CORRELATION

1.1: Bivariate data:

So far we have confined our discussion to the distributions involving only one variable. Sometimes, in practical applications, we might come across certain set of data, where each observation of the set may comprise of the values of two or more variables.

Suppose we have a set of 30 students in a class and we want to measure the heights and weights of all the students. We observe that each individual (unit) of the set assumes two values – one relating to the height and the other to the weight. Such a distribution in which each individual or unit of the set is made up of two values is called a bivariate distribution. The following examples will illustrate clearly the meaning of bivariate distribution.

(i) In a class of 60 students the series of marks obtained in two subjects by all of them.

(ii) The series of sales revenue and advertising expenditure of two companies in a particular year.

(iii) The series of ages of husbands and wives in a sample of selected married couples.

Thus in a bivariate distribution, we are given a set of pairs of observations, wherein each pair represents the values of two variables. In a bivariate distribution, we are interested in finding a relationship (if it exists) between the two variables under study.

The concept of 'correlation' is a statistical tool which studies the relationship between two variables and Correlation Analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

Definition:

"Two variables are said to be in correlation if the change in one of the variables results in a change in the other variable".

OR

"Degree of linear relationship between two variables is called correlation"

OR

"Mutual relationship between two variables is called correlation"

Correlation is denoted by r and range of the correlation is -1 to +1

Some examples of correlation are:

(i) Hours spent studying Vs Marks scored by students

(ii) Amount of rainfall Vs Agricultural yield

(iii) Electricity usage Vs Electricity bill

(iv) Suicide rates Vs Number of stressful people

(v) Years of experience Vs Salary

(vi) Demand Vs Product price

(vii) Age Vs Beauty

(viii) Age Vs Health issues

(ix) Number of Degrees Vs Salary

(x) Number of Degrees Vs Education expenditure

1.1.2: Types of Correlation

There are two important types of correlation. They are (1) Positive and Negative correlation and (2) Linear and Non – Linear correlation.

1.1.2 (a): Positive Correlation

If the values of the two variables deviate in the same direction i.e. if an increase (or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.

Some examples of series of positive correlation are:

(i) Heights and weights;

(ii) Household income and expenditure;

(iii) Price and supply of commodities;

(iv) Amount of rainfall and yield of crops.

1.1.2 (b): Negative Correlation

Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if the increase in the variables deviate in opposite direction. That is, if increase (or decrease) in the values of one variable results on an average, in corresponding decrease (or increase) in the values of other variable.

Some examples of series of negative correlation are:

(i) Volume and pressure of perfect gas;

(ii) Price and demand of goods.

Note:

(i) If the points are very close to each other, a fairly good amount of correlation can be expected between the two variables. On the other hand if they are widely scattered a poor correlation can be expected between them.

(ii) If the points are scattered and they reveal no upward or downward trend then we say the variables are uncorrelated.

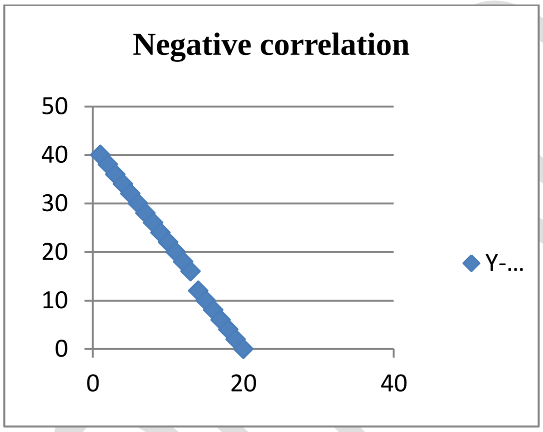
(iii) If there is an upward trend rising from the lower left hand corner and going upward to the upper right hand corner, the correlation obtained from the graph is said to be positive. Also, if there is a downward trend from the upper left hand corner the correlation obtained is said to be negative.

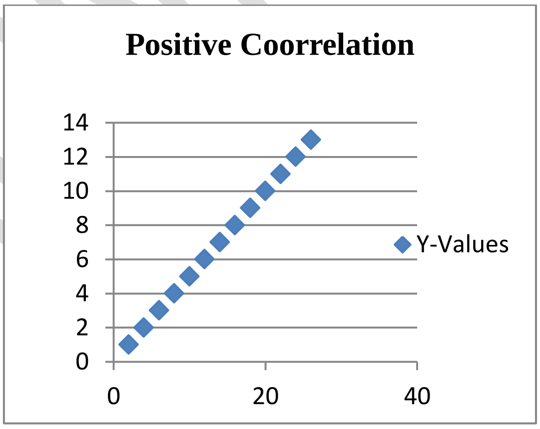
1.1.3 A Scatter diagram:

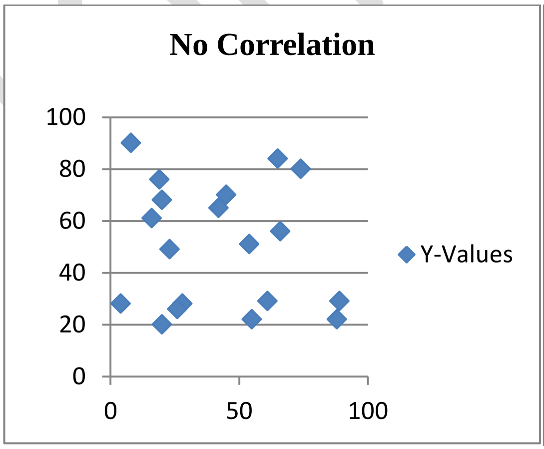
It is the graphical representation of the pairs of data (X; Y) in an orthogonal
coordinate system. The values x of the input variable X are represented on the OX axis, and the values y of the output variable Y are represented on the OY axis.

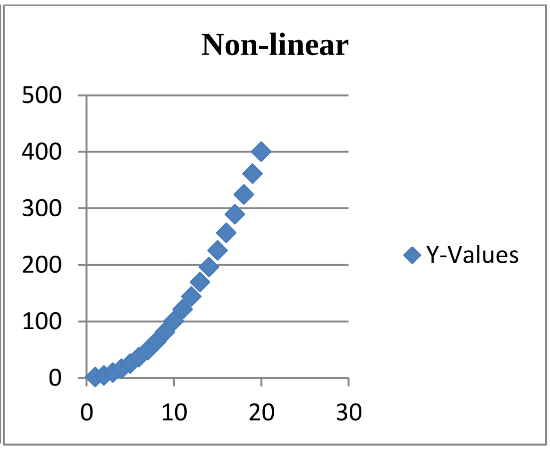
If for the increasing values x of the input variable X there is no definite displacement of the values y of the variable Y, we then say that there is **no Correlation between X and Y**.

If for the increasing values x of the input variable X there is a definite displacement of the values y of the variable Y, we then say that there is a **correlation.** We have a positive correlation if the values y tends to increase and a negative correlation if the values y tends to decrease.









