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:

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

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

   $\sum$ = Symbol for summation

   X_i = Value of the *i*th item X, $i = 1, 2, ..., n$

   n = total number of items

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

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

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:

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

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

   w_i = weight of *i*th item X

   X_i = value of the *i*th 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:

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

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

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 *n*th root of the product of the values of *n* times in a given series. Symbolically, we can put it thus:

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

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

where

G.M. = geometric mean,

n = number of items.

X_i = *i*th 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:

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

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

where

H.M. = Harmonic mean

Rec. = Reciprocal

X_i = ith 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,

$$\text{Range} = \left(\begin{array}{c}\text{Highest value of an}\\ \text{item in a series}\end{array}\right) - \left(\begin{array}{c}\text{Lowest value of an}\\ \text{item in a series}\end{array}\right)$$

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 $\left(\delta_{\bar{X}}\right) = \dfrac{\Sigma\left|X_i - \bar{X}\right|}{n}$,   if deviations, $\left|X_i - \bar{X}\right|$, are obtained from

or   arithmetic average.

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

or   from median

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

mode.

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

X_i = ith 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:

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

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

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

Or

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

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

Or

$$\text{Standard deviation}(\sigma) = \sqrt{\frac{\Sigma f_i\left(X_i - \overline{X}\right)^2}{\Sigma f_i}} \text{ , in case of frequency distribution}$$
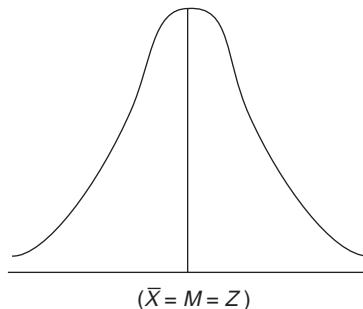
where f_i means the frequency of the ith item.

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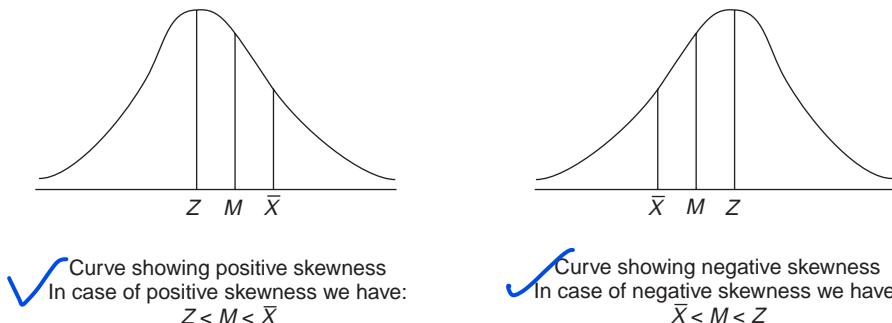
When the distribution of item in a series happens to be perfectly symmetrical, we then have the following type of curve for the distribution:

$(\overline{X} = M = Z)$

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

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 $\overline{X}$ 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:
$Z < M < \overline{X}$

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

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 $Z < M < \overline{X}$ and in case of negative skewness we have $\overline{X} < M < Z$. Usually we measure skewness in this way:

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

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

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

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

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

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:

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

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.

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

---

[*] Alternatively, the formula can be written as:

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

Or

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

Or

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

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

where X_i = ith value of X variable

$\overline{X}$ = mean of X

Y_i = ith value of Y variable

$\overline{Y}$ = 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:

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

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

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:

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

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

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

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

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

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

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

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

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,

$R_{y \cdot x_1 x_2}$ defined as under:

$$R_{y \cdot x_1 x_2} = \sqrt{\frac{b_1 \sum Y_i X_{1i} - n\bar{Y}\,\bar{X}_1 + b_2 \sum Y_i X_{2i} - n\bar{Y}\,\bar{X}_2}{\sum Y_i^2 - n\bar{Y}^2}}$$

