Correlation

Correlation

Correlation is defined as the degree of relationship between two or more variables. The correlation may be simple, multiple or partial. If only two variables are considered the correlation is called a simple correlation.

Simple Correlation

If the change of the value of one variable resulted in the change of the value of the other variable we say that the variables are correlated. This relation may be of two types. If the increase in the value of one variable result in the increase in the value of the other variable, then we say that the variables are directly correlated or positively correlated. On the other had if increase in one variable results in a decrease of the other variable the correlation is called inverse correlation or negative correlation. One way to get a rough idea about the correlation is the scatter diagram

Scatter Diagram

Let (x_1,y_1), (x_2,y_2) ….(x_n,y_n) be the set of observation on two characteristics. A diagram obtained by plotting points with (x_i,y_i) as co-ordinate in xy-plane is called scatter diagram. It consists of n points scattered over the (x,y) plane. A rough idea of the relationship may be had by an inspection of the scatter diagram. If the points closely cluster round a well defined line a high degree of linear relationship may be inferred. But if the points are widely scattered without clustering round any line or curve lack of relationship between them may be suspected.

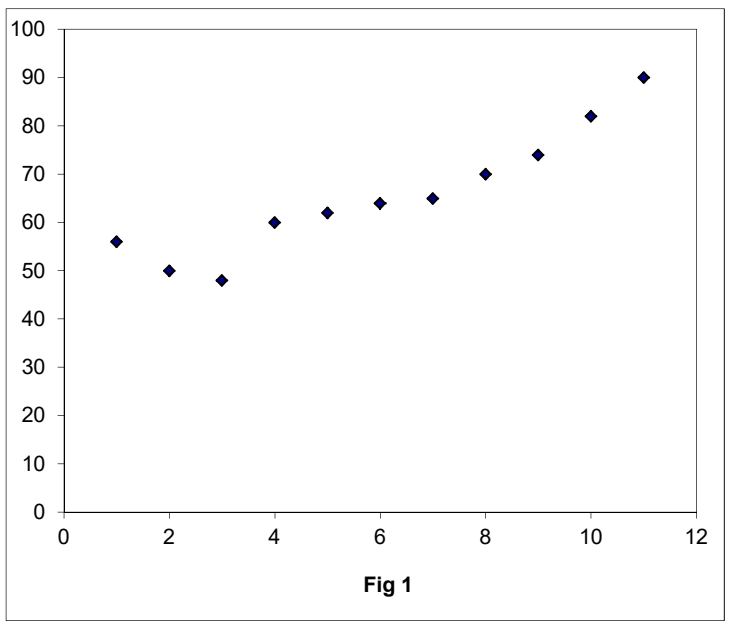

Fig 1

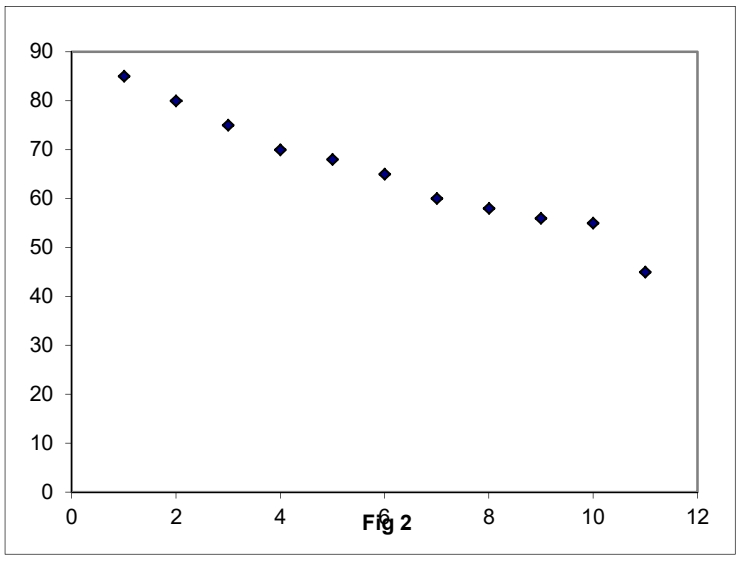

Fig 2

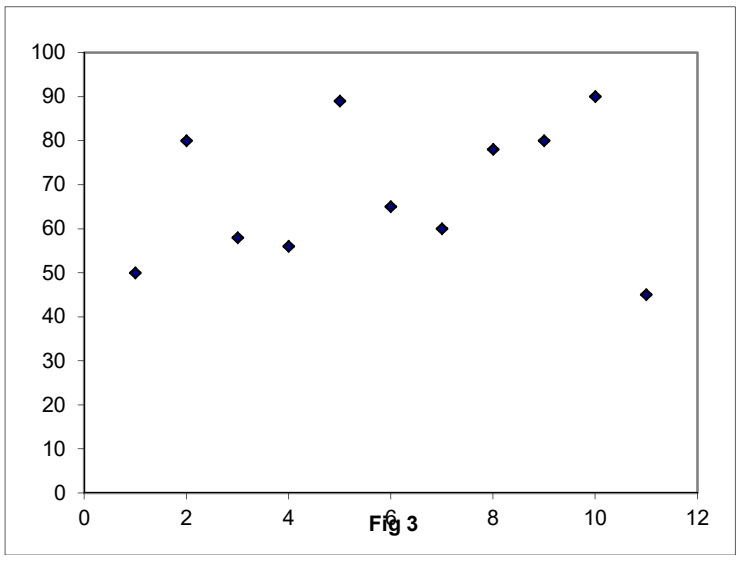

Fig 3

Figure 1 give a scatter diagram in which a positive correlation holds. Figure 2 indicates a perfect negative correlation. Figure 3 is indicative of lack of any relationship between the variables.

Karl Pearson Coefficient of Correlation

Karl Pearson Coefficient of Correlation between x and y is defined as $\frac{Cov(x,y)}{\sigma_x \sigma_y}$. The sample correlation coefficient between x and y is denoted by r_{xy} and is given by

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{p_{xy}}{\sigma_x \sigma_y}$$

where $p_{xy} = \sum (x-\bar{x})(y-\bar{y})$ $\sigma_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$, $\bar{x} = \frac{\sum x}{n}$

$\sigma_y = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}$, $\bar{y} = \frac{\sum y}{n}$.

Note

1. Let r_{xy} be the correlation coefficient between the variable (x,y) and a, b, c and d are real constant. Define u = $\frac{x-a}{c}$ and v = $\frac{y-b}{d}$ and let r_{uv} be the correlation between (u,v). Then $r_{xy} = r_{uv}$.
2. The above result is very useful in calculating the correlation coefficient when the observations are numerically large. We may reduce the magnitude of the observations by a suitable linear transformation and the correlation coefficient calculated from the transformed observations will be the same as that calculated from the original observations.
3. $-1<r_{xy}<1$
