...if we have information on more than one variables, we might be interested in seeing if there is any connection - any association - between them.

In many business research situations, the key to decision making lies in understanding the relationships between two or more variables. *For example,* in an effort to predict the behavior of the bond market, a broker might find it useful to know whether the interest rate of bonds is related to the prime interest rate. While studying the effect of advertising on sales, an account executive may find it useful to know whether there is a strong relationship between advertising dollars and sales dollars for a company.

In all these cases involving two or more variables, we may be interested in seeing:

- if there is any association between the variables;
- if there is an association, is it strong enough to be useful;
- if so, what form the relationship between the two variables takes;
- how we can make use of that relationship for predictive purposes, that is, forecasting;
- and how good such predictions will be.

WHAT IS CORRELATION?

Correlation is a measure of association between two or more variables. When two or more variables very in sympathy so that movement in one tends to be accompanied by corresponding movements in the other variable(s), they are said to be correlated.
"The correlation between variables is a measure of the nature and degree of association between the variables".
As *a measure of the degree of relatedness of two variables,* correlation is widely used in exploratory research when the objective is to locate variables that might be related in some way to the variable of interest.

TYPES OF CORRELATION

Correlation can be classified in several ways. The important ways of classifying correlation are:
(i) Positive and negative,
(ii) Linear and non-linear (curvilinear) and
(iii) Simple, partial and multiple.

Positive and Negative Correlation

If both the variables move in the same direction, we say that there is a positive correlation, *i.e.,* if one variable increases, the other variable also increases on an average or if one variable decreases, the other variable also decreases on an average. On the other hand, if the variables are varying in opposite direction, we say that it is a case of negative correlation; *e.g.,* movements of demand and supply.

Linear and Non-linear (Curvilinear) Correlation

If the change in one variable is accompanied by change in another variable in a constant ratio,
X : 10 20 30 40 50
Y : 25 50 75 100 125
On the other hand, if the amount of change in one variable does not follow a constant ratio with the change in another variable, it is a case of non-linear or curvilinear correlation. If a couple of figures in either series *X* or series *Y* are changed, it would give a non-linear correlation.

Simple, Partial and Multiple Correlation

The distinction amongst these three types of correlation depends upon the number of variables involved in a study. If only two variables are involved in a study, then the correlation is said to be simple correlation. When three or more variables are involved in a study, then it is a problem of either partial or multiple correlation. In multiple correlation, three or more variables are studied simultaneously. But in partial correlation we consider only two variables influencing each other while the effect of other variable(s) is held constant. Suppose we have a problem comprising three variables *X, Y* and *Z. X* is the number of hours studied, *Y* is I.Q. and *Z* is the number of marks obtained in the examination. In a multiple correlation, we will study the relationship between the marks obtained *(Z)* and the two variables, number of hours studied *(X)* and I.Q. (*Y*). In contrast, when we study the relationship between *X* and *Z,* keeping an average I.Q. (*Y*) as constant, it is said to be a study involving partial correlation.

CORRELATION DOES NOT NECESSARILY MEAN CAUSATION

The correlation analysis, in discovering the nature and degree of relationship between variables, does not necessarily imply any cause and effect relationship between the variables. Two variables may be related to each other but this does not mean that one variable causes the other. *For example*, we may find that logical reasoning and creativity are correlated, but that does not mean if we could increase peoples' logical reasoning ability, we would produce greater creativity. We need to conduct an actual experiment to unequivocally demonstrate a causal relationship. But if it is true that influencing someones' logical reasoning ability does influence their creativity, then the two variables must be correlated with each other. **In other words, *causation always implies correlation, however converse is not true.***

Let us see some situations-

1. The correlation may be due to chance particularly when the data pertain to a small sample. A small sample bivariate series may show the relationship but such a relationship may not exist in the universe.
2. It is possible that both the variables are influenced by one or more other variables. For example, expenditure on food and entertainment for a given number of households show a positive relationship because both have increased over time. But, this is due to rise in family incomes over the same period. In other words, the two variables have been influenced by another variable - increase in family incomes.

