Faculty Science

Everest Shiwach

Department: Botany

B Sc III -Paper I (Plant Resource Utilization, Palynology, Plant Pathology and Biostatistics)

Unit- IV Topic- Correlation

Correlation is a statistical tool that helps to measure and analyse the degree of relationship between the two variables. According to Simpson and Kafka correlation is an analysis to determine the relationship between two or more variables. The measure of correlation is called Correlation coefficient and it is denoted by '*r*'.

For example- Relationship between height and weight of people, relationship between haemoglobin percentage and RBC number etc.

Significance of Correlation

1. Correlation are used to determine the relationship between two or more variables whether the variables are related or not.

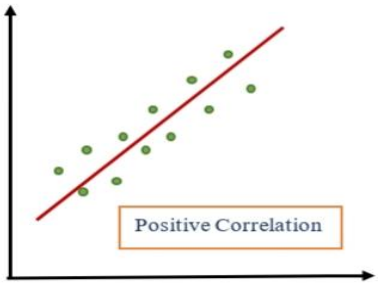
2. Reduce the range of uncertainty it gives a clear idea about the uncertainty among the
variables.

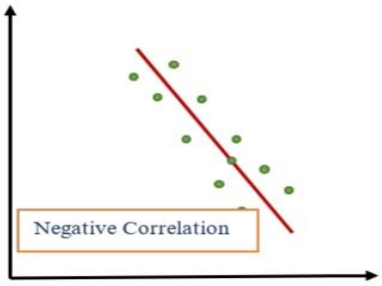
Degree of Correlation

Nature	Positive Correlation	Negative Correlation
Perfect Correlation	+1	-1
High Correlation	Between+0.75 and +1	Between -0.75 and -1
No Correlation	0	0

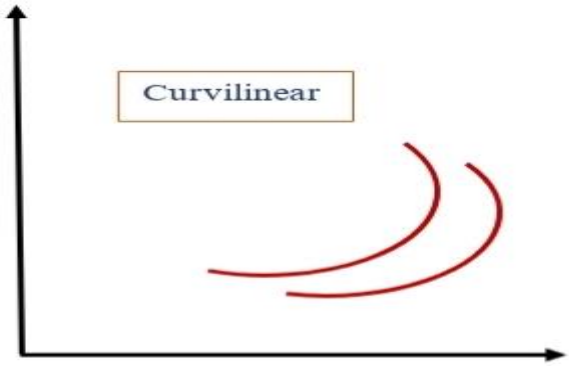
Types of correlation

1. Positive and Negative Correlation- When both variables change simultaneously. Then it is called positive correlation change may be either increasing or decreasing for example relationship between temperature and water percentage in the body of fish. It is also called direct correlation. In negative correlation one variable increases while other variable decreases. Both variables move in reverse direction.





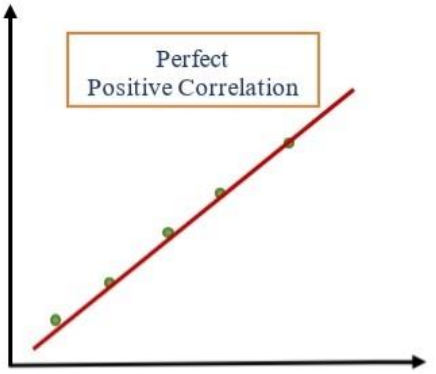
2. Linear and Curvilinear Correlation -When ratios of changes of both variables are uniform. If a graph is plotted of ratio on graph paper straight line would be obtained it would be obtained in linear correlation (e.g Perfect Positive and Negative Correlation). If the graph plotted of ratio of changes is curved then it is called curvilinear.

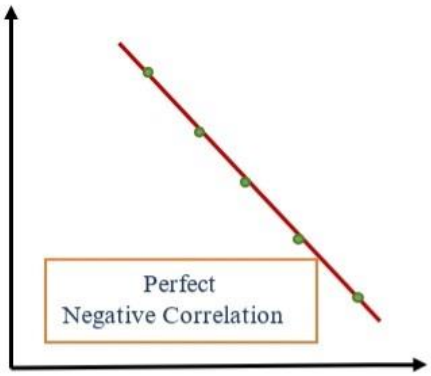


3. Simple and Multiple Correlation- When the number of variables in correlation is two then it is called simple correlation. For example- RBC and haemoglobin percentage. When number of variables in correlation more than two then it is called to be multiple correlation. For example- relationship between haemoglobin percentage, RBC number and iron intake.

