percentages must not be averaged unless each is weighted by the group size from which it has been derived.
2. Use of too large percentages should be avoided, since a large percentage is difficult to understand and tends to confuse, defeating the very purpose for which percentages are used.
3. Percentages hide the base from which they have been computed. If this is not kept in view, the real differences may not be correctly read.

4. Percentage decreases can never exceed 100 per cent and as such for calculating the percentage of decrease, the higher figure should invariably be taken as the base.
5. Percentages should generally be worked out in the direction of the causal-factor in case of two-dimension tables and for this purpose we must select the more significant factor out of the two given factors as the causal factor.

7.3 MISSING VALUES AND OUTLIERS

Missing values are the observations which the researcher plan to collect but could not collect or lost due to some reason. Many statistical tools cannot be employed when the data set has one or more missing values. In data collection through asking questions 'Don't know' response may also creep the problem of missing values. Utmost care should be taken by the researcher to avoid the missing values in the data set. Most common methods to deal with the problem of missing value while conducting the analysis is either to leave the observation, if possible, or to replace the missing value by the arithmetic mean of other collected observation.

Outliers are the observations which are quite different to other observations in the data set. Although all the statistical techniques can be employed when data set has outliers, their interpretations may be misleading. The most common reason of outliers being present in the data set is the recording error. This error should be corrected while editing and cleaning the data. Consider an example of survey of 100 customers in a mall.

If few bulk customers purchasing very large amounts are among the 100 surveyed customers. In this survey having outliers (bulk customers) may not be posing any error as bulk customers are always there in the mall along with small customers. However, in a similar survey at a nearby grocerry shop on a day when there is strike in the mall may include some bulk customers which could be misleading. Thus outliers should not be ignored as they might have some relevant information or pose to a serious risk.

Before detecting the outliers, we need to define them first. Commonly, an observation with a value that is more than 3 standard deviations from the mean is considered as an outlier. A scatter plot (discussed later) can also be helpful in identifying the outliers. After identifying an outlier, the researcher has to decide what to do with it. The researcher may like to delete it or modify the value of it or retain it as it is. It depends on the knowledge about the cause of that outlier.

7.4 TYPES OF ANALYSIS

As stated earlier, by analysis we mean the computation of certain indices or measures along with searching for patterns of relationship that exist among the data groups. Analysis, particularly in case of survey or experimental data, involves estimating the values of unknown parameters of the population and testing of hypotheses for drawing inferences. Analysis may, therefore, be categorised as descriptive analysis and inferential analysis (Inferential analysis is often known as statistical analysis). Descriptive analysis is largely the study of the distributions of one or more variables involed in the study. In this context we work out various measures that show the size and shape of a distribution(s) along with the study of measuring relationships between two or more variables.

We may as well talk of correlation analysis and causal analysis. *Correlation analysis* studies the joint variation of two or more variables for determining the amount of correlation between two or more variables. *Causal analysis* is concerned with the study of how one or more variables affect

changes in another variable. It is thus a study of functional relationships existing between two or more variables. This analysis can be termed as regression analysis. Causal analysis is considered relatively more important in experimental researches, whereas in most social and business researches our interest lies in understanding and controlling relationships between variables then with determining causes *per se* and as such we consider correlation analysis as relatively more important.

In modern times, with the availability of computer facilities, there has been a rapid development of *multivariate analysis*. Usually the following analyses are involved when we make a reference of multivariate analysis:

(a) *Multiple regression analysis:* This analysis is adopted when the researcher has one dependent variable which is presumed to be a function of two or more independent variables. The objective of this analysis is to make a prediction about the dependent variable based on its covariance with all the concerned independent variables.

(b) *Multiple discriminant analysis:* This analysis is appropriate when the researcher has a single dependent variable that cannot be measured, but can be classified into two or more groups on the basis of some attribute. The object of this analysis happens to be to predict an entity's possibility of belonging to a particular group based on several predictor variables.

(c) *Multivariate analysis of variance (or multi-ANOVA):* This analysis is an extension of two-way ANOVA, wherein the ratio of among group variance to within group variance is worked out on a set of variables.

(d) *Canonical analysis:* This analysis can be used in case of both measurable and non-measurable variables for the purpose of simultaneously predicting a set of dependent variables from their joint covariance with a set of independent variables.

Inferential analysis is concerned with the various tests of significance for testing hypotheses in order to determine with what validity data can be said to indicate some conclusion or conclusions. It is also concerned with the estimation of population values. It is mainly on the basis of inferential analysis that the task of interpretation (i.e., the task of drawing inferences and conclusions) is performed.

7.5 STATISTICS IN RESEARCH

The role of statistics in research is to function as a tool in designing research, analysing its data and drawing conclusions therefrom. Most research studies result in a large volume of raw data which must be suitably reduced so that the same can be read easily and for further analysis. Clearly the science of statistics cannot be ignored by any research worker, even though he may not have occasion to use statistical methods in all their details and ramifications. Classification and tabulation, as stated earlier, achieve this objective to some extent, but we have to go a step further and develop certain indices or measures to summarise the collected/classified data. Only after this we can adopt the process of generalisation from small groups (i.e., samples) to population. If fact, there are two major areas of statistics viz., descriptive statistics and inferential statistics. *Descriptive statistics* concern the development of certain indices from the raw data, whereas inferential statistics concern with the process of generalisation. *Inferential statistics* are also known as sampling statistics and are mainly concerned with two major type of problems: (i) the estimation of population parameters, and (ii) the testing of statistical hypotheses.

The important statistical measures that are used to summarise the survey/research data are: (1) measures of central tendency or statistical averages; (2) measures of dispersion; (3) measures of asymmetry (skewness); (4) measures of relationship; and (5) other measures.

Amongst the measures of central tendency, the three most important ones are the arithmetic average or mean, median and mode. Geometric mean and harmonic mean are also sometimes used.

From among the measures of dispersion, variance, and its square root—the standard deviation are the most often used measures. Other measures such as mean deviation, range, etc. are also used. For comparison purpose, we use mostly the coefficient of standard deviation or the coefficient of variation.

In respect of the measures of skewness and kurtosis, we mostly use the first measure of skewness based on mean and mode or on mean and median. Other measures of skewness, based on quartiles or on the methods of moments, are also used sometimes. Kurtosis is also used to measure the peakedness of the curve of the frequency distribution.

Amongst the measures of relationship, Karl Pearson's coefficient of correlation is the frequently used measure in case of statistics of variables, whereas Yule's coefficient of association is used in case of statistics of attributes. Multiple correlation coefficient, partial correlation coefficient, regression analysis, etc., are other important measures often used by a researcher.

Index numbers, analysis of time series, coefficient of contingency, etc., are other measures that may as well be used by a researcher, depending upon the nature of the problem under study.

We give below a brief outline of some important measures (our of the above listed measures) often used in the context of research studies.

PROBLEMS

1. "Processing of data implies editing, coding, classification and tabulation". Describe in brief these four operations pointing out the significance of each in context of research study.
2. Classification according to class intervals involves three main problems viz., h\How many classes should be there? How to choose class limits? How to determine class frequency? State how these problems should be tackled by a researcher.
3. Why tabulation is considered essential in a research study? Narrate the characteristics of a good table.
4. (a) How the problem of DK responses should be dealt with by a researcher? Explain.
   (b) What points one should observe while using percentages in research studies?
5. Write short notes on the followings:
   (i) Data adjusting
   (ii) Data cleaning
   (iii) Questionnaire checking
6. Why scale transformations are made? Explain.
7. How will you treat the missing data?
8. What are the reasons of weight assigning?
9. What are dummy variables? Give example.