ASSOCIATION. CORRELATION.

Correlation is measured by a correlation coefficient, and requires that there is an association between two variables, but association does not imply correlation. Technically, association refers to any relationship between two variables, whereas correlation is often used to refer only to a linear relationship between two variables. The terms are used interchangeably in this guide, as is common in most statistics texts.

The **correlation coefficient** is a numerical value that refers to the extent to which two variables move together in an increasing or decreasing trend. Correlations can be linear or non-linear. For quantitative and ordinal data, there are two primary measures of correlation: Pearson's correlation (r), which measures linear trends, and Spearman's (rank) correlation (s), which measures increasing and decreasing trends that are not necessarily linear, but can be U-shaped or J-shaped. Pearson's correlation assumes that both variables are normally distributed, whereas Spearman's (rank) correlation is non-parametric.

Correlation coefficients