3. There may be another situation where both the variables may be influencing each other so that we cannot say which is the cause and which is the effect. *For example*, take the case of price and demand. The rise in price of a commodity may lead to a decline in the demand for it. Here, price is the cause and the demand is the effect. In yet another situation, an increase in demand may lead to a rise in price. Here, the demand is the cause while price is the effect, which is just the reverse of the earlier situation. In such situations, it is difficult to identify which variable is causing the effect on which variable, as both are influencing each other.

The foregoing discussion clearly shows that correlation does not indicate any causation or functional relationship. ***Correlation coefficient is merely a mathematical relationship and this has nothing to do with cause and effect relation.*** It only reveals co-variation between two variables. Even when there is no cause-and-effect relationship in bivariate series and one interprets the relationship as causal, such a correlation is called ***spurious*** or ***non-sense correlation***. Obviously, this will be misleading. As such, one has to be very careful in correlation exercises and look into other relevant factors before concluding a cause-and-effect relationship.

CORRELATION ANALYSIS

Correlation Analysis is a statistical technique used to indicate the nature and degree of relationship existing between one variable and the other(s). It is also used along with regression analysis to measure how well the regression line explains the variations of the dependent variable with the independent variable. It often becomes necessary to examine how two paired data sets are related.

For example, we may have data on the sales of a product and the expenditure incurred on its advertisement for a specified number of years. Given that sales and advertisement expenditure are related to each other, it is useful to examine the nature of relationship between the two and quantify the degree of that relationship. As this requires use of appropriate statistical methods, these falls under the purview of what we call regression and correlation analysis

The commonly used methods for studying linear relationship between two variables involve both graphic and algebraic methods. Some of the widely used methods include:

1. Scatter Diagram
2. Pearson's Coefficient of Correlation
3. Spearman's Rank Correlation

SCATTER DIAGRAM

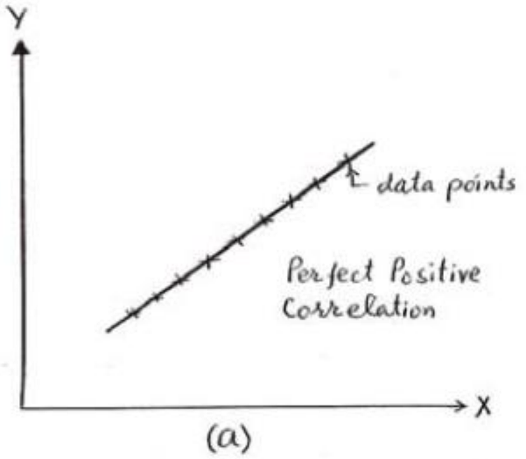
This method is also known as Dotogram or Dot diagram. Scatter diagram is one of the simplest methods of diagrammatic representation of a bivariate distribution.

if the plotted points are very close to each other, it indicates high degree of correlation. If the plotted points are away from each other, it indicates low degree of correlation.

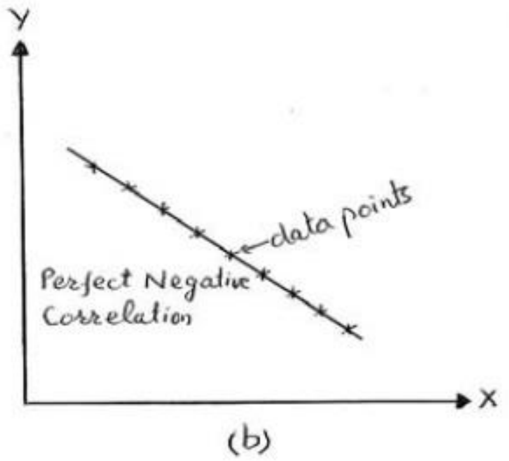
if the points on the diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.

If there is an upward trend rising from lower left hand corner and going upward to the upper right hand corner, the correlation *is positive*since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left hand corner to the lower right hand corner, the correlation *is negative* since in this case the values of the two variables move in the opposite directions.

