7

Processing and Analysis of Data

The data, after collection, has to be processed and analysed in accordance with the outline laid down for the purpose at the time of developing the research plan. This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis. Technically speaking, processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. The term analysis refers to the computation of certain measures along with searching for patterns of relationship that exist among data-groups. Thus, "in the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions".[1] But there are persons (Selltiz, Jahoda and others) who do not like to make difference between processing and analysis. They opine that analysis of data in a general way involves a number of closely related operations which are performed with the purpose of summarising the collected data and organising these in such a manner that they answer the research question(s). We, however, shall prefer to observe the difference between the two terms as stated here in order to understand their implications more clearly.

PROCESSING OPERATIONS

With this brief introduction concerning the concepts of processing and analysis, we can now proceed with the explanation of all the processing operations.

1. 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

[1] G.B. Giles, *Marketing*, p. 44.

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.

2. 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.

3. 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 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.

*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.

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 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

* 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.

which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

4. 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 availaibility 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.
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 emphasise 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.

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

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.

SOME PROBLEMS IN PROCESSING

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 or 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.

ELEMENTS/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 distributions of one variable. This study provides us with profiles of companies, work groups, persons and other subjects on any of a multiple of characteristics such as size. Composition, efficiency, preferences, etc."[2]. this sort of analysis may be in respect of one variable (described as unidimensional analysis), or in respect of two variables (described as bivariate analysis) or in respect of more than two variables (described as multivariate analysis). 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* which may be defined as "all statistical methods which simultaneously analyse more than two variables on a sample of observations"[3]. 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.

[2] C. William Emory, *Business Research Methods*, p. 356.

[3] Jagdish N. Sheth, "The Multivariate Revolution in Marketing Research", *Journal of Marketing*, Vol. 35, No. 1 (Jan. 1971), pp. 13–19.

* Readers are referred to standard texts for more details about these analyses.

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.

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 can be used 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.

* One may read any standard text book on statistical methods for details about these measures.

MEASURES OF CENTRAL TENDENCY

Measures of central tendency (or statistical averages) tell us the point about which items have a tendency to cluster. Such a measure is considered as the most representative figure for the entire mass of data. Measure of central tendency is also known as statistical average. Mean, median and mode are the most popular averages. *Mean*, also known as arithmetic average, is the most common measure of central tendency and may be defined as the value which we get by dividing the total of the values of various given items in a series by the total number of items. we can work it out as under:

$$\text{Mean (or } \bar{X})^{*} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

where $\bar{X}$ = The symbol we use for mean (pronounced as X bar)

$\sum$ = Symbol for summation

X_i = Value of the ith item X, $i = 1, 2, \ldots, n$

n = total number of items

In case of a frequency distribution, we can work out mean in this way:

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_n X_n}{f_1 + f_2 + \ldots + f_n = n}$$

Sometimes, instead of calculating the simple mean, as stated above, we may workout the weighted mean for a realistic average. The weighted mean can be worked out as follows:

$$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$

where $\bar{X}_w$ = Weighted item

w_i = weight of ith item X

X_i = value of the ith item X

Mean is the simplest measurement of central tendency and is a widely used measure. Its chief use consists in summarising the essential features of a series and in enabling data to be compared. It is amenable to algebraic treatment and is used in further statistical calculations. It is a relatively stable measure of central tendency. But it suffers from some limitations viz., it is unduly affected by extreme items; it may not coincide with the actual value of an item in a series, and it may lead to wrong impressions, particularly when the item values are not given with the average. However, mean is better than other averages, specially in economic and social studies where direct quantitative measurements are possible.

Median is the value of the middle item of series when it is arranged in ascending or descending order of magnitude. It divides the series into two halves; in one half all items are less than median, whereas in the other half all items have values higher than median. If the values of the items arranged in the ascending order are: 60, 74, 80, 90, 95, 100, then the value of the 4th item viz., 88 is the value of median. We can also write thus:

* If we use assumed average A, then mean would be worked out as under:

$\bar{X} = A + \frac{\sum(X_i - A)}{n}$ or $\bar{X} = A + \frac{\sum f_i(X_i - A)}{\sum f_i}$, in case of frequency distribution. This is also known as short cut method of finding $\bar{X}$.

$$\text{Median } (M) = \text{Value of } \left(\frac{n+1}{2}\right)\text{th item}$$

Median is a positional average and is used only in the context of qualitative phenomena, for example, in estimating intelligence, etc., which are often encountered in sociological fields. Median is not useful where items need to be assigned relative importance and weights. It is not frequently used in sampling statistics.

Mode is the most commonly or frequently occurring value in a series. The mode in a distribution is that item around which there is maximum concentration. In general, mode is the size of the item which has the maximum frequency, but at items such an item may not be mode on account of the effect of the frequencies of the neighbouring items. Like median, mode is a positional average and is not affected by the values of extreme items. it is, therefore, useful in all situations where we want to eliminate the effect of extreme variations. Mode is particularly useful in the study of popular sizes. For example, a manufacturer of shoes is usually interested in finding out the size most in demand so that he may manufacture a larger quantity of that size. In other words, he wants a modal size to be determined for median or mean size would not serve his purpose. but there are certain limitations of mode as well. For example, it is not amenable to algebraic treatment and sometimes remains indeterminate when we have two or more model values in a series. It is considered unsuitable in cases where we want to give relative importance to items under consideration.

