Module 2

Correlation and Regression: Correlation - Significance, Types, and Methods, Scatter diagram, Karl Pearson correlation, Spearman's Rank correlation, Regression, Significance, Linear Regression Analysis, Types of regression models, Lines of Regression, Standard error of Estimate (Theory and Problems).

CORRELATION

Introduction:

In today's business world we come across many activities, which are dependent on each other. In businesses we see large number of problems involving the use of two or more variables. Identifying these variables and its dependency helps us in resolving the many problems. Many times there are problems or situations where two variables seem to move in the same direction such as both are increasing or decreasing. At times an increase in one variable is accompanied by a decline in another. For example, family income and expenditure, price of a product and its demand, advertisement expenditure and sales volume etc. If two quantities vary in such a way that movements in one are accompanied by movements in the other, then these quantities are said to be correlated.

Meaning:

Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables.

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Uses of correlations:

1. Correlation analysis helps inn deriving precisely the degree and the direction of such relationship.
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
3. Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing farces may become effective
4. Economic theory and business studies show relationships between variables like price and quantity demanded advertising expenditure and sales promotion measures etc.
5. The measure of coefficient of correlation is a relative measure of change.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:

I. Positive and Negative
II. Simple, Partial and Multiple
III. Linear and non-linear

I. Positive, Negative and Zero Correlation:

Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation. For example, the correlation between heights and weights of a group of persons is a positive correlation.

Height (cm): X	158	160	163	166	168	171	174	176
Weight (kg): Y	60	62	64	65	67	69	71	72

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.

Price of Product (Rs. Per Unit): X	6	5	4	3	2	1
Demand (In Units): Y	75	120	175	250	215	400

Zero Correlation: Actually, it is not a type of correlation but still it is called as zero or no correlation. When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation.

II. Simple, Partial and Multiple Correlation:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation: In case of partial correlation, one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

Multiple Correlation: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc., it is a case of multiple correlation.

III. Linear and Non-linear Correlation:

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line. For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.

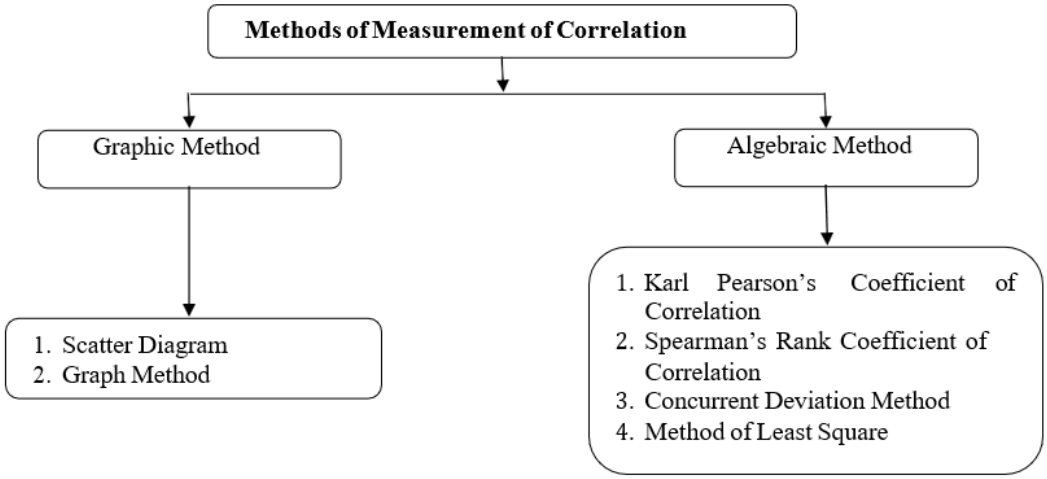
Raw material: X	10	20	30	40	50	60
Finished Product: Y	2	4	6	8	10	12

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

Advertisement Expenses: X	10	20	30	40	50	60
Sales Volume: Y	2	4	6	8	10	12

Methods of measurement of correlation:

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below:



Among these methods we will discuss only the following methods:

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

Scatter Diagram:

This is graphic method of measurement of correlation. It is a diagrammatic representation of bivariate data to ascertain the relationship between two variables. Under this method the given data are plotted on a graph paper in the form of dot. i.e. for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. Usually, an independent variable is shown on the X-axis whereas the dependent variable is shown on the Y-axis. Once the values are plotted on the graph it reveals the type of the correlation between variable X and Y. A scatter diagram reveals whether the movements in one series are associated with those in the other series.