Alternatively, we can write

$$R_{y \cdot x_1 x_2} = \sqrt{\frac{b_1 \sum x_{1i} y_i + b_2 \sum x_{2i} y_i}{\sum Y_i^2}}$$

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:

$$r_{yx_1 \cdot x_2} = \frac{R^2_{y \cdot x_1 x_2} - r^2_{yx_2}}{1 - r^2_{yx_2}}$$

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,

$$r_{yx_2 \cdot x_1} = \frac{R^2_{y \cdot x_1 x_2} - r^2_{yx_1}}{1 - r^2_{yx_1}}$$

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:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{1 - r_{yx_2}^2}\ \sqrt{1 - r_{x_1 x_2}^2}}$$

and

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{1 - r_{yx_1}^2}\ \sqrt{1 - r_{x_1 x_2}^2}}$$

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

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

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

Where (AB) = frequency of class AB and

$\dfrac{(A)}{N} \times \dfrac{(B)}{N} \times N$ = 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,

$\quad Q_{AB}$ = 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,

$\quad Q_{AB.C}$ = 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:

$$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \qquad i = 1, 2, 3 \dots$$

where
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, 2, 3 \dots$

$\qquad O_{ij}$ = observed frequencies

$\qquad E_{ij}$ = 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 4×4 contingency table with two attributes *A* and *B*, each one of which has been further classified into four sub-categories.

Table 7.2: 4×4 Contingency Table

		Attribute A				
		A_1	A_2	A_3	A_4	Total
	B_1	$(A_1 B_1)$	$(A_2 B_1)$	$(A_3 B_1)$	$(A_4 B_1)$	(B_1)
Attribute B	B_2	$(A_1 B_2)$	$(A_2 B_2)$	$(A_3 B_2)$	$(A_4 B_2)$	(B_2)
	B_3	$(A_1 B_3)$	$(A_2 B_3)$	$(A_3 B_3)$	$(A_4 B_3)$	(B_3)
	B_4	$(A_1 B_4)$	$(A_2 B_4)$	$(A_3 B_4)$	$(A_4 B_4)$	(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 2×2 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 2×2 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 $= \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$

N = number of items.

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

Read!