Geometric mean is also useful under certain conditions. It is defined as the nth root of the product of the values of n times in a given series. Symbolically, we can put it thus:

$$\text{Geometric mean (or G.M.)} = \sqrt[n]{\pi X_i}$$

$$= \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \ldots X_n}$$

where

G.M. = geometric mean,

n = number of items.

X_i = ith value of the variable X

π = conventional product notation

For instance, the geometric mean of the numbers, 4, 6, and 9 is worked out as

$$\text{G.M.} = \sqrt[3]{4.6.9}$$

$$= 6$$

The most frequently used application of this average is in the determination of average per cent of change i.e., it is often used in the preparation of index numbers or when we deal in ratios.

Harmonic mean is defined as the reciprocal of the average of reciprocals of the values of items of a series. Symbolically, we can express it as under:

$$\text{Harmonic mean (H.M.)} = \text{Rec.}\frac{\sum \text{Rec} X_i}{n}$$

$$= \text{Rec.}\frac{\text{Rec.}\, X_1 + \text{Rec.}\, X_2 + \ldots + \text{Rec.}\, X_n}{n}$$

where

H.M. = Harmonic mean

Rec. = Reciprocal

X_i = *i*th value of the variable X

n = number of items

For instance, the harmonic mean of the numbers 4, 5, and 10 is worked out as

$$\text{H.M.} = \text{Rec}\frac{1/4 + 1/5 + 1/10}{3} = \text{Rec}\frac{\frac{15+12+6}{60}}{3}$$

$$= \text{Rec}\left(\frac{33}{60} \times \frac{1}{3}\right) = \frac{60}{11} = 5.45$$

Harmonic mean is of limited application, particularly in cases where time and rate are involved. The harmonic mean gives largest weight to the smallest item and smallest weight to the largest item. As such it is used in cases like time and motion study where time is variable and distance constant.

From what has been stated above, we can say that there are several types of statistical averages. Researcher has to make a choice for some average. There are no hard and fast rules for the selection of a particular average in statistical analysis for the selection of an average mostly depends on the nature, type of objectives of the research study. One particular type of average cannot be taken as appropriate for all types of studies. The chief characteristics and the limitations of the various averages must be kept in view; discriminate use of average is very essential for sound statistical analysis.

MEASURES OF DISPERSION

An averages can represent a series only as best as a single figure can, but it certainly cannot reveal the entire story of any phenomenon under study. Specially it fails to give any idea about the scatter of the values of items of a variable in the series around the true value of average. In order to measure this scatter, statistical devices called measures of dispersion are calculated. Important measures of dispersion are (a) range, (b) mean deviation, and (c) standard deviation.

(a) *Range* is the simplest possible measure of dispersion and is defined as the difference between the values of the extreme items of a series. Thus,

$$\text{Range} = \begin{pmatrix}\text{Highest value of an}\\ \text{item in a series}\end{pmatrix} - \begin{pmatrix}\text{Lowest value of an}\\ \text{item in a series}\end{pmatrix}$$

The utility of range is that it gives an idea of the variability very quickly, but the drawback is that range is affected very greatly by fluctuations of sampling. Its value is never stable, being based on only two values of the variable. As such, range is mostly used as a rough measure of variability and is not considered as an appropriate measure in serious research studies.

(b) *Mean deviation* is the average of difference of the values of items from some average of the series. Such a difference is technically described as deviation. In calculating mean deviation we ignore the minus sign of deviations while taking their total for obtaining the mean deviation. Mean deviation is, thus, obtained as under:

Mean deviation from mean $(\delta_{\bar{X}}) = \dfrac{\Sigma|X_i - \bar{X}|}{n}$, if deviations, $|X_i - \bar{X}|$, are obtained from arithmetic average.

or

Mean deviation from median $(\delta_m) = \dfrac{\Sigma|X_i - M|}{n}$, if deviations, $|X_i - M|$, are obtained from median

or

Mean deviation from mode $(\delta_z) = \dfrac{\Sigma|X_i - Z|}{n}$, if deviations, $|X_i - Z|$, are obtained from mode.

where δ = Symbol for mean deviation (pronounced as delta);

X_i = *i*th values of the variable X;

n = number of items;

$\bar{X}$ = Arithmetic average;

M = Median;

Z = Mode.

When mean deviation is divided by the average used in finding out the mean deviation itself, the resulting quantity is described as the *coefficient of mean deviation*. Coefficient of mean deviation is a relative measure of dispersion and is comparable to similar measure of other series. Mean deviation and its coefficient are used in statistical studies for judging the variability, and thereby render the study of central tendency of a series more precise by throwing light on the typicalness of an average. It is a better measure of variability than range as it takes into consideration the values of all items of a series. Even then it is not a frequently used measure as it is not amenable to algebraic process.

(c) *Standard deviation* is most widely used measure of dispersion of a series and is commonly denoted by the symbol 'σ' (pronounced as sigma). Standard deviation is defined as the square-root of the average of squares of deviations, when such deviations for the values of individual items in a series are obtained from the arithmetic average. It is worked out as under:

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n}}$$

*If we use assumed average, A, in place of $\bar{X}$ while finding deviations, then standard deviation would be worked out as under:

$$\sigma = \sqrt{\frac{\Sigma(X_i - A)^2}{n} - \left(\frac{\Sigma(X_i - A)}{n}\right)^2}$$

Or

$$\sigma = \sqrt{\frac{\Sigma f_i(X_i - A)^2}{\Sigma f_i} - \left(\frac{\Sigma f_i(X_i - A)}{\Sigma f_i}\right)^2}, \text{ in case of frequency distribution.}$$

This is also known as the short-cut method of finding σ.