- Perfect Positive Correlation: In this case, the points will form on a straight line falling from the lower left-hand corner to the upper right-hand corner.
- Perfect Negative Correlation: In this case, the points will form on a straight line rising from the upper left-hand corner to the lower right-hand corner.
- High Degree of Positive Correlation: In this case, the plotted points fall in a narrow band, wherein points show a rising tendency from the lower left-hand corner to the upper right-hand corner.

- High Degree of Negative Correlation: In this case, the plotted points fall in a narrow band, wherein points show a declining tendency from upper left-hand corner to the lower right-hand corner.
- Low Degree of Positive Correlation: If the points are widely scattered over the diagrams, wherein points are rising from the left-hand corner to the upper right-hand corner.
- Low Degree of Negative Correlation: If the points are widely scattered over the diagrams, wherein points are declining from the upper left-hand corner to the lower right-hand corner.
- Zero (No) Correlation: When plotted points are scattered over the graph haphazardly, then it indicates that there is no correlation or zero correlation between two variables.

Perfect Positive Correlation

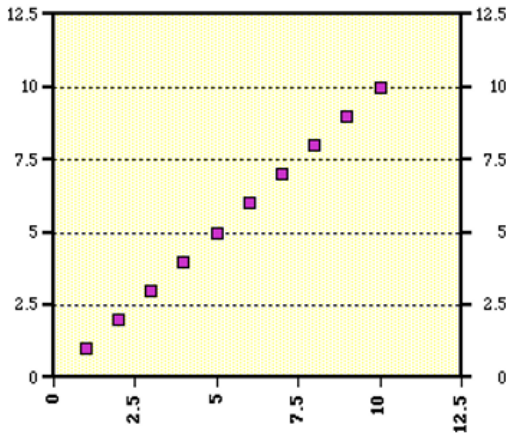


Diagram – I

Perfect Negative Correlation

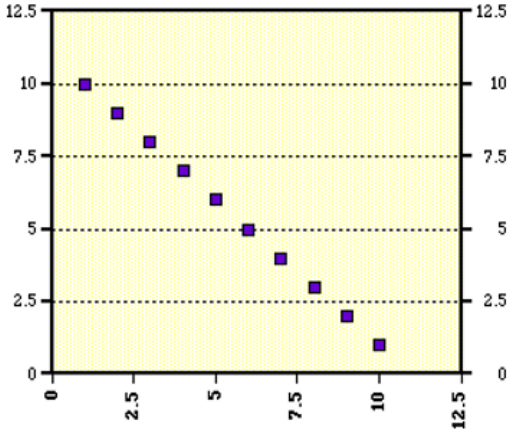


Diagram – II

High Positive Correlation

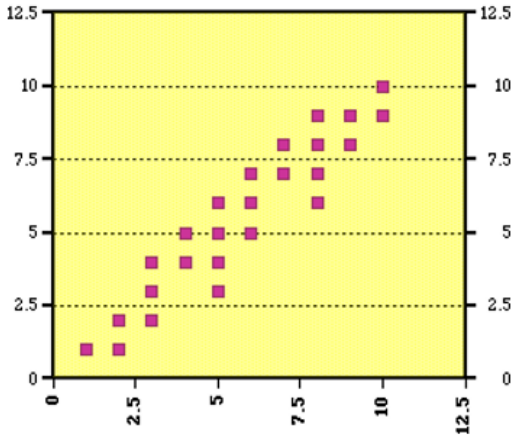


Diagram – III

High Negative Correlation

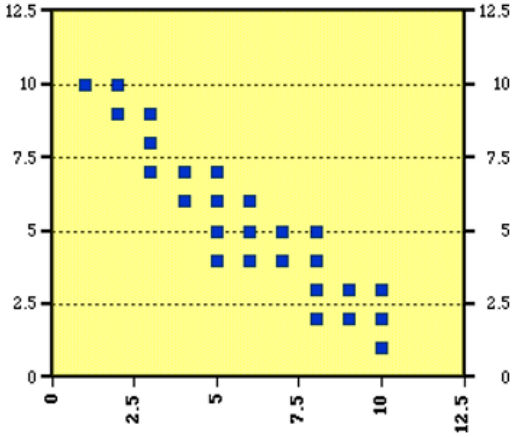


Diagram – IV

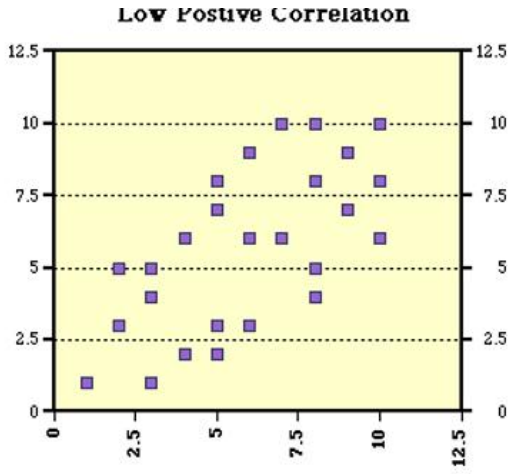


Diagram – V

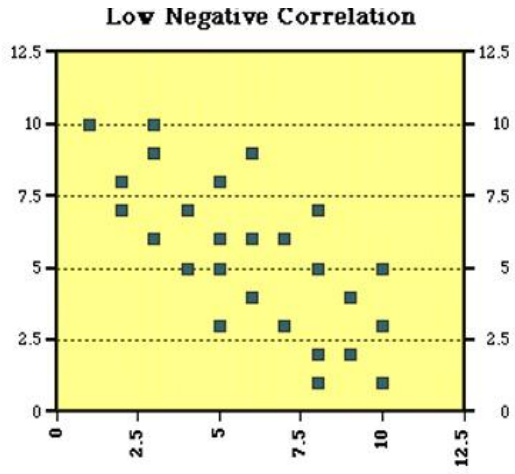
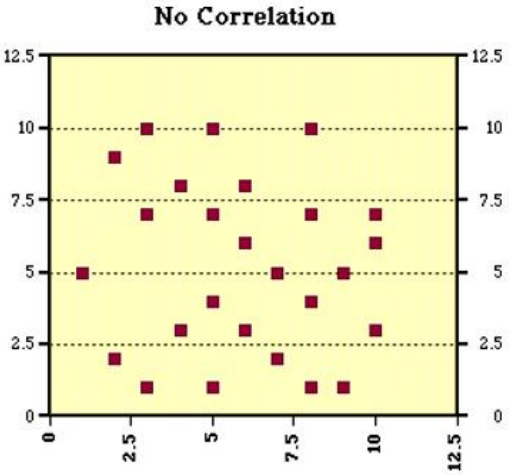


Diagram – VI



What is Karl Pearson's Coefficient of Correlation?

The first person to give a mathematical formula for the measurement of the degree of relationship between two variables in 1890 was Karl Pearson. Karl Pearson's Coefficient of Correlation is also known as **Product Moment Correlation** or **Simple Correlation Coefficient.** This method of measuring the coefficient of correlation is the most popular and is widely used. It is denoted by **'r',** where r is a pure number which means that r has no unit.