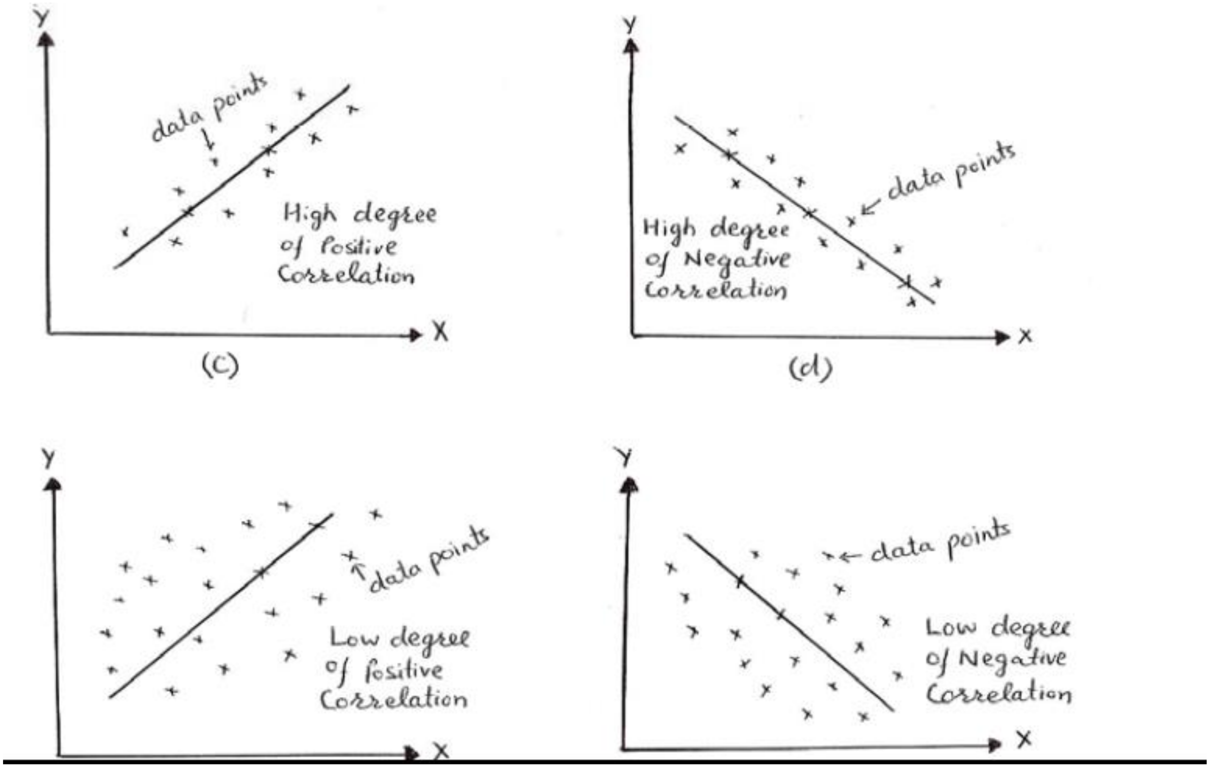
In particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive(fig: a), and if all the points like on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative(fig: b).



(a)



(b)


Y
data points
High degree of Positive Correlation
X
(c)
Y
data points
High degree of Negative Correlation
X
(d)
Y
data points
Low degree of Positive Correlation
X
Y
data points
Low degree of Negative Correlation
X

shows that the two curves move in the same direction and, moreover, they are very close to each other, suggesting a close relationship between price yield per plot (qtls) and quantity of fertilizer used (kg)

PEARSON’S COEFFICIENT OF CORRELATION

A mathematical method for measuring the intensity or the magnitude of *linear relationship* between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice. Karl Pearson’s measure, known as Pearsonian correlation coefficient between two variables *X* and *Y*, usually denoted by *r(X,Y)* or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between *X* and *Y,* to the product of the standard deviations of *X* and *Y*.(formula we have discussed in class)