Or

$$\text{Standard deviation}(\sigma) = \sqrt{\frac{\Sigma f_i (X_i - \bar{X})^2}{\Sigma f_i}}$$, in case of frequency distribution

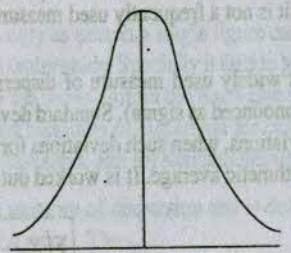
where f_i means the frequency of the ith item.

When we divide the standard deviation by the arithmetic average of the series, the resulting quantity is known as *coefficient of standard deviation which* happens to be a relative measure and is often used for comparing with similar measure of other series. When this coefficient of standard deviation is multiplied by 100, the resulting figure is known as *coefficient of variation*. Sometimes, we work out the square of standard deviation, known as *variance*, which is frequently used in the context of analysis of variation.

The standard deviation (along with several related measures like variance, coefficient of variation, etc.) is used mostly in research studies and is regarded as a very satisfactory measure of dispersion in a series. It is amenable to mathematical manipulation because the algebraic signs are not ignored in its calculation (as we ignore in case of mean deviation). It is less affected by fluctuations of sampling. These advantages make standard deviation and its coefficient a very popular measure of the scatteredness of a series. It is popularly used in the context of estimation and testing of hypotheses.

MEASURES OF ASYMMETRY (SKEWNESS)

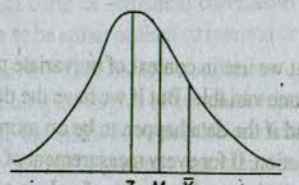
When the distribution of item in a series happens to be perfectly symmetrical, we then have the following type of curve for the distribution:



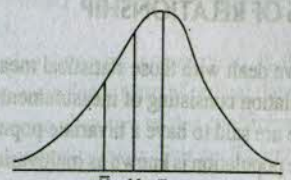
Curve showing no skewness in which case we have $\bar{X} = M = Z$

Fig. 7.1

Such a curve is technically described as a *normal curve* and the relating distribution as normal distribution. Such a curve is perfectly bell shaped curve in which case the value of $\bar{X}$ or M or Z is just the same and skewness is altogether absent. But if the curve is distorted (whether on the right side or on the left side), we have asymmetrical distribution which indicates that there is skewness. If the curve is distorted on the right side, we have positive skewness but when the curve is distorted towards left, we have negative skewness as shown here under:



Curve showing positive skewness
In case of positive skewness we have:
$Z < M < \bar{X}$



Curve showing negative skewness
In case of negative skewness we have:
$\bar{X} < M < Z$

Fig. 7.2

Skewness is, thus, a measure of asymmetry and shows the manner in which the items are clustered around the average. In a symmetrical distribution, the items show a perfect balance on either side of the mode, but in a skew distribution the balance is thrown to one side. The amount by which the balance exceeds on one side measures the skewness of the series. The difference between the mean, median or the mode provides an easy way of expressing skewness in a series. In case of positive skewness, we have $Z < M < \bar{X}$ and in case of negative skewness we have $\bar{X} < M < Z$. Usually we measure skewness in this way:

Skewness = $\bar{X} - Z$ and its coefficient (j) is worked

out as $j = \dfrac{\bar{X} - Z}{\sigma}$

In case Z is not well defined, then we work out skewness as under:

Skewness = $3(\bar{X} - M)$ and its coefficient (j) is worked

out as $j = \dfrac{3(\bar{X} - M)}{\sigma}$

The significance of skewness lies in the fact that through it one can study the formation of series and can have the idea about the shape of the curve, whether normal or otherwise, when the items of a given series are plotted on a graph.

Kurtosis is the measure of flat-toppedness of a curve. A bell shaped curve or the normal curve is Mesokurtic because it is kurtic in the centre; but if the curve is relatively more peaked than the normal curve, it is called Leptokurtic whereas a curve is more flat than the normal curve, it is called Platykurtic. In brief, Kurtosis is the humpedness of the curve and points to the nature of distribution of items in the middle of a series.

It may be pointed out here that knowing the shape of the distribution curve is crucial to the use of statistical methods in research analysis since most methods make specific assumptions about the nature of the distribution curve.

MEASURES OF RELATIONSHIP

So far we have dealt with those statistical measures that we use in context of univariate population i.e., the population consisting of measurement of only one variable. But if we have the data on two variables, we are said to have a bivariate population and if the data happen to be on more than two variables, the population is known as multivariate population. If for every measurement of a variable, X, we have corresponding value of a second variable, Y, the resulting pairs of values are called a bivariate population. In addition, we may also have a corresponding value of the third variable, Z, or the forth variable, W, and so on, the resulting pairs of values are called a multivariate population. In case of bivariate or multivariate populations, we often wish to know the relation of the two and/or more variables in the data to one another. We may like to know, for example, whether the number of hours students devote for studies is somehow related to their family income, to age, to sex or to similar other factor. There are several methods of determining the relationship between variables, but no method can tell us for certain that a correlation is indicative of causal relationship. Thus we have to answer two types of questions in bivariate or multivariate populations viz.,

(i) Does there exist association or correlation between the two (or more) variables? If yes, of what degree?