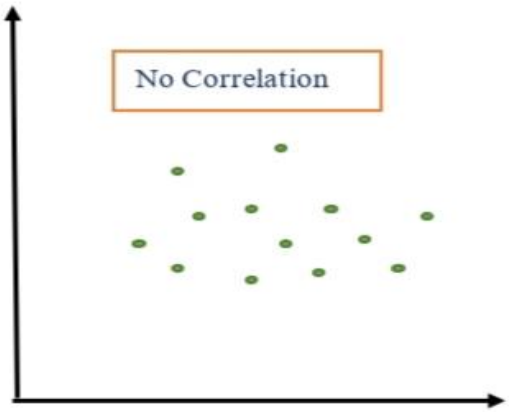
4. Perfect Positive and Perfect Negative Correlation- When both variables are directly proportional to each other. When proportionality occur in the same direction. It is called perfect positive correlation then r is + 1. When proportionality occur in reverse direction then it is called perfect negative

correlation then r is -1.





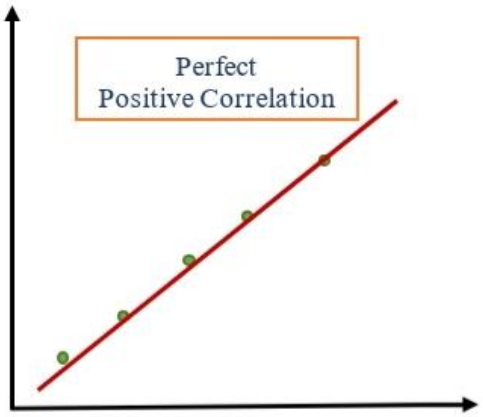
5. No Correlation- When variables are not dependent on each other then there is no correlation it is called zero correlation. Its value is 0 and r = 0.

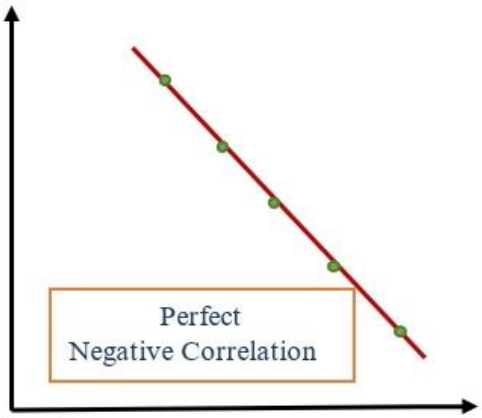


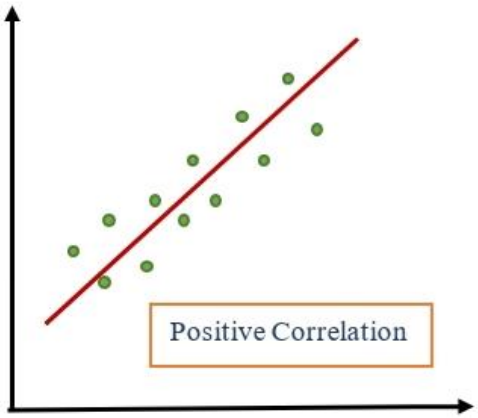
Methods of studying correlation

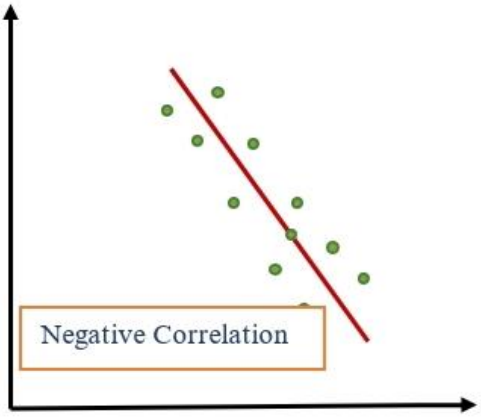
1. Scatter diagram
2. Karl Pearson's coefficient of correlation
3. Rank correlation

1.Scatter diagram- It is the simplest graphical method of showing correlation between two variables X and Y here statistical data are plotted against each pair of values of two variables by Dot and then cluster of dots make a shape this diagrammatic representation of bivariate data is known as a scatter diagram it indicate both the degree and the type of correlation.