> *According to **Karl Pearson,** "Coefficient of Correlation is calculated by dividing the sum of products of deviations from their respective means by their number of pairs and their standard deviations."*

Features of Karl Pearson's Coefficient of Correlation

The main features of Karl Pearson's Coefficient of Correlation are as follows:

1. Knowledge of Direction of Correlation: This method of measuring coefficient of correlation gives us knowledge about the direction of the relationship between two variables. In other words, it tells us whether the relationship between two variables is positive or negative.

2. Size of Correlation: Karl Pearson's Coefficient of Correlation indicates the size of the relationship between two variables. Besides, Correlation Coefficient ranges between -1 and +1.

3. Indicates Magnitude and Direction: This method not only specifies the magnitude of the correlation between two variables but also specifies its direction. It means that, if two variables are directly related, then the correlation coefficient between the variables will be a positive value. However, if two variables are inversely related, then the correlation coefficient between the variables will be a negative value.

4. Ideal Measure: As this method is based on the most essential statistical measure, such as standard deviation and mean, it is an ideal/appropriate measure.

Note: The value of the Correlation Coefficient should always lie between -1 and +1.

- *When r = +1, it means that there is perfect positive correlation.*
- *When r = -1, it means that there is perfect negative correlation.*
- *When r = 0, it means that there is no or zero correlation.*

Properties of Coefficient of Correlation

1. Coefficient of Correlation is Independent of change of origin and scale of measurements: Coefficient of Correlation is not affected by the change of origin and scale of measurement.

2. Coefficient of Correlation lies between -1 and +1: The property of r also serves as a useful check on the correctness of the calculations. If the value of r lies outside the range, then it would mean that there is some error in the calculations.