Definition:

Scatter plot is a mathematical diagram to display values of two variables for a set of data. It is used to investigate the possible relationship between two variables.

- **Merits and Demerits of Scatter Diagram:-**
- **Merits:-**

[1] Scatter diagram is the simplest method of studying correlation.

[2] It is easy to understand.

[3] It is not influenced by extreme values.

- **Demerits:-**

[1] It does not give a numerical measure of correlation.

[2] It is a subjective method.

[3] It cannot be applied to qualitative data.

- **Covariance:-**

The shape of the data points is a graphical representation of the spread for two jointly related variables. The numerical measure for this sort of spread is called covariance

Definition: If $\{(x_i, y_i); i = 1,2,3,...,n\}$ are bivariate data on (X, Y), then covariance between X and Y is given by

$$\text{Cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\text{Cov}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy}$$

Remark:-

[1] Cov (X, Y) = Cov (Y, X)

[2] Cov (X, constant) = 0

[3] Covariance may be negative

- **Properties of Covariance:**

[1] $\text{Cov}(x, x) = \text{Var}(x)$

Proof: By definition,

$$\mathrm{Cov}(x,x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{n}$$

$$\mathrm{Cov}(x,x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

$$\mathrm{Cov}(x,x) = \mathrm{Var}(x)$$

[2] Effect of change of origin on covariance. i. e. $\mathrm{Cov}(X-a, Y-b) = \mathrm{Cov}(x,y)$

Proof: By definition of covariance

$$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\mathrm{Cov}(X-a, Y-b) = \frac{\sum_{i=1}^{n}\left[(x_i-a)-(\bar{x}-a)\right]\left[(y_i-b)-(\bar{y}-b)\right]}{n}$$

$$\mathrm{Cov}(X-a, Y-b) = \frac{\sum_{i=1}^{n}\left[(x_i-a-\bar{x}+a)\right]\left[(y_i-b-\bar{y}+b)\right]}{n}$$

$$\mathrm{Cov}(X-a, Y-b) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\mathrm{Cov}(X-a, Y-b) = \mathrm{Cov}(x,y)$$

[3] Effect of change of origin and scale on covariance. i. e.

$\mathrm{Cov}\left(\frac{X-a}{h}, \frac{Y-b}{k}\right) = \frac{1}{hk}\mathrm{Cov}(X,Y)$, **a, b, h and k are constants.**

Proof: By definition of covariance

$$\mathrm{Cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\mathrm{Cov}\left(\frac{X-a}{h}, \frac{Y-b}{k}\right) = \frac{\sum_{i=1}^{n}\left[\left(\frac{x_i-a}{h}\right)-\left(\frac{\bar{x}-a}{h}\right)\right]\left[\left(\frac{y_i-b}{k}\right)-\left(\frac{\bar{y}-b}{k}\right)\right]}{n}$$

$$\mathrm{Cov}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{1}{hk}\frac{\sum_{i=1}^{n}\left[x_i-a-\bar{x}+a\right]\left[y_i-b-\bar{y}+b\right]}{n}$$

$$\mathrm{Cov}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{1}{hk}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

$$\mathrm{Cov}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{1}{hk}\mathrm{Cov}(X,Y)$$

1.2 Karl Pearson's Coefficient of Correlation for ungrouped data

$$\mathrm{Corr}(X, X)=r=\frac{\mathrm{Cov}(X,X)}{\sigma_x\sigma_x}$$

Where,

$$\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}x_iy_i}{n}-\bar{x}\bar{y};\qquad \bar{x}=\frac{\sum_{i=1}^{n}x_i}{n};\qquad \bar{y}=\frac{\sum_{i=1}^{n}x_i}{n}$$

$$\sigma_X=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-[\bar{x}]^2}\qquad ;\sigma_Y=\sqrt{\frac{\sum_{i=1}^{n}y_i^2}{n}-[\bar{y}]^2}$$

$$r=\frac{\frac{\sum_{i=1}^{n}x_iy_i}{n}-\bar{x}\bar{y}}{\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-[\bar{x}]^2}\sqrt{\frac{\sum_{i=1}^{n}y_i^2}{n}-[\bar{y}]^2}}\quad\Rightarrow r=\frac{\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{n}}{\sqrt{\frac{\sum_{i=1}^{n}x_i^2-n[\bar{x}]^2}{n}}\sqrt{\frac{\sum_{i=1}^{n}y_i^2-n[\bar{y}]^2}{n}}}$$

$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n[\bar{x}]^2}\sqrt{\sum_{i=1}^{n}y_i^2-n[\bar{y}]^2}}$$

The Pearson correlation coefficient, r, measures the **strength** and the **direction** of a straight-line relationship.

•The **strength** of the relationship is determined by the closeness of the points to a straight line.

•The **direction** is determined by whether one variable generally increases or generally decreases when the other variable increases.

•**r** is always between –1 and +1

•**Magnitude** indicates the strength

•**r = –1 or +1** indicates a perfect linear relationship

•**sign** indicates the direction

•**r = 0** indicates no linear relationship

➢ **Properties of Correlation:**

[1] Corr(X, X) = 1

Proof: By definition of correlation,

$$\text{Corr}(X, X) = r = \frac{\text{Cov}(X,X)}{\sigma_x \sigma_x}$$

But, $\text{Cov}(X,X) = \text{Var}(X)$

$$\text{Corr}(X, X) = \frac{\text{Var}(X)}{\sigma_x^2}$$

$$\text{Corr}(X, X) = \frac{\sigma_x^2}{\sigma_x^2}$$

$$\text{Corr}(X, X) = 1$$

[2] Effect of change of origin and scale on correlation. i. e.

$$\text{Corr}\left(\frac{X-a}{h}, \frac{Y-b}{k}\right) = \frac{1}{hk}\text{Corr}(X,Y)$$ **. a, b, h and k are constants.**

Proof: By definition of covariance and standard deviation

$$\text{Cov}\left(\frac{X-a}{h}, \frac{Y-b}{k}\right) = \frac{1}{hk}\text{Cov}(X,Y)$$

$$\sigma_{\left(\frac{x-a}{h}\right)} = \frac{1}{|h|}\sigma_x \quad \text{and} \quad \sigma_{\left(\frac{y-b}{k}\right)} = \frac{1}{|k|}\sigma_y$$

Now,

$$\mathrm{Corr}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{\frac{1}{hk}\mathrm{Cov}(X,Y)}{\frac{1}{|h|}\sigma_x\frac{1}{|k|}\sigma_y}$$

$$\mathrm{Corr}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{\frac{1}{hk}\mathrm{Cov}(X,Y)}{\frac{1}{|hk|}\sigma_x\sigma_y}$$

If h and k have same algebraic signs, then

$$\mathrm{Corr}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}=\mathrm{Corr}(x,y)$$

If h and k have opposite algebraic signs, then

$$\mathrm{Corr}\left(\frac{X-a}{h},\frac{Y-b}{k}\right)=\frac{-\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}=-\mathrm{Corr}(x,y)$$

[3] Correlation coefficient r always lies between -1 and 1 .i. e.