High Degree Positive Correlation

High Degree Negative Correlation

Merits- 1. This method is a simple and non- mathematical to study correlation.
2.It is easy to understand.
3. It is not influenced by the size of the extreme items.
4. It may be treated as the first step in investing the relationship between two variables.
Demerits- 1. It gives only a rough idea about the direction and degree of correlation between two variables.
2. It cannot establish exact degree of correlation between two variables.

2 Karl Pearson's coefficient of correlation- It is the most widely used mathematical method of measuring correlation. It is a single number that tells us to what extent two things are related and to what extent variations is in one go with variation in other.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Where r = Correlation coefficient, x = Deviation of x variable and y = Deviation of y variable.

$x = X - \bar{x}, \qquad y = Y - \bar{y}.$

e.g. Data given of length and weight of fishes of seven groups as follows

Groups	A	B	C	D	E	F	G
Length in cm	6	7	8	9	10	11	12
Weight in gm	10	11	12	13	14	15	16

Groups	Length X	Weight Y	x	Y	X^2	Y^2	$x.y$
A	6	10	-3	-3	9	9	9
B	7	11	-2	-2	4	4	4
C	8	12	-1	-1	1	1	1
D	9	13	0	0	0	0	0
E	10	14	1	1	1	1	1
F	11	15	2	2	4	4	4
G	12	16	3	3	9	9	9
	$\sum x$ =63	$\sum y$ =91			29	29	29

$$\bar{x} = \frac{63}{7} = 9 \qquad \bar{y} = \frac{91}{7} = 13$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \qquad r = \frac{29}{\sqrt{29.29}} = 1$$

There is Perfect Positive Correlation between the body weight and length of fishes.

Merits- 1. It is widely used method to measure the relationship between two variables.

2. It gives direction and degree between two variables.

3. It summarizes in one precise and quantitative value, both degree and direction of correlation between two variables.

Demerits- 1. It always assumes the linear relationship.

2. The value of coefficient of correlation is affected by the extreme values.

3. It is difficult to calculate.

3.Spearman Rank Correlation Coefficient

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad d = R_1 - R_2 \qquad \text{n = number of pairs}$$

e.g. Number of ponds (*X*) in a town and number of fishes (*Y*) were given below in the table. Find the rank correlation(rho)?

X	24	24	25	26	26	27	28	29	30
Y	250	230	310	250	350	340	380	360	340

X	Rank of *X*	*Y*	Rank of *Y*	*d*	d^2
24	8+9/2 =8.5	250	7+8/2 =7.5	1	1
24	8+9/2 =8.5	230	9	-0.5	0.25
25	7	310	6	1	1
26	5+6/2 =5.5	250	7+8/2 =7.5	-2	4
26	5+6/2 =5.5	350	3	2.5	6.25
27	4	340	4+5/2 =4.5	-0.5	0.25
28	3	380	1	2	4
29	2	360	2	0	0
30	1	340	4+5/2 =4.5	-3.5	12.25
n =9		*n* =9			$\sum d^2$ =29

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad = 1 - \frac{6\times 29}{9(81-1)} = 0.758$$

Here *n* =9 calculated value of $\rho = 0.758$ Table value of significance at 0.1 level is 0.712. It means fish production in all ponds of observation has got same rate of growth.

Merits- 1. It is easy to understand and simple to calculate.
2. It is useful when factors under study are qualitative in nature.
3. It does not require the assumption of the normality of the population from which the sample observations are taken.

Demerits- 1. If bivariate grouped data is given, rank correlation coefficient cannot be calculated.
2. If *n* >30, this formula is time consuming.