(ii) Is there any cause and effect relationship between the two variables in case of the bivariate population or between one variable on one side and two or more variables on the other side in case of multivariate population? If yes, of what degree and in which direction?

The first question is answered by the use of correlation technique and the second question by the technique of regression. There are several methods of applying the two techniques, but the important ones are as under:

In case of bivariate population: Correlation can be studied through (a) cross tabulation; (b) Charles Spearman's coefficient of correlation; (c) Karl Pearson's coefficient of correlation; whereas cause and effect relationship can be studied through simple regression equations.

In case of multivariate population: Correlation can be studied through (a) coefficient of multiple correlation; (b) coefficient of partial correlation; whereas cause and effect relationship can be studied through multiple regression equations.

We can now briefly take up the above methods one by one.

Cross tabulation approach is specially useful when the data are in nominal form. Under it we classify each variable into two or more categories and then cross classify the variables in these sub-categories. Then we look for interactions between them which may be symmetrical, reciprocal or asymmetrical. A symmetrical relationship is one in which the two variables vary together, but we assume that neither variable is due to the other. A reciprocal relationship exists when the two variables mutually influence or reinforce each other. Asymmetrical relationship is said to exist if one variable (the independent variable) is responsible for another variable (the dependent variable). The cross classification procedure begins with a two-way table which indicates whether there is or there is not an interrelationship between the variables. This sort of analysis can be further elaborated in which case a third factor is introduced into the association through cross-classifying the three variables. By doing so we find conditional relationship in which factor X appears to affect factor Y only when factor Z is held constant. The correlation, if any, found through this approach is not considered a very

powerful form of statistical correlation and accordingly we use some other methods when data happen to be either ordinal or interval or ratio data.

Charles Spearman's coefficient of correlation (or rank correlation) is the technique of determining the degree of correlation between two variables in case of ordinal data where ranks are given to the different values of the variables. The main objective of this coefficient is to determine the extent to which the two sets of ranking are similar or dissimilar. This coefficient is determined as under:

$$\text{Spearman's coefficient of correlation (or } r_s) = 1 - \left[\frac{6\sum d_i^2}{n\left(n^2-1\right)}\right]$$

where d_i = difference between ranks of ith pair of the two variables;

n = number of pairs of observations.

As rank correlation is a non-parametric technique for measuring relationship between paired observations of two variables when data are in the ranked form, we have dealt with this technique in greater details later on in the book in chapter entitled 'Hypotheses Testing II (Non-parametric tests)'.

Karl Pearson's coefficient of correlation (or simple correlation) is the most widely used method of measuring the degree of relationship between two variables. This coefficient assumes the following:

(i) that there is linear relationship between the two variables;

(ii) that the two variables are casually related which means that one of the variables is independent and the other one is dependent; and

(iii) a large number of independent causes are operating in both variables so as to produce a normal distribution.

Karl Pearson's coefficient of correlation can be worked out thus.

$$\text{Karl Pearson's coefficient of correlation (or } r)^* = \frac{\Sigma\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n\cdot\sigma_X\cdot\sigma_Y}$$

*Alternatively, the formula can be written as:

$$r = \frac{\Sigma\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\Sigma\left(X_i-\bar{X}\right)^2\cdot\Sigma\left(Y_i-\bar{Y}\right)^2}}$$

Or

$$r = \frac{\text{Covariance between } X \text{ and } Y}{\sigma_x\cdot\sigma_y} = \frac{\Sigma\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)/n}{\sigma_x\cdot\sigma_y}$$

Or

$$r = \frac{\Sigma X_iY_i - n\cdot\bar{X}\cdot\bar{Y}}{\sqrt{\Sigma X_i^2 - n\bar{X}^2}\,\sqrt{\Sigma Y_i^2 - n\bar{Y}^2}}$$

(This applies when we take zero as the assumed mean for both variables, X and Y.)

where X_i = ith value of X variable

$\bar{X}$ = mean of X

Y_i = ith value of Y variable

$\bar{Y}$ = Mean of Y

n = number of pairs of observations of X and Y

σ_X = Standard deviation of X

σ_Y = Standard deviation of Y

In case we use assumed means (A_x and A_y for variables X and Y respectively) in place of true means, then Karl Person's formula is reduced to:

$$\frac{\frac{\sum dx_i \cdot dy_i}{n} - \left(\frac{\sum dx_i \cdot \sum dy_i}{n}\right)}{\sqrt{\frac{\sum dx_i^2}{n} - \left(\frac{\sum dx_i}{n}\right)^2}\sqrt{\frac{\sum dy_i^2}{n} - \left(\frac{\sum dy_i}{n}\right)^2}}$$

$$\frac{\frac{\sum dx_i \cdot dy_i}{n} - \left(\frac{\sum dx_i \cdot \sum dy_i}{n}\right)}{\sqrt{\frac{\sum dx_i^2}{n} - \left(\frac{\sum dx_i}{n}\right)^2}\sqrt{\frac{\sum dy_i^2}{n} - \left(\frac{\sum dy_i}{n}\right)^2}}$$

where $\sum dx_i = \sum (X_i - A_x)$

$\sum dy_i = \sum (Y_i - A_y)$

$\sum dx_i^2 = \sum (X_i - A_x)^2$

$\sum dy_i^2 = \sum (Y_i - A_y)^2$

$\sum dx_i \cdot dy_i = \sum (X_i - A_x)(Y_i - A_y)$

n = number of pairs of observations of X and Y.