$-1\le \mathrm{Corr}(X,Y)\le 1$.

Proof: With usual notation

$$\left[\left(\frac{x_i-\bar{x}}{\sigma_x}\right)-\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\right]^2\ge 0$$

$$\frac{\sum_{i=1}^{n}\left[\left(\frac{x_i-\bar{x}}{\sigma_x}\right)-\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\right]^2}{n}\ge 0$$

$$\frac{1}{n}\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma_x^2}+\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma_x^2}-2\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y}\right]\ge 0$$

$$\left[\frac{\sigma_x^2}{\sigma_x^2}+\frac{\sigma_y^2}{\sigma_x^2}-2\frac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}\right]\ge 0$$

$$\left[1+1-2\mathrm{Corr}(X,Y)\right]\ge 0 \qquad \left(\because \mathrm{Corr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sigma_x\sigma_y}\right)$$

$$\left[1+1\text{-}2\text{Corr}\left(X,Y\right)\right]\geq 0$$

$$1\text{-}\text{Corr}\left(X,Y\right)\geq 0$$

$$1\geq \text{Corr}\left(X,Y\right)$$

$$\text{Corr}\left(X,Y\right)\leq 1 \qquad \ldots(\text{i})$$

Similarly,

$$\left[\left(\frac{x_i\text{-}\bar{x}}{\sigma_x}\right)+\left(\frac{y_i\text{-}\bar{y}}{\sigma_y}\right)\right]^2\geq 0$$

$$\frac{\sum_{i=1}^{n}\left[\left(\frac{x_i\text{-}\bar{x}}{\sigma_x}\right)+\left(\frac{y_i\text{-}\bar{y}}{\sigma_y}\right)\right]^2}{n}\geq 0$$

$$\frac{1}{n}\left[\frac{\sum_{i=1}^{n}\left(x_i\text{-}\bar{x}\right)^2}{\sigma_x^2}+\frac{\sum_{i=1}^{n}\left(y_i\text{-}\bar{y}\right)^2}{\sigma_x^2}+2\frac{\sum_{i=1}^{n}\left(x_i\text{-}\bar{x}\right)\left(y_i\text{-}\bar{y}\right)}{\sigma_x\sigma_y}\right]\geq 0$$

$$\left[\frac{\sigma_x^2}{\sigma_x^2}+\frac{\sigma_y^2}{\sigma_x^2}+2\frac{\text{Cov}\left(X,Y\right)}{\sigma_x\sigma_y}\right]\geq 0$$

$$\left[1+1+2\text{Corr}\left(X,Y\right)\right]\geq 0 \qquad \left(\because \text{Corr}\left(X,Y\right)=\frac{\text{Cov}\left(X,Y\right)}{\sigma_x\sigma_y}\right)$$

$$\left[1+1+2\text{Corr}\left(X,Y\right)\right]\geq 0$$

$$1+\text{Corr}\left(X,Y\right)\geq 0$$

$$\text{Corr}\left(X,Y\right)\geq -1$$

$$-1\leq \text{Corr}\left(X,Y\right) \qquad \ldots(\text{ii})$$

From (i) and (ii), we can write

$$-1\leq \text{Corr}\left(X,Y\right)\leq 1$$

1.3: Spearman’s Rank Correlation Coefficient

Data which are arranged in numerical order, usually from largest to smallest and numbered 1,2,3 ---- are said to be in **ranks** or **ranked data.**

These ranks prove useful at certain times when two or more values of one variable are the same. The coefficient of correlation for such type of data is given by **Spearman rank difference correlation coefficient** and is denoted by R.

In order to calculate R, we arrange data in ranks computing the difference in rank 'd' for each pair. The following example will explain the usefulness of R. R is given by the formula

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

Derivation

If $\{(x_i, y_i); i = 1,2,3,...,n\}$ are bivariate data on (X, Y) and are permutations of 1,2,3,…,n. X and Y have same arithmetic mean and standard deviations.

$$\bar{x}= \bar{y} = \frac{1+2+3+...+n}{n}$$

$$\bar{x}= \bar{y} = \frac{\text{Sum of first n natural numbers}}{n}$$

$$\bar{x}= \bar{y} = \frac{n(n+1)}{2n}$$

$$\bar{x}= \bar{y} = \frac{(n+1)}{2}$$

$$\sigma_x^2 = \sigma_y^2 = \frac{1^2+2^2+3^2+...+n^2}{n} - \left[\frac{(n+1)}{2}\right]^2$$

$$\sigma_x^2 = \sigma_y^2 = \frac{\text{Sum of squares of first n natural numbers}}{n} - \left[\frac{(n+1)}{2}\right]^2$$

$$\sigma_x^2 = \sigma_y^2 = \frac{n(n+1)(2n+1)}{6n} - \left[\frac{(n+1)}{2}\right]^2$$

$$\sigma_x^2 = \sigma_y^2 = \frac{(n+1)}{2}\left[\frac{(2n+1)}{3} - \frac{(n+1)}{2}\right]$$

$$\sigma_x^2 = \sigma_y^2 = \frac{(n+1)}{2}\left[\frac{(4n+2-3n-3)}{6}\right]$$

$$\sigma_x^2 = \sigma_y^2 = \frac{(n+1)}{2}\left[\frac{(n-1)}{6}\right]$$

$$\sigma_x^2 = \sigma_y^2 = \frac{(n^2-1)}{12}$$

$$\text{Cov}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy}$$

$$\text{Cov}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \left(\frac{n+1}{2}\right)^2 \qquad \ldots(i)$$

Let

$$d_i = x_i - y_i$$

$$d_i^2 = (x_i - y_i)^2 = x_i^2 + y_i^2 - 2x_i y_i$$

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 + 2\sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} d_i^2 = \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)(2n+1)}{6} + 2\sum_{i=1}^{n} x_i y_i$$

$$2\sum_{i=1}^{n} x_i y_i = \frac{2n(n+1)(2n+1)}{6} - \sum_{i=1}^{n} d_i^2$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

Put this value in equation (i), we get

$$\text{Cov}(x,y) = \frac{(n+1)(2n+1)}{6} - \frac{1}{2n}\sum_{i=1}^{n} d_i^2 - \left(\frac{n+1}{2}\right)^2$$

$$\text{Cov}(x,y) = \left(\frac{n+1}{2}\right)\left[\frac{(2n+1)}{3} - \left(\frac{n+1}{2}\right)\right] - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

$$\text{Cov}(x,y) = \left(\frac{n+1}{2}\right)\left[\left(\frac{n-1}{6}\right)\right] - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

$$\text{Cov}(x,y) = \left(\frac{n^2-1}{12}\right) - \frac{1}{2n}\sum_{i=1}^{n} d_i^2$$

$$\text{Cov}(x,y) = \left(\frac{1}{12n}\right)\left[n\left(n^2-1\right) - 6\sum_{i=1}^{n} d_i^2\right]$$