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;
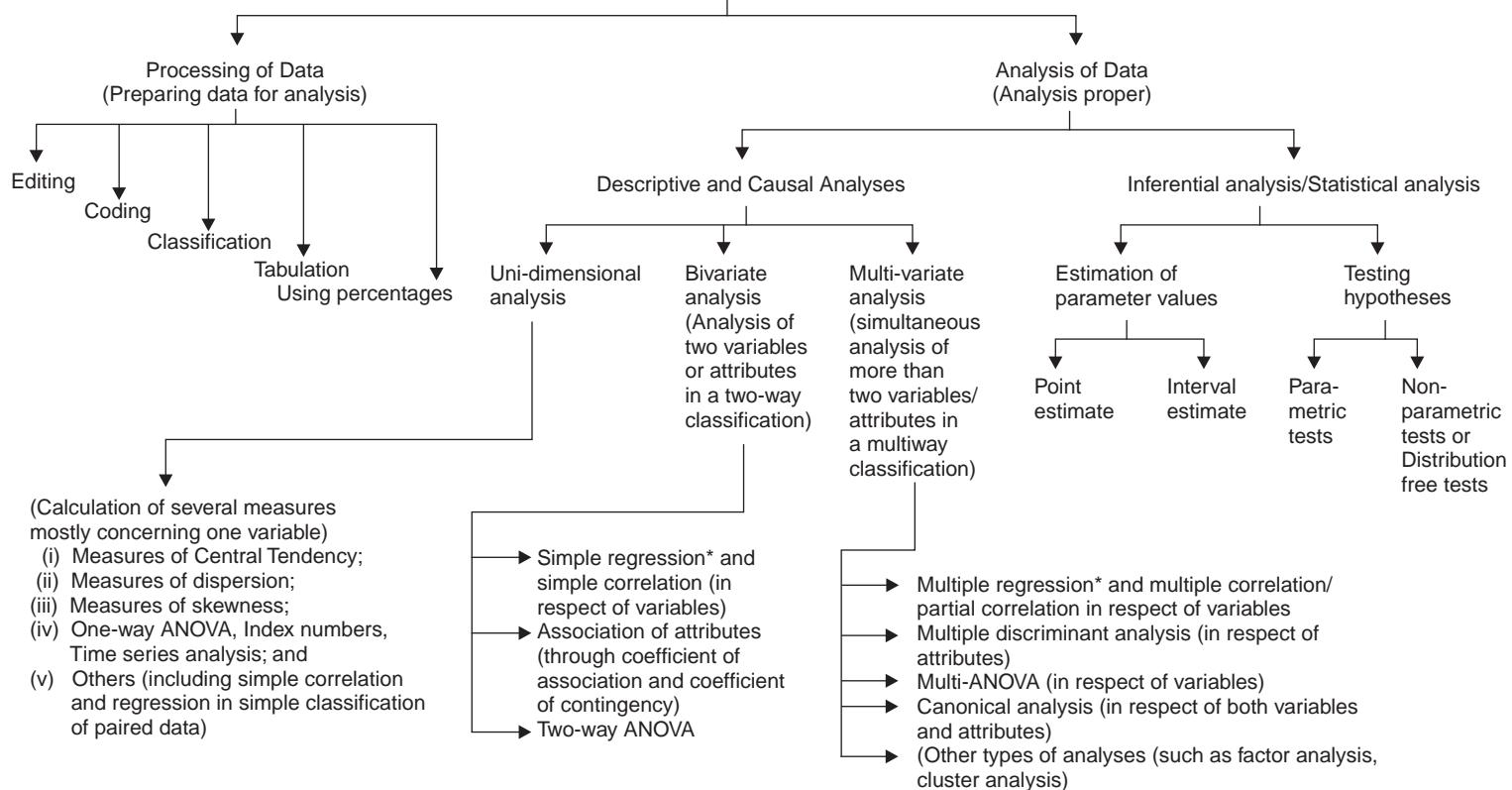
    (iii) Coefficient of contingency;

    (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 $r_{yx_2 \cdot x_1}$.

Appendix
(Summary chart concerning analysis of data)

Analysis of Data
(in a broad general way can be categorised into)

Processing of Data
(Preparing data for analysis)

- Editing
- Coding
- Classification
- Tabulation
- Using percentages

Analysis of Data
(Analysis proper)

Descriptive and Causal Analyses

- Uni-dimensional analysis
- Bivariate analysis (Analysis of two variables or attributes in a two-way classification)
- Multi-variate analysis (simultaneous analysis of more than two variables/attributes in a multiway classification)

Inferential analysis/Statistical analysis

- Estimation of parameter values
  - Point estimate
  - Interval estimate
- Testing hypotheses
  - Parametric tests
  - Non-parametric tests or Distribution free tests

(Calculation of several measures mostly concerning one variable)
- (i) Measures of Central Tendency;
- (ii) Measures of dispersion;
- (iii) Measures of skewness;
- (iv) One-way ANOVA, Index numbers, Time series analysis; and
- (v) Others (including simple correlation and regression in simple classification of paired data)

→ Simple regression* and simple correlation (in respect of variables)
→ Association of attributes (through coefficient of association and coefficient of contingency)
→ Two-way ANOVA

→ Multiple regression* and multiple correlation/partial correlation in respect of variables
→ Multiple discriminant analysis (in respect of attributes)
→ Multi-ANOVA (in respect of variables)
→ Canonical analysis (in respect of both variables and attributes)
→ (Other types of analyses (such as factor analysis, cluster analysis)

* Regression analysis (whether simple or multiple) is termed as Causal analysis whereas correlation analysis indicates simply co-variation between two or more variables.