This is the short cut approach for finding 'r' in case of ungrouped data. If the data happen to be grouped data (i.e., the case of bivariate frequency distribution), we shall have to write Karl Pearson's coefficient of correlation as under:

$$\frac{\frac{\sum f_{ij} \cdot dx_i \cdot dy_j}{n} - \left(\frac{\sum f_i dx_i}{n} \cdot \frac{\sum f_j dy_j}{n}\right)}{\sqrt{\frac{\sum f_i dx_i^2}{n} - \left(\frac{\sum f_i dx_i}{n}\right)}\sqrt{\frac{\sum f_i dy_j^2}{n} - \left(\frac{\sum f_j dy_j}{n}\right)^2}}$$

where f_{ij} is the frequency of a particular cell in the correlation table and all other values are defined as earlier.

Karl Pearson's coefficient of correlation is also known as the product moment correlation coefficient. The value of 'r' lies between ± 1. Positive values of r indicate positive correlation between the two variables (i.e., changes in both variables take place in the statement direction), whereas negative values of 'r' indicate negative correlation i.e., changes in the two variables taking place in the opposite directions. A zero value of 'r' indicates that there is no association between the two variables. When $r = (+)1$, it indicates perfect positive correlation and when it is $(-)1$, it indicates perfect negative correlation, meaning thereby that variations in independent variable (X) explain 100% of the variations in the dependent variable (Y). We can also say that for a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then correlation will be termed as perfect positive. But if such change occurs in the opposite direction, the correlation will be termed as perfect negative. The value of 'r' nearer to +1 or –1 indicates high degree of correlation between the two variables.

SIMPLE REGRESSION ANALYSIS

Regression is the determination of a statistical relationship between two or more variables. In simple regression, we have only two variables, one variable (defined as independent) is the cause of the behaviour of another one (defined as dependent variable). Regression can only interpret what exists physically i.e., there must be a physical way in which independent variable X can affect dependent variable Y. The basic relationship between X and Y is given by

$$\hat{Y} = a + bX$$

where the symbol $\hat{Y}$ denotes the estimated value of Y for a given value of X. This equation is known as the regression equation of Y on X (also represents the regression line of Y on X when drawn on a graph) which means that each unit change in X produces a change of b in Y, which is positive for direct and negative for inverse relationships.

Then generally used method to find the 'best' fit that a straight line of this kind can give is the least-square method. To use it efficiently, we first determine

$$\sum x_i^2 = \sum X_i^2 - n\bar{X}^2$$

$$\sum y_i^2 = \sum Y_i^2 - n\bar{Y}^2$$

$$\sum x_i y_i = \sum X_i Y_i - n\bar{X} \cdot \bar{Y}$$

Then $$b = \frac{\sum x_i y_i}{\sum x_i^2}, \; a = \bar{Y} - b\bar{X}$$

These measures define a and b which will give the best possible fit through the original X and Y points and the value of r can then be worked out as under:

$$r = \frac{b\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}}$$

Thus, the regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables which can be used for the purpose of prediction of the values of dependent variable, given the values of the independent variable.

[Alternatively, for fitting a regression equation of the type $\hat{Y} = a + bX$ to the given values of X and Y variables, we can find the values of the two constants viz., a and b by using the following two normal equations:

$$\sum Y_i = na + b\sum X_i$$

$$\sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

and then solving these equations for finding a and b values. Once these values are obtained and have been put in the equation $\hat{Y} = a + bX$, we say that we have fitted the regression equation of Y on X to the given data. In a similar fashion, we can develop the regression equation of X and Y viz., $\hat{X} = a + bX$, presuming Y as an independent variable and X as dependent variable].

MULTIPLE CORRELATION AND REGRESSION

When there are two or more than two independent variables, the analysis concerning relationship is known as multiple correlation and the equation describing such relationship as the multiple regression equation. We here explain multiple correlation and regression taking only two independent variables and one dependent variable (Convenient computer programs exist for dealing with a great number of variables). In this situation the results are interpreted as shown below:

Multiple regression equation assumes the form

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

where X_1 and X_2 are two independent variables and Y being the dependent variable, and the constants a, b_1 and b_2 can be solved by solving the following three normal equations:

$$\sum Y_i = na + b_1\sum X_{1i} + b_2\sum X_{2i}$$

$$\sum X_{1i}Y_i = a\sum X_{1i} + b_1\sum X_{1i}^2 + b_2\sum X_{1i}X_{2i}$$

$$\sum X_{2i}Y_i = a\sum X_{2i} + b_1\sum X_{1i}X_{2i} + b_2\sum X_{2i}^2$$

(It may be noted that the number of normal equations would depend upon the number of independent variables. If there are 2 independent variables, then 3 equations, if there are 3 independent variables then 4 equations and so on, are used.)