The rank correlation R is given by

$$R = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

$$R = \frac{\left(\frac{1}{12n}\right)\left[n\left(n^2-1\right) - 6\sum_{i=1}^{n} d_i^2\right]}{\sqrt{\frac{(n^2-1)}{12}}\sqrt{\frac{(n^2-1)}{12}}}$$

$$R = \left(\frac{1}{12n}\right)\left[n\left(n^2-1\right) - 6\sum_{i=1}^{n} d_i^2\right] \times \frac{12}{(n^2-1)}$$

$$R = \left[\frac{n\left(n^2-1\right)}{n\left(n^2-1\right)} - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}\right]$$

$$R = \left[1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}\right]$$

1.4: Spearman's Rank Correlation Coefficient with ties

Tie: Tie is said to occur in ranking if two or more items have same merit**.**

$$R = 1 - \frac{\left[6\sum d_i^2 + T_x + T_y\right]}{n(n^2-1)}$$

Where,

$$T_x = \frac{\sum m_i(m_i^2-1)}{12} = \text{sum of correction due to ties in x}$$

$$T_y = \frac{\sum m_i(m_i^2-1)}{12} = \text{sum of correction due to ties in y}$$

NUMERICAL EXAMPLES:

Example 1 : Compute the correlation coefficient between x and y from the following data n =10 ,$\sum xy = 220$,$\sum x^2 = 200$, $\sum y^2 = 262$,$\sum x = 40$ and $\sum y = 50$

Solution:

From the given data we have formula of Correlation Coefficient,

$$r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$r = \frac{10\times220-40\times50}{\sqrt{10\times200-(40)^2}\sqrt{10\times262-(50)^2}}$$

$$r = \frac{2200-2000}{\sqrt{2000-1600}\sqrt{2620-2500}}$$

$$r = \frac{200}{20\times10.9545}$$

$$r = 0.91$$

Interpretation: Thus there is a good amount of positive correlation between the two variables X and Y.

Alternately, as given,

$$\bar{x} = \frac{\sum x}{n} \qquad \Rightarrow \bar{x} = \frac{40}{10} \qquad \Rightarrow \bar{x} = 4$$

$$\bar{y} = \frac{\sum y}{n} \qquad \Rightarrow \bar{y} = \frac{50}{10} \qquad \Rightarrow \bar{y} = 5$$

$$\mathrm{Cov}(x,y) = \frac{\sum xy}{n} - \bar{x}\bar{y} \qquad \Rightarrow \mathrm{Cov}(x,y) = \frac{220}{10} - 4.5 \qquad \Rightarrow \mathrm{Cov}(x,y) = 2$$

$$\sigma_x = \sqrt{\frac{\sum x_i^2}{n} - (\bar{x})^2} \qquad \Rightarrow \sigma_x = \sqrt{\frac{200}{10} - (4)^2} \qquad \sigma_x = 2$$

$$\sigma_y = \sqrt{\frac{\sum y_i^2}{n} - (\bar{y})^2} \qquad \Rightarrow \sigma_y = \sqrt{\frac{262}{10} - (5)^2}$$

$$\sigma_y = \sqrt{26.20 - 25} \qquad \sigma_y = 1.0954$$

Thus applying formula of Correlation Coefficient, we get

$$r = \frac{\mathrm{Cov}(x,y)}{\sigma_x . \sigma_y} \qquad \Rightarrow r = \frac{2}{2 \times 1.0954} \qquad \Rightarrow r = 0.91$$

As before, we draw the same conclusion.

Example 2: Find product moment correlation coefficient from the following information:

X	2	3	5	5	6	8
Y	9	8	8	6	5	3

Solution:

In order to find the covariance and the two standard deviations, we prepare the following table:

Computation of Correlation Coefficient

x_i	y_i	x_iy_i	x_i^2	y_i^2
2	9	18	4	81
3	8	24	9	64
5	8	40	25	64
5	6	30	25	36
6	5	30	36	25
8	3	24	64	9
29	39	166	163	279

In some cases, there may be some confusion about selecting the pair of variables for which correlation is wanted. This is explained in the following problem.

Example 3: Examine whether there is any correlation between age and blindness on the basis of the following data:

Age in years	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons (in thousands)	90	120	140	100	80	60	40	20
No. of blind persons	10	15	18	20	15	12	10	06

Solution:

Let us denote the mid-value of age in years as x and the number of blind persons per lakh as y .Then as before, we compute correlation coefficient between X and Y.

Computation of correlation between age and blindness

Age in years	Mid-value	No. of persons	No. of blind	No. of blind per lakh	xy	x^2	y^2

	x	('000)P	B	Y=B/P×1 lakh			
0-10	5	90	10	11	55	25	121
10-20	15	120	15	12	180	225	144
20-30	25	140	18	13	325	625	169
30-40	35	100	20	20	700	1225	400
40-50	45	80	15	19	855	2025	361
50-60	55	60	12	20	1100	3025	400
60-70	65	40	10	25	1625	4225	625
70-80	75	20	6	30	2250	5625	900
Total	320	-	-	150	7090	17000	3120

The correlation coefficient between age and blindness is given by

$$r = \frac{n\sum xy - \sum x.\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

$$r = \frac{8\times7090 - 320\times150}{\sqrt{8\times17000 - (320)^2} \times \sqrt{8\times3120 - (150)^2}}$$

$$r = \frac{8720}{183.3030\times49.5984}$$

$$r = \frac{8720}{9091.5355}$$

$$r = 0.96$$

Which exhibits a very high degree of positive correlation between age and blindness.

Example 4 : Coefficient of correlation between x and y for 20 items is 0.4 . The AM's and SD's of x and y are know to be 12 and 15 and 3 and 4 respectively. Later on, it was found that the pair (20,15) was wrongly taken as (15,20) . Find the correct value of the correlation coefficient.