4. $r_{xy}=0$ does not means that there is no relation between the variables. It only indicate that there is no linear relation ship between the variables.

Spearman Rank Correlation Coefficient

The Pearson's coefficient of correlation calculated for the ranks (or order) of the observations is called rank correlation coefficient. The main utility of such a coefficient is that it can be evaluated even when the characteristic under consideration is not numerically measurable but can be ranked. The formula to calculate the rank correlation coefficient

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

where $d_i = (x_i - y_i)$ is the difference in rank

Probable Error

Probable Error use to test the validity of the correlation coefficient. The Probable Error of the coefficient of correlation is an amount, which, if added to or subtracted from the mean correlation coefficient, produces amount with in which the chances are even that a coefficient of correlation from a series selected random will fall. The formula for calculating Probable Error is

$$\text{Probable Error} = 0.6745\ \frac{1-r^2}{N}$$

Where r is the coefficient of correlation and n is the number of Pairs

The confidence interval for the population coefficient of correlation are $[r\text{-}PE, r\text{+}PE]$

Functions of probable error

1. If the value of **r** is less than the probable error, the value of **r** is not significant.
2. If the value of **r** is more than six times the probable error(r=6PE), the value of **r** is significant.
3. If the probable error is less than 0.3, the correlation should not be considered at all.
4. If the probable error is small, the correlation definitely existing.

Conditions for the use of Probable Error

1. The number of items should be large enough. When the number of pairs of observation is small, the probable error may lead to fallacious conclusions.
2. The distribution should have a normal distribution. That is, bell shaped curve.
3. The items in the sample must have been selected by random sample method and unbiased manner.
4. The statistical measure for which probable error is computed must have been from a sample.

Tied Ranks

It may be noted that the formula for Spearman's rank correlation coefficient is derived on the assumptions that all the ranks are different. But there are situation in which more than one individual is given the same rank. Fore example in an examination more than one candidate may get the same marks and so equal ranks. In such a situation the convention is to assign the average of the ranks they would have got if the marks were different to all those who have got the same rank. For example if three people get the second rank we assign the rank $\frac{2+3+4}{3} = 3$ to each of them. In such case it is more accurate to calculate the Pearson's coefficient of correlation between the ranks directly after assigning the average rank to those with same rank. The formula is modified to suit such situation as

$$r = 1 - \frac{6\sum d_i^2 + \frac{1}{2}\sum m_i(m_i^2 - 1)}{n(n^2 - 1)}$$

where m_i is the number of times i^{th} rank repeated in x and y.