In multiple regression analysis, the regression coefficients (viz., b_1 b_2) become less reliable as the degree of correlation between the independent variables (viz., X_1, X_2) increases. If there is a high degree of correlation between independent variables, we have a problem of what is commonly described as the *problem of multicollinearity*. In such a situation we should use only one set of the independent variable to make our estimate. In fact, adding a second variable, say X_2, that is correlated with the first variable, say X_1, distorts the values of the regression coefficients. Nevertheless, the prediction for the dependent variable can be made even when multicollinearity is present, but in such a situation enough care should be taken in selecting the independent variables to estimate a dependent variable so as to ensure that multi-collinearity is reduced to the minimum.

With more than one independent variable, we may make a difference between the collective effect of the two independent variables and the individual effect of each of them taken separately. The collective effect is given by the coefficient of multiple correlation,

$R_{y\cdot x_1x_2}$ defined as under:

$$R_{y\cdot x_1x_2} = \sqrt{\frac{b_1\sum Y_iX_{1i} - n\bar{Y}\bar{X}_1 + b_2\sum Y_iX_{2i} - n\bar{Y}\bar{X}_2}{\sum Y_i^2 - n\bar{Y}^2}}$$

Alternatively, we can write

$$R_{y\cdot x_1x_2} = \sqrt{\frac{b_1\sum x_{1i}y_i + b_2\sum x_{2i}y_i}{\sum Y_i^2}}$$

where

$$x_{1i} = (X_{1i} - \bar{X}_1)$$

$$x_{2i} = (X_{2i} - \bar{X}_2)$$

$$y_i = (Y_i - \bar{Y})$$

and b_1 and b_2 are the regression coefficients.

PARTIAL CORRELATION

Partial correlation measures separately the relationship between two variables in such a way that the effects of other related variables are eliminated. In other words, in partial correlation analysis, we aim at measuring the relation between a dependent variable and a particular independent variable by holding all other variables constant. Thus, each partial coefficient of correlation measures the effect of its independent variable on the dependent variable. To obtain it, it is first necessary to compute the simple coefficients of correlation between each set of pairs of variables as stated earlier. In the case of two independent variables, we shall have two partial correlation coefficients denoted $r_{yx_1\cdot x_2}$ and $r_{yx_2\cdot x_1}$ which are worked out as under:

$$r_{yx_1\cdot x_2} = \frac{R^2_{y\cdot x_1x_2} - r^2_{yx_2}}{1 - r^2_{yx_2}}$$

This measures the effort of X_1 on Y, more precisely, that proportion of the variation of Y not explained by X_2 which is explained by X_1. Also,

$$r_{yx_2\cdot x_1} = \frac{R^2_{y\cdot x_1x_2} - r^2_{yx_1}}{1 - r^2_{yx_1}}$$

in which X_1 and X_2 are simply interchanged, given the added effect of X_2 on Y.

Alternatively, we can work out the partial correlation coefficients thus:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{1 - r^2_{yx_2}} \sqrt{1 - r^2_{x_1x_2}}}$$

and

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{1 - r^2_{yx_1}} \sqrt{1 - r^2_{x_1x_2}}}$$

These formulae of the alternative approach are based on simple coefficients of correlation (also known as zero order coefficients since no variable is held constant when simple correlation coefficients are worked out). The partial correlation coefficients are called first order coefficients when one variable is held constant as shown above; they are known as second order coefficients when two variables are held constant and so on.

ASSOCIATION IN CASE OF ATTRIBUTES

When data is collected on the basis of some attribute or attributes, we have statistics commonly termed as statistics of attributes. It is not necessary that the objects may process only one attribute; rather it would be found that the objects possess more than one attribute. In such a situation our interest may remain in knowing whether the attributes are associated with each other or not. For example, among a group of people we may find that some of them are inoculated against small-pox and among the inoculated we may observe that some of them suffered from small-pox after inoculation. The important question which may arise for the observation is regarding the efficiency of inoculation for its popularity will depend upon the immunity which it provides against small-pox. In other words, we may be interested in knowing whether inoculation and immunity from small-pox are associated. Technically, we say that the two attributes are associated if they appear together in a greater number of cases than is to be expected if they are independent and not simply on the basis that they are appearing together in a number of cases as is done in ordinary life.

The association may be positive or negative (negative association is also known as disassociation). If class frequency of *AB*, symbolically written as (*AB*), is greater than the expectation of *AB* being together if they are independent, then we say the two attributes are positively associated; but if the class frequency of *AB* is less than this expectation, the two attributes are said to be negatively associated. In case the class frequency of *AB* is equal to expectation, the two attributes are considered as independent i.e., are said to have no association. It can be put symbolically as shown hereunder:

If $(AB) > \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then *AB* are positively related/associated.

If $(AB) < \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then *AB* are negatively related/associated.

If $(AB) = \frac{(A)}{N} \times \frac{(B)}{N} \times N$, then *AB* are independent i.e., have no association.

Where (*AB*) = frequency of class *AB* and

$\frac{(A)}{N} \times \frac{(B)}{N} \times N$ = Expectation of *AB*, if *A* and *B* are independent, and *N* being the number of items

In order to find out the degree or intensity of association between two or more sets of attributes, we should work out the coefficient of association. Professor Yule's coefficient of association is most popular and is often used for the purpose. It can be mentioned as under:

$$Q_{AB} = \frac{(AB)(ab) - (Ab)(aB)}{(AB)(ab) + (Ab)(aB)}$$

where,

Q_{AB} = Yule's coefficient of association between attributes *A* and *B*.