Solution:

We are given that n = 20 and the original

$\bar{x}=12, \bar{y}=15, \sigma_x=3, \sigma_y=4$ and $r = 0.4$,

$r=\frac{cov(x,y)}{\sigma_x \times \sigma_y}=$ $\Rightarrow 0.4=\frac{cov(x,y)}{3\times 4}$

$cov(x,y)= 4.8$ $\Rightarrow 4.8 =\frac{\sum xy}{n}-\bar{x}\bar{y}$

$4.8 =\frac{\sum xy}{20}-12\times 15$ $\Rightarrow \sum xy =3696$

Hence, corrected $\sum xy =3696-20\times 15+15\times 20=3696$

$\sigma_x{}^2=9$ $\Rightarrow \sigma_x{}^2=\left(\frac{\sum x_i^2}{20}\right)-12^2$

$9=\left(\frac{\sum x_i^2}{20}\right)-144$ $\Rightarrow \left(\frac{\sum x_i^2}{20}\right)=153$

$\sum x_i^2=3060$

Similarly,

$\sigma_y{}^2=16$ $\Rightarrow \sigma_y{}^2=\left(\frac{\sum y_i^2}{20}\right)-15^2$

$16=\left(\frac{\sum x_i^2}{20}\right)-225$ $\Rightarrow \left(\frac{\sum x_i^2}{20}\right)=241$

$\sum y_i^2=4820$

Thus corrected $\sum x = n\bar{x}$ -wrong x value + correct x value

Thus corrected $\sum x = 20\times 12-15+20$

Thus corrected $\sum x = 245$

Similary, corrected $\sum y = 20 \times 15 - 20 + 15$

Corrected $\sum x^2 = 3060 - 15^2 + 20^2 = 3235$

Corrected $\sum y^2 = 4820 - 20^2 + 15^2 = 4645$

Thus corrected value of the correlaction coefficent is

$$r = \frac{20 \times 3696 - 245 \times 295}{\sqrt{20 \times 3235 - (245)^2} \times \sqrt{20 \times 4645 - (295)^2}}$$

$$r = \frac{73920 - 72275}{68.3740 \times 76.6480}$$

$r = 0.31$

Example 5: Compute the coefficient of correlation between marks in statistics and mathematics for the bivariate frequency distribution

Solution:

For the sake of computational advantage, we effect change of origin and scale for the both variables X and Y.

Define $u_i = \frac{x_i - a}{b} = \frac{x_i - 10}{4}$ and $v_i = \frac{y_i - c}{d} = \frac{y_i - 10}{4}$

Where x_i and y_i denote respectively the mid-value of the X-class interval and y-class interval respectively. The following table shows the necessary calculation on the right top coroner each of

Cell, the [product of the cell frequency, corresponding u value and the respective v value has been

Shown. They add up in a particular row or column to provide the value of f_{ij} u_i v_j for that particular row or column.

Computation of correlation coefficient between marks of mathematics and statistics

A single formula for computing correlation coefficient from bivariate frequency distribution is given by

$$r=\frac{N\sum_{i,j} f_{ij}u_iv_j-\sum f_{io}u_i\times\sum f_{oj}v_i}{\sqrt{N\sum f_{io}u_i^2}-(\sum f_{io}u_i)^2\times\sum f_{oj}v_j^2-(\sum f_{oj}v_j)^2}\text{........(3.10)}$$

$$r=\frac{50\times44-8\times20}{\sqrt{50\times76-8^2}\sqrt{50\times74-20^2}}$$

$$r=\frac{2040}{61.1228\times57.4456}$$

$$r=0.58$$

The value of r shown a good amount of positive correction between the marks in statistics and mathematics on the basis of give data.

Since n=8 and $\sum d_i^2 = 4$, applying formula, we get.

$$R=1-\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \Rightarrow R=1-\frac{6\times4}{8(8^2-1)}$$

$$R=1-0.0476 \qquad \Rightarrow R=0.95$$

The high positive value of the rank correlation coefficient indicates that there is a very good amount of agreement between sales and advertisement.

Example 6: Compute rank correlation from the following data relating to ranks given by two judges in a contest:

Serial No. of Candidate:	1	2	3	4	5	6	7	8	9	10
Rank by judge B	10	5	6	1	2	3	4	7	9	8
Rank by judge B	5	6	9	2	8	7	3	4	10	1

Solution:

We directly apply rank correlation coefficient formula as ranks are already given.

Serial No.	**Rank by A(x_i)**	**Rank by B(y_i)**	**$d_i = x_i - y_i$**	**d_i^2**
1	10	5	5	25
2	5	6	-1	1
3	6	9	-3	9
4	1	2	-1	1
5	2	8	-6	36
6	3	7	-4	16
7	4	3	1	1
8	7	4	3	9
9	8	10	-2	4
10	9	1	8	64
Total	-	-	0	166

The rank correlation coefficient is given by

$$R=1-\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \Rightarrow R=1-\frac{6\times 166}{10(10^2-1)}$$

R= -0.006

The very low value (almost 0) indicates that there is hardly any agreement between the ranks given by the two judges in the contest.

Example 7: Compute the coefficient of rank correlation between Economics mark and statistics marks as given below:

Economics Marks	80	56	50	48	50	62	60
Statistics Marks	90	75	75	65	65	50	65

Solution: This is a case of tied ranks as more than one student share the same mark both for Economics and statistics. For Economics The student receiving 80 marks gets rank 1 one getting 62 marks receives rank 2, the student with 60 receives rank 3, student with 56 marks gets rank 4 and since there are two student, each getting 50 marks ,each would be receiving a common rank , the average of the next two ranks 5 and 6 i.e. $\frac{5+6}{2}$ i.e. 5.50 and lastly the last rank……

7 go to the student getting the lowest Economics Marks. In the Computation of rank correlation between Eco marks and stats marks with tied marks

Eco Mark (x_i)	**Stats Mark (y_i)**	**Rank for Eco (x_i)**	**Rank for Stats (y_i)**	**$d_i = x_i - y_i$**	**d_i^2**
80	90	1	1	0	0
56	75	4	2.50	1.50	2.25
50	75	5.50	2.50	3	9
48	65	7	5	2	4
50	65	5.50	5	0.50	0.25
62	50	2	7	-5	25
60	65	3	5	-2	4
Total	-	-	-	0	44.50

$$\frac{\sum\left(t_j^3 - t_j\right)}{12} = \frac{\left(2^3 - 2\right) + \left(2^3 - 2\right) + \left(3^3 - 3\right)}{12} = 3$$

$$\text{R} = 1 - \frac{6\left[\sum_i d_i^2 + \sum_j \frac{\left(t_j^3 - t_j\right)}{12}\right]}{n\left(n^2 - 1\right)}$$

$$\text{R} = 1 - \frac{6 \times \left(44.50 + 3\right)}{7\left(7^2 - 1\right)}$$

$$\text{R} = 0.15$$

[A] Write the correct answer from the following.

[1] ______is concerned with the measurement of the 'strength of associations' between variables

(a) correlation (b) regression

(c) both (d) none

Answer: (a) correlation

[2] When high value of one variable are associated with the high values of the other and low value of one variable are associated with the low values of another then they are said to be

(a) positively correlated
(b) directly correlated
(c) both
(d) none

Answer: (c) both

[3] If high value of one tend to low values of the other, they are said to be

(a) negatively correlated
(b) inversely correlated
(c) both
(d) none

Answer: (c) both

[4] Correlation coefficient between to variable is a measure of the linear relationship

(a) true
(b) false
(c) both
(d) none

Answer: (a) true

[5] Correlation coefficient are depend of the choice of both origin and the scale of observations.

(a) true
(b) false
(c) both
(d) none

Answer: (b) false

[6] Correlation coefficient is a pure number.