3. Zero Correlation: If two variables (say X and Y) are independent of each other, in that case, the coefficient of correlation between them will be zero.

4. Measure of Linear Relationship: The coefficient of correlation is a measure that helps in determining the linear relationship between two variables. If both the variables (say X and Y) increase or decrease together, then r will be positive. However, if one variable increases when the other variable decreases or vice-versa, then r will be negative.

Merits of Karl Pearson's Coefficient of Correlation

Various advantages of Karl Pearson's Coefficient of Correlation are as follows:

1. Popular Method: Karl Pearson's Coefficient of Correlation is the most popular and widely used mathematical method to study the correlation between two variables.

2. Degree and Direction of Correlation: The value of correlation coefficient not only summarises the degree of correlation but also its direction.

Demerits of Karl Pearson's Coefficient of Correlation

Various disadvantages of Karl Pearson's Coefficient of Correlation are as follows:

1. Affected by Extreme Values: If the values of the two variables are extreme, then it would have a great impact on the value of correlation coefficient.

2. Assumption of Linear Relationship: While determining correlation coefficient, it is always assumed that there is a linear relationship without thinking whether the assumption is correct or not.

3. Time-Consuming Method: In comparison to other methods of determining correlation coefficients, this method takes more time.

4. Possibility of Wrong Interpretation: While interpreting the value of coefficient of correlation using this method, one has to be very careful. It is because the chances of misinterpreting the coefficient are more.

Formula of Karl Pearson's Coefficient of Correlation

$$Karl\ Pearson's\ Coefficient\ of\ Correlation(r) = \frac{Sum\ of\ Products\ of\ Deviations\ from\ their\ respective\ means}{Number\ of\ Pairs \times Standard\ Deviations\ of\ both\ Series}$$

Or

$$r = \frac{\sum xy}{N \times \sigma_x \times \sigma_y}$$

Where,

N = Number of Pair of Observations

x = Deviation of X series from Mean $(X - \bar{X})$

y = Deviation of Y series from Mean $(Y - \bar{Y})$

σ_x = Standard Deviation of X series $(\sqrt{\frac{\sum x^2}{N}})$

σ_y = Standard Deviation of Y series $(\sqrt{\frac{\sum y^2}{N}})$

r = Coefficient of Correlation

Methods of Calculating Karl Pearson's Coefficient of Correlation

1. Actual Mean Method
2. Direct Method
3. Short-Cut Method/Assumed Mean Method/Indirect Method
4. Step-Deviation Method

1. Actual Mean Method

The steps involved in the calculation of coefficient of correlation by using Actual Mean Method are:

1. The first step is to calculate the mean of the given two series (say X and Y).
2. Now, take the deviation of X series from $\bar{X}$ and denote the deviations by x.
3. Square the deviations of x and obtain the total; i.e., $\sum x^2$
4. Take the deviation of Y series from $\bar{Y}$ and denote the deviations by y.
5. Square the deviations of y and obtain the total; i.e., $\sum y^2$
6. Multiply the respective deviations of Series X and Y and obtain the total; i.e., $\sum xy$.
7. Now, use the following formula to determine the Coefficient of Correlation:

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

2. Direct Method

The steps involved in the calculation of coefficient of correlation by using Direct Method are:

1. The first step is to calculate the sum of Series X ($\sum X$).
2. Now, calculate the sum of Series Y ($\sum Y$).
3. Square the values of X Series and calculate their total; i.e., $\sum X^2$.
4. Square the values of Y Series and calculate their total; i.e., $\sum Y^2$.
5. Multiply the values of Series X and Y and calculate their total; i.e., $\sum XY$.
6. Now, use the following formula to determine Coefficient of Correlation:

$$r = \frac{N\sum XY - \sum X.\sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

3. Short-Cut Method/Assumed Mean Method

Actual Mean can sometimes come in fractions which can make the calculation of standard deviation complicated and difficult. In those cases, it is suggested to use Short-Cut Method to simplify the calculations. The steps involved in the calculation of coefficient of correlation by using Assumed Mean Method are:

1. First of all, take the deviations of X Series from the assumed mean and denote the values by dx. Calculate their total; i.e., ∑dx.
2. Now, square the deviations of X series and calculate their total; i.e., $\sum dx^2$.
3. Take the deviations of Y Series from the assumed mean and denote the values by dy. Calculate their total; i.e., ∑dy.
4. Square the deviations of Y series and calculate their total; i.e., $\sum dy^2$.
5. Multiply dx and dy and calculate their total; i.e., ∑dxdy.
6. Now, use the following formula to determine Coefficient of Correlation:

$$r = \frac{N\sum dxdy - \sum dx.\sum dy}{\sqrt{N\sum dx^2 - (\sum dx)^2}\sqrt{N\sum dy^2 - (\sum dy)^2}}$$