(*AB*) = Frequency of class *AB* in which *A* and *B* are present.

(*Ab*) = Frequency of class *Ab* in which *A* is present but *B* is absent.

(*aB*) = Frequency of class *aB* in which *A* is absent but *B* is present.

(*ab*) = Frequency of class *ab* in which both *A* and *B* are absent.

The value of this coefficient will be somewhere between +1 and –1. If the attributes are completely associated (perfect positive association) with each other, the coefficient will be +1, and if they are completely disassociated (perfect negative association), the coefficient will be –1. If the attributes are completely independent of each other, the coefficient of association will be 0. The varying degrees of the coefficients of association are to be read and understood according to their positive and negative nature between +1 and –1.

Sometimes the association between two attributes *A* and *B*, may be regarded as unwarranted when we find that the observed association between *A* and *B* is due to the association of both *A* and *B* with another attribute *C*. For example, we may observe positive association between inoculation and exemption for small-pox, but such association may be the result of the fact that there is positive association between inoculation and richer section of society and also that there is positive association between exemption from small-pox and richer section of society. The sort of association between *A* and *B* in the population of *C* is described as *partial association* as distinguished from *total association* between *A* and *B* in the overall universe. We can workout the coefficient of partial association between *A* and *B* in the population of *C* by just modifying the above stated formula for finding association between *A* and *B* as shown below:

$$Q_{AB.C} = \frac{(ABC)(abC) - (AbC)(aBC)}{(ABC)(abC) + (AbC)(aBC)}$$

where,

$Q_{AB.C}$ = Coefficient of partial association between *A* and *B* in the population of *C*; and all other values are the class frequencies of the respective classes (*A*, *B*, *C* denotes the presence of concerning attributes and *a*, *b*, *c* denotes the absence of concerning attributes).

At times, we may come across cases of *illusory association*, wherein association between two attributes does not correspond to any real relationship. This sort of association may be the result of

some attribute, say C with which attributes A and B are associated (but in reality there is no association between A and B). Such association may also be the result of the fact that the attributes A and B might not have been properly defined or might not have been correctly recorded. Researcher must remain alert and must not conclude association between A and B when in fact there is no such association in reality.

In order to judge the significance of association between two attributes, we make use of *Chi-square test** by finding the value of Chi-square (χ_2) and using Chi-square distribution the value of χ_2 can be worked out as under:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad i = 1, 2, 3 \ldots$$

where $j = 1, 2, 3 \ldots$

O_{ij} = observed frequencies

E_{ij} = expected frequencies.

Association between two attributes in case of manifold classification and the resulting contingency table can be studied as explained below:

We can have manifold classification of the two attributes in which case each of the two attributes are first observed and then each one is classified into two or more subclasses, resulting into what is called as contingency table. The following is an example of 4 x 4 contingency table with two attributes A and B, each one of which has been further classified into four sub-categories.

Table 7.2: 4 x 4 Contingency Table

		Attribute A				
		A_1	A_2	A_3	A_4	Total
Attribute B	B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	(A_4B_1)	(B_1)
	B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	(A_4B_2)	(B_2)
	B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	(A_4B_3)	(B_3)
	B_4	(A_1B_4)	(A_2B_4)	(A_3B_4)	(A_4B_4)	(B_4)
	Total	(A_1)	(A_2)	(A_3)	(A_4)	N

Association can be studied in a contingency table through Yule's coefficient of association as stated above, but for this purpose we have to reduce the contingency table into 2 x 2 table by combining some classes. For instance, if we combine $(A_1) + (A_2)$ to form (A) and $(A_3) + (A_4)$ to form (a) and similarly if we combine $(B_1) + (B_2)$ to form (B) and $(B_3) + (B_4)$ to form (b) in the above contingency table, then we can write the table in the form of a 2 x 2 table as shown in Table 4.3

*See Chapter "Chi-square test" for all details.

Table 7.3

		Attribute		
		A	a	Total
Attribute	B	(AB)	(aB)	(B)
	b	(Ab)	(ab)	(b)
	Total	(A)	(a)	N

After reducing a contingency table in a two-by-two table through the process of combining some classes, we can work out the association as explained above. But the practice of combining classes is not considered very correct and at times it is inconvenient also, Karl Pearson has suggested a measure known as *Coefficient of mean square contingency* for studying association in contingency tables. This can be obtained as under:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

where

C = Coefficient of contingency

χ^2 = Chi-square value which is $= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

N = number of items.

This is considered a satisfactory measure of studying association in contingency tables.

OTHER MEASURES

1. Index numbers: When series are expressed in same units, we can use averages for the purpose of comparison, but when the units in which two or more series are expressed happen to be different, statistical averages cannot be used to compare them. In such situations we have to rely upon some relative measurement which consists in reducing the figures to a common base. Once such method is to convert the series into a series of index numbers. This is done when we express the given figures as percentages of some specific figure on a certain data. We can, thus, define an index number as a number which is used to measure the level of a given phenomenon as compared to the level of the same phenomenon at some standard date. The use of index number weights more as a special type of average, meant to study the changes in the effect of such factors which are incapable of being measured directly. But one must always remember that index numbers measure only the relative changes.