(a) True
(b) false
(c) both
(d) none

Answer: (a) True

[7] Correlation coefficient is -__________ of the units measurement

(a) dependent
(b) independent
(c) both
(d) none

Answer: (b) independent

[8] The value of correlations coefficient lies between

(a) -1 and +1 (b) -1and 0

(c) 0 and 1 (d) none

Answer: (a) -1 and +1

[9] Correlation coefficient can be found out by

(a) scattar diagram (b) rank method

(c) both (a) and (b) (d) none

Answer: (c) both

[10] Covariance measures ________variations of two variables

(a) joint (b) single

(c) both (a) and (b) (d) none

Answer: (b) single

[11] In calculating the Karl Pearson's coefficient of correlation it is necessary that the data should be of numerical measurements

(a) valid (b) not valid

(c) both (d) none

Answer: (a) valid

[12] Rank correlation coefficient lies between

(a) 0 to 1 (b) -1to +1 inclusive of this value

(c) -1to 0 (d) both

Answer: (a) 0 to 1

[13] A coefficient near +1 indicates tendency for the larger values of variable to be associated with the larger of the other

(a) true (b) false

(c) both (d)none

Answer: (b) false

[14] In rank correlation coefficient the association need not be linear

(a) true (b) false

(c) both (d) none

Answer: (a) true

[15] In rank correlation coefficient only an increasing/decreasing relation is required

(a) false (b) true

(c) both (d) none

Answer: (a) false

[16] Great advantage of…………. is that it can be used to rank attributes which cannot be express by way of numerical value.

(a) Concurrent correlation (b) Regression

(c) Rank correlation (d) none

Answer: (c) Rank correlation

[17] The sum of difference of rank is

(a) 1 (b) -1

(c) 0 (d) none

Answer: (c) 0

[18] Karl Pearson's coefficient is defined from

(a) Ungrouped data (b) group data

(c) both (d) none

Answer: (b) group data

[19] Correlation methods are used to study the relationship between two time series of data which are recorded annually, monthly, weekly, daily and so on

(a) true (b) false

(c) both (d) none

Answer: (a) true

[20] Age of applicants for life insurance and the premium of insurance then the correlation is

(a) positive (b) negative

(c) zero (d) none

Answer: (a) positive

[21] “Unemployment index and the purchasing power of the common man” then the correlations is

(a) positive (b) negative

(c) zero (d) none

Answer: (b) negative

[22] Production of pig iron and soot content in Durgapur- correlations are

(a) positive (b) negative

(c) zero (d) none

Answer: (a) positive

[23] Demand for goods and their prices under normal times____ correlations is

(a) positive (b) negative

(c) zero (d) none

Answer: (b) negative

[24]_____is relative measure of associations between two or more variables.

(a) Coefficient of correlation (b) Coefficient of regression

(c) both (d) none

Answer: (a) Coefficient of correlation

[25] The lines of regression pass through the points bearing_____no. of points on both sides

(a) equal (b) unequal

(c) zero (d) none

Answer: (d) none

[26] The square of coefficient of correlation ‘r’ is called the coefficient of

(a) determination (b) regressionn

(c) both (d) none

Answer: (a) determination

[27] A relationship $r^2=1-\frac{5}{3}$ is not possible.

(a) true (b) false (c) both (d) none

Answer: (a) true

[28] Whatever may be the value of r positive or negative, its square will be

(a) negative only (b) positive only

(c) zero only (d) none only

Answer: (b) positive only

[29] Simple correlation is called

(a) linear correlation (b) non-linear correlation

(c) both (d) none

Answer: (a) linear correlation

[30] A scatter diagram indicates the type of correlation between two variables

(a) true (b) false (c) both (d) none

Answer: (a) true

[31] If the pattern of points or dots on the scatter diagram show a linear path diagonally accross the graph paper trom the bottom left- hand corner to the top right, correlation will be,

(a) negative (b) zero (c) positive (d) none

Answer: (c) positive

[32] Variance may be positive, negative or zero

(a) true (b) false (c) both (d) none

Answer: (b) false

[33] Covariance may be positive, negative or zero.

(a) true (b) false (c) both (d) none

Answer: (a) true

[34] Correlation coefficient between x and y is equal to correlation coefficient betveen u and v.

(a) true (b) false (c) both (d) none

Answer: (a) true

[35] In case 'The ages of husbands and wives' --------type of correlation.

(a) positive (b) negative (c) zero (d) none

Answer: (a) positive

[36] In case "Shoe size and intelligence"

(a) positive correlation (b) negative correlation

(c) no correlation (d) none

Answer: (c) no correlation

[37] In case Insurance companies profits and the number of claims they have to pay.

(a) positive correlation (b) negative correlation

(c) no correlation (d) none

Answer: (b) negative correlation

[38] In case 'Years of education and income'

(a) positive correlation (b) negative correlation

(c) no correlation (d) none

Answer: (a) positive correlation

[39] Correlation between 'Amount of rainfall and yield of crop' is

(a) positive correlation (b) negative correlation

(c) no correlation (d) none

Answer: (a) positive correlation

[40] For calculation of correlatom coefficient, change of origin is

(a) not possable (b) possshle

(c) both (d) none

Answer: (b) possshle

[41] The relation, $r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$, is

(a) true (b) false

(c) both (d) none

Answer: (a) true

[42] A small value of r indicates only a ------linear type of relationship between the variables.

(a) good (b) poor

(c) maximum (d) highest

Answer: (b) poor

[43] In case of employed person's age and income ' correlation is

(a) positive (b) negative (c) zero (d) none

Answer: (a) positive

[44] In case 'speed of an automobile and the distance required to stop the car often applying brakes then the correlation is

(a) positive (b)negative (c) zero (d) none

Answer: (a) positive

[45] In case 'sale of woolen garments and day temperature' correlation is

(a) positive (b) negative

(c) zero (d) none

Answer: (b) negative

[46] In case sale of cold drink and day temperature correlation is

(a) positive (b) negative

(c) zero (d) none

Answer: (a) positive

[47] In case of ‘Production and price per unit’ then the correlation is

(a) positive (b) negative

(c) zero (d) none

Answer: (b) negative

[48] If slopes at two regression line are equal then r is equal to

(a) 1 (b) + or -1 (c) 0 (d) none

Answer: (b) + or -1

[49] Covariance measures the joint variations of two variables

(a) true (b) false (c) both (d) none

Answer: (a) true

[50] The minimum value of correlation coefficient is

(a) 0 (b)-2 (c) 1 (d) -1

Answer: (d) -1

[51] The maximum value of correlation coefficient is

(a) 0 (b) 2 (c) 1 (d) -1

Answer: (c) 1

[52] In method of concurrent deviations, only the directions of change (positive direction/negative directions) in the variables are taken into account for calculation of

(a) coefficient of S.D (b) coefficient of regression

(c) coefficient of correlation (d) none

Answer: (c) coefficient of correlation

[53] Bivariate Data are the data collected for

(a) Two variables

(b) More than two variables

(c) Two variables at the same point of time

(d) Two variables at different points of time.