Where,

N = Number of pair of observations

∑dx = Sum of deviations of X values from assumed mean

∑dy = Sum of deviations of Y values from assumed mean

$\sum dx^2$ = Sum of squared deviations of X values from assumed mean

$\sum dy^2$ = Sum of squared deviations of Y values from assumed mean

∑dxdy = Sum of the products of deviations dx and dy

4. Step Deviation Method

This method simplifies the calculation of coefficient of correlation as the deviations are taken from assumed means and are divided by a common factor. The steps involved in the calculation of coefficient of correlation by using Step Deviation Method are:

1. First of all, take the deviations of Series X from the assumed mean and divide them by Common Factor (C) to determine step deviation (dx') . Calculate the total of step deviations; i.e., $\sum dx'$
2. Take the deviations of Series Y from the assumed mean and divide them by Common Factor (C) to determine step deviation (dy') . Calculate the total of step deviations; i.e., $\sum dy'$
3. Square the step deviation of Series X and determine their total; i.e., $\sum dx'^2$
4. Square the step deviation of Series Y and determine their total; i.e., $\sum dy'^2$
5. Multiply (dx') and (dy') , and determine their total; i.e., $\sum dx'dy'$
6. Now, use the following formula to determine Coefficient of Correlation:

$$r = \frac{N\sum dx'dy' - \sum dx'.\sum dy'}{\sqrt{N\sum dx'^2 - (\sum dx')^2}\sqrt{N\sum dy'^2 - (\sum dy')^2}}$$

...

Where,

N = Number of pair of observations

$\sum dx'$ = Sum of deviations of X values from assumed mean

$\sum dy'$ = Sum of deviations of Y values from assumed mean

$\sum dx'^2$ = Sum of squared deviations of X values from assumed mean

$\sum dy'^2$ = Sum of squared deviations of Y values from assumed mean

$\sum dx'dy'$ = Sum of the products of deviations (dx') and (dy')

Spearman's Rank Correlation Coefficient

Spearman's Rank Correlation Coefficient or Spearman's Rank Difference Method or Formula is a method of calculating the correlation coefficient of qualitative variables and was developed in **1904 by Charles Edward Spearman**. In other words, the formula determines the correlation coefficient of variables like beauty, ability, honesty, etc., whose quantitative measurement is not possible. Therefore, these attributes are ranked or put in the order of their preference.

$$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$$

In the given formula,

r_k = Coefficient of rank correlation

D = Rank differences

N = Number of variables

Case 1: When Ranks are given

In this case, the ranks of the frequency distribution or variables are already given, and the coefficient of rank correlation is calculated based on those ranks. The formula for calculating Spearman's Rank Correlation is

$$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$$

Case 2: When Ranks are not given

When the ranks of the variables or distribution are not given, then the individual has to rank the values themselves. While ranking the values, one has to adopt a uniform procedure for both series of distribution. For instance, if 1st rank is given to the lowest value of one series, then the same pattern should be followed for the second series as well. Once the rank has been determined, the coefficient of rank correlation is determined as the first case. The formula for calculating Spearman's rank correlation coefficient is

$$r_k = 1 - \frac{6\sum D^2}{N^3 - N}$$

Case 3: When Ranks are equal

When two or more values of a series have an equal rank, then in such cases, each value is given the average of the two ranks. To avoid any mistake, the formula for calculating Spearman's Rank Correlation Coefficient is

$$r_k = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots]}{N^3 - N}$$

Here, m_1, m_2, ... are the number of times a value has repeated in the given X, Y, ... series, respectively.

What do the values of Spearman's Rank Correlation Coefficient indicate?

r_k = +1: Perfect Positive Monotonic Correlation

r_K = −1: Perfect Negative Monotonic Correlation

r_K = 0: No Monotonic Correlation

What is the significance of a positive or negative Spearman's Rank Correlation Coefficient?

Positive r_K: *As one variable increases, the other variable tends to increase.*

Negative r_K: *As one variable increases, the other variable tends to decrease.*

Can Spearman's Rank Correlation be used for small sample sizes?

Yes, Spearman's Rank Correlation is appropriate for small sample sizes. However, the reliability of the correlation increases with larger sample sizes.