Changes in various economic and social phenomena can be measured and compared through index numbers. Different indices serve different purposes. Specific commodity indices are to serve as a measure of changes in the phenomenon of that commodity only. Index numbers may measure cost of living of different classes of people. In economic sphere, index numbers are often termed as

'economic barometers measuring the economic phenomenon in all its aspects either directly by measuring the same phenomenon or indirectly by measuring something else which reflects upon the main phenomenon.

But index numbers have their own limitations with which researcher must always keep himself aware. For instance, index numbers are only approximate indicators and as such give only a fair idea of changes but cannot give an accurate idea. Chances of error also remain at one point or the other while constructing an index number but this does not diminish the utility of index numbers for they still can indicate the trend of the phenomenon being measured. However, to avoid fallacious conclusions, index numbers prepared for one purpose should not be used for other purposes or for the same purpose at other places.

2. Time series analysis: In the context of economic and business researches, we may obtain quite often data relating to some time period concerning a given phenomenon. Such data is labelled as 'Time Series'. More clearly it can be stated that series of successive observations of the given phenomenon over a period of time are referred to as time series. Such series are usually the result of the effects of one or more of the following factors:

(i) *Secular trend* or long term trend that shows the direction of the series in a long period of time. The effect of trend (whether it happens to be a growth factor or a decline factor) is gradual, but extends more or less consistently throughout the entire period of time under consideration. Sometimes, secular trend is simply stated as trend (or T).

(ii) *Short time oscillations* i.e., changes taking place in the short period of time only and such changes can be the effect of the following factors:

(a) *Cyclical fluctuations (or C)* are the fluctuations as a result of business cycles and are generally referred to as long term movements that represent consistently recurring rises and declines in an activity.

(b) *Seasonal fluctuations (or S)* are of short duration occurring in a regular sequence at specific intervals of time. Such fluctuations are the result of changing seasons. Usually these fluctuations involve patterns of change within a year that tend to be repeated from year to year. Cyclical fluctuations and seasonal fluctuations taken together constitute short-period regular fluctuations.

(c) *Irregular fluctuations (or I)*, also known as Random fluctuations, are variations which take place in a completely unpredictable fashion.

All these factors stated above are termed as components of time series and when we try to analyse time series, we try to isolate and measure the effects of various types of these factors on a series. To study the effect of one type of factor, the other type of factor is eliminated from the series. The given series is, thus, left with the effects of one type of factor only.

For analysing time series, we usually have two models; (1) multiplicative model; and (2) additive model. Multiplicative model assumes that the various components interact in a multiplicative manner to produce the given values of the overall time series and can be stated as under:

$$Y = T \times C \times S \times I$$

where

Y = observed values of time series, T = Trend, C = Cyclical fluctuations, S = Seasonal fluctuations, I = Irregular fluctuations.

Additive model considers the total of various components resulting in the given values of the overall time series and can be stated as:

$$Y = T + C + S + I$$

There are various methods of isolating trend from the given series viz., the free hand method, semi-average method, method of moving averages, method of least squares and similarly there are methods of measuring cyclical and seasonal variations and whatever variations are left over are considered as random or irregular fluctuations.

The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of business firm(s). The past trends can be used to evaluate the success or failure of management policy or policies practiced hitherto. On the basis of past trends, the future patterns can be predicted and policy or policies may accordingly be formulated. We can as well study properly the effects of factors causing changes in the short period of time only, once we have eliminated the effects of trend. By studying cyclical variations, we can keep in view the impact of cyclical changes while formulating various policies to make them as realistic as possible. The knowledge of seasonal variations will be of great help to us in taking decisions regarding inventory, production, purchases and sales policies so as to optimize working results. Thus, analysis of time series is important in context of long term as well as short term forecasting and is considered a very powerful tool in the hands of business analysts and researchers.

Questions

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., 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 a brief note on different types of analysis of data pointing out the significance of each.
6. What do you mean by multivariate analysis? Explain how it differs from bivariate analysis.
7. How will you differentiate between descriptive statistics and inferential statistics? Describe the important statistical measures often used to summarise the survey/research data.
8. What does a measure of central tendency indicate? Describe the important measures of central tendency pointing out the situation when one measure is considered relatively appropriate in comparison to other measures.
9. Describe the various measures of relationships often used in context of research studies. Explain the meaning of the following correlation coefficients:

   (i) r_{yx}, (ii) $r_{yx_1 \cdot x_2}$, (iii) $R_{y \cdot x_1 x_2}$
10. Write short notes on the following:
    (i) Cross tabulation;
    (ii) Discriminant analysis;

(iii) Coefficient of contingency;

(iv) Multicollinearity;

(v) Partial association between two attributes.

11. "The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of business firms." Discuss.

12. "Changes in various economic and social phenomena can be measured and compared through index numbers". Explain this statement pointing out the utility of index numbers.

13. Distinguish between:

(i) Field editing and central editing;

(ii) Statistics of attributes and statistics of variables;

(iii) Exclusive type and inclusive type class intervals;

(iv) Simple and complex tabulation;

(v) Mechanical tabulation and cross tabulation.

14. "Discriminate use of average is very essential for sound statistical analysis". Why? Answer giving examples.

15. Explain how would you work out the following statistical measures often used by researchers?

(i) Coefficient of variation;

(ii) Arithmetic average;

(iii) Coefficient of skewness;

(iv) Regression equation of X on Y;

(v) Coefficient of $r_{yx_2 \cdot x_1}$.