Answer: (c) Two variables at the same point of time

[54] For a bivariate frequency table having (p + q) classification the total number of cells is

(a) P (b) p + q (c) Q (d) p q

Answer: (d) p q

[55] Some of the cell frequencies in a bivariate frequency table may be

(a) Negative (b) zero

(c) (a) or (b) (d) None of these

Answer: (b) zero

[56] For a p x q bivariate frequency table, the maximum number of marginal distribution is

(a) p (b) p + q (c)1 (d) 2

Answer: (d) 2

[57] For a p x q classification of bivariate data, the maximum number of conditional distributions is

(a) p (b) p + q (c) pq (d) p or q

Answer: (b) p + q

[58] Correlation analysis aims at

(a) Predicting one variable for a given value of the other variable

(b) Establishing relation between two variables

(c) Measuring the extent of relation between two variables

(d) Both [b] and [c]

Answer: (d) Both [b] and [c]

[59] What is spurious correlation?

[a] It is a bad relation between two variables.

[b] It is very low correlation between two variables.

[c] It is the correlation between two variables having no causal relation.

[d] It is a negative correlation.

Answer: [c]

[60] Scatter diagram is considered for measuring

[a] Linear relationship between two variables

[b] Curvilinear relationship between two variables

[c] Neither (a) nor (b)

[d] Both (a) and (b)

Answer: [d]

[61] If the plotted points in a scatter diagram lie from upper left to lower right, then the correlation is

(a) Positive (b) Zero (c) Negative (d) none of these

Answer: [c]

[62] If all the plotted in a scatter diagram are evenly distributed , then the correlation is

(a) Zero (b) negative

(c) Positive (d) (a) or (b).

Answer:[a]

[63] If all the plotted point in a scatter diagram lie on a single line, then the correlation is

(a) Perfect positive (b) Perfect negative

(c) Both (a) and (b) (d) Either (a) or (b) .

Answer:[d]

[64] The correlation between shoe-size and intelligence is

(a) zero (b) Positive

(c) Negative (d) None of these

Answer:[a]

[65] The correlation between the speed of an automobile and the distance travelled by it after applying the brakes is

(a) Negative (b) zero

(c) Positive (d) None of these

Answer:[a]

[66] Scatter diagram helps us to

(a) Find the nature correlation between two variables

(b) Compute the extent of correlation between two variables

(c) Obtain the mathematical relationship between two variables

(d) Both (a) and (c)

Answer:[a]

[67] Pearson's correlation coefficient is used for finding

(a) Correlation for any type of relation

(b) correlation for linear relation any

(c) Correlation for curvilinear relation only

(d) Both (b) and (c)

Answer: [b]

[68] Product moment correlation coefficient is considered for

(a) Finding the nature of correlation

(b) Finding the amount of correlation

(c) Both (a) and (b)

(d) Either (a) and (b)

Answer: (c) Both (a) and (b)

[69] If the value of correlation coefficient is positive, then the points in a scatter diagram tend to cluster

(a) From lower left corner to upper right corner

(b) From lower left corner to lower right corner

(c) From lower right corner to upper left corner

(d) From lower right corner to upper right

Answer: (a) From lower left corner to upper right corner

[70] When r =1, all the point in a scatter diagram would lie

(a) On a straight line directed from lower left to upper right

(b) On a straight line directed from upper left to lower right

(c) On a straight line

(d) Both (a) and (b)

Answer: (a) On a straight line directed from lower left to upper right

[71] Product moment correlation coefficient may be defined as the ration of

(a) The product of standard deviations of the two variables to the covariance between them

(b) The covariance between the variables to the product of the variance of them

(c) The covariance between the variables to the product of their standard deviations

(d) Either (b) or (c)

Answer: (c)

[72] The covariance between two variables is

(a) Strictly positive (b) Strictly negative

(c) Always 0 (d) Either positive or negative or zero

Answer: (d) Either positive or negative or zero

[73] The coefficient of correlation between two variables

(a) can have any unit .

(b) Is expressed as the product of units of the two variables

(c) Is a unit free measure

(d) None of these

Answer: (c) Is a unit free measure

[74] What are the limits of the correlation coefficient?

(a) No

(b) -1 and 1

(c) 0 and 1, including the limits

(d) -1 and 1, including the limits

Answer: (d) -1 and 1, including the limits

[75] For finding correlation between two attributes, we consider

(a) Pearson’s correlation coefficient

(b) Scatter diagram

(c) Spearman’s rank correlation Coefficient

(d) Coefficient of concurrent deviations.

Answer: (c) Spearman’s rank correlation Coefficient

[76] For finding the degree of agreement about beauty between two Judges in a Beauty contest, we use

(a) Scatter diagram

(b) Coefficient of rank correlation

(c) Coefficient of correlation

(d) Coefficient of concurrent deviation.

Answer: (b) Coefficient of rank correlation

[77] If there is a perfect disagreement between the marks in Geography and Statistics, then what would be the value of rank correlation coefficient?

(a) Any value (b) Only +1

(c) Only -1 (d) (b) or (c)

Answer: (c) Only -1

[78] When we are not concerned with the magnitude of the two variables under discussion, we consider

(a) Rank correlation coefficient

(b) Product moment correlation coefficient

(c) Coefficient of concurrent deviation

(d) (a) or (b)

Answer: (c) Coefficient of concurrent deviation

[79] What is the quickest method to find correlation between two variables?

(a) Scatter diagram

(b) Method of concurrent deviation

(c) Method of rank correlation

(d) Method of product moment correlation

Answer: (b) Method of concurrent deviation

[80] What are the limits of the coefficient of concurrent deviations?

(a) No limit

(b) Between -1 and 0, including the limiting values

(c) Between0 and 1 , including the limiting values inclusive

(d) Between -1 and 1, the limiting values inclusive

Answer: (d) Between -1 and 1, the limiting values inclusive

[81] If for two variable x and y, the covariance, variance of X and variance of y are 40, 16 and 256 respectively, what is the value of the correlation coefficient?

(a) 0.01 (b) 0.625

(c) 0.4 (d) 0.5

Answer: (b) 0.625

[82] If cov(x,y) =15 , what restrictions should be put for the standard deviations of x and y ?

(a) No restriction.

(b) The product of the standard deviations should be more then 15.

(c) The product of the standard deviations should be less than 15.

(d) The sum of the standard deviations be less than 15.

Answer: (b) The product of the standard deviations should be more then 15

[83] If the covariance between two variables is 20 and the variance of one of the variables is 16, what would be the variance of the other variance?

(a) More then 100 (b) More then 10

(c) Less than 10 (d) More then

Answer: (a) More then 100

[84] If r = 0.6 then the coefficient of non-determination is

(a) 0.4 (b)-0.6

(c) 0.36 (d) 0.64

Answer: (d) 0.64

[85] From the following data

x:	2	3	5	4	7
y:	4	6	7	8	10

Coefficient of correlation was found to be 0.93. What is the correlation between u and v as given below?

u:	-3	-2	0	-1	2
v:	-4	-2	-1	0	2

(a) -0.93 (b) 0.93 (c) -0.93 (d) -0.57

Answer: (b) 0.93

[86] From the following data

x:	2	3	5	4	7
y:	4	6	7	8	10

Coefficient of correlation was found to be 0.93. What is the correlation between u and v as given below?

u:	10	15	25	20	35
v:	-24	-36	-42	-48	-60

(a) -0.6 (b)0.6 (c) -0.93 (d) -0.93

Answer: (c) -0.93

[87] If the sum of squares of difference of ranks, given by two judges A and B, of 8 students is 21, what is the value of rank correlation coefficient?

(a) 0.7 (b) 0.65 (c) 0.75 (d) 0.8

Answer: (c) 0.75

[88] If the rank correlation coefficient between marks in Statistics and mathematics for a group of student is 0.6 and the sum of the differences in ranks is 66, what is the number of students in the group?

(a) 1 (b) 9 (c) 8 (d) 11

Answer: (a) 1

[89] While computing rank correlation coefficient between profit and investment for the last 6 years of a company the difference in rank for a year was taken 3 instead of 4. What is the rectified rank correlation coefficient if it is known that the original value of rank correlation coefficient was 0.4?

(a) 0.3 (b) 0.2 (c) 0.25 (d) 0.28

Answer: (b) 0.2

[90] For 10 pairs of observations, number of concurrent deviations was found to be 4.What is the value of the coefficient of concurrent deviations?

(a) $\sqrt{0.2}$ (b) $-\sqrt{0.2}$ (c) $\frac{1}{3}$ (d) $-\frac{1}{3}$

Answer: (d) $-\frac{1}{3}$

[91] The coefficient of concurrent deviation for p pairs of observations was found to be $1/\sqrt{3}$. If the number of concurrent deviations was found to be 6, then the value of p is.

(a) 10 (b) 9 (c) 8 (d) none of these

Answer: (a) 10

[92] What is the value of correlation coefficient due to Pearson on the basis of the following data:

X:	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y:	27	18	11	6	3	2	3	6	11	18	27

(a) 1 (b) -1 (c) 0 (d) -0.5

Answer: (c) 0

[A] THEORY QUESTIONS:

[1] Define correlation coefficient. Show that it always lies between -1 and +1.

[2] Describe scatter diagram and explain how it is used to measure correlation.

[3] Define Spearman's rank correlation coefficients and state its formula.

[4] What is correlation? Explain its different types with illustrations.

[5] Show that correlation coefficient is independent of change of origin and scale.

[6] If X and Y are uncorrelated variables then show that: Var (X + Y) = Var (X - Y).

[7] State the properties of product moment correlation coefficient.

[B] NUMERICAL EXAMPLES:

[1] What is the coefficient of correlation from the following data:

X:	1	2	3	4	5
Y:	8	6	7	5	5

Answer:-0.85

[2] The coefficient of correlation between x and y where

X:	64	60	67	59	69
Y:	57	60	73	62	68

Answer: 0.655

[3] What is the coefficient of correlation between the ages of husbands and wives from the following data?

Age of husband (year)	46	45	42	40	38	35	32	30	27	25
Age of wife (year)	37	35	31	28	30	25	23	19	19	18

Answer:0.98

[4] The following result relate to bivariate data on (x,y):

$\sum xy = 414; \quad \sum x = 120 ; \quad \sum y = 90; \quad n = 30.$

$\sum x_i^2 = 600; \quad \sum y_i^2 = 300;$

Later on , It was known that two pairs of observations (12,11) and (6,8) were wrongly taken ,the correct pairs of observations being (10,9) and (8,10).The corrected value of the correlation coefficient is

Answer:0.846

[5] The following table provides the distribution of items according to size groups and also the number of defectives:

Size group:	9-11	11-13	13-15	15-17	17-19
No. of items:	250	350	400	300	150
No. of defective items:	25	70	60	45	20

The correlation coefficient between size and defectives is

Answer:0.07

[6] For two variables x and y ,it is known that cov (x,y) = 8 and r = 0.4 variance of x is16 and sum of squares of deviation of y from its mean is 250 .The number of observation for this bivariate data is

Answer:10

[7] Eight contestants in a musical contest were ranked by two judges A and B in the following manner:

Serial number of the contestants	1	2	3	4	5	6	7	8
Rank by judge A	7	6	2	4	5	3	1	8
Rank by judge B	5	4	6	3	8	2	1	7

The rank correlation coefficient is

Answer: 0.57

[8] Following are the marks of 10 students in Botany and Zoology:

Botany marks	58	43	50	19	28	24	77	34	29	75
Zoology marks	62	63	79	56	65	54	70	59	55	69

The coefficient of rank correlation between marks in botany and zoology is

Answer: 0.75

[9] What is the value of rank correlation between the following marks in physics and Chemistry?

Roll No	1	2	3	4	5	6
Marks in physics	25	30	46	30	55	80
Marks in chemistry	30	25	50	40	50	78

Answer: o.857

[10] What is the coefficient of concurrent deviations for the following data:

Supply	68	43	38	78	66	83	38	23	83	63	53
Demand	65	60	55	61	35	75	45	40	85	80	85

Answer: o.89

[11] What is the coefficient of concurrent deviations for the following data:

Year	1996	1997	1998	1999	2000	2001	2002	2003
Price	35	38	40	33	45	48	49	52
Demand	36	35	31	36	30	29	27	24

Answer: –o.43

[12] Karl Pearson’s coefficient of correlation between X and Y obtained from 10 pairs of items is 0.5. Means of X and Y are 12 and 15 respectively. Standard deviations of X and Y are 3 and 4 respectively. While checking it is

noticed that one of the observation was wrongly entered as 16 instead of 26 for X series and as 9 instead of 18 for Y series. Calculate the correct coefficient of correlation.

[13] Spearman's rank correlation coefficients between the marks in Accountancy and Statistics for a group of students is 0.5. If the sum of squares of differences between the ranks is 42 find the number of students in the group assume that no rank is repeated.

[14] Spearman's rank correlation coefficients between X and Y is 2/3. If the sum of squares of differences between the ranks is 55 find the number of pairs in the group assume that no rank is repeated.

[15] From the following data compute the coefficient of correlation:

Number of pairs of observation	= 10
Sum of X series	= 9
Sum of Y series	= 5
Sum of squares of X series	= 653
Sum of squares of Y series	= 595
Sum of products of X and Y series	= 534

[16] From the following data compute the coefficient of correlation:

Number of pairs of observation	= 10
Sum of X series	= 25
Sum of Y series	= 30
Sum of squares of X series	= 1500
Sum of squares of Y series	= 1800
Sum of products of X and Y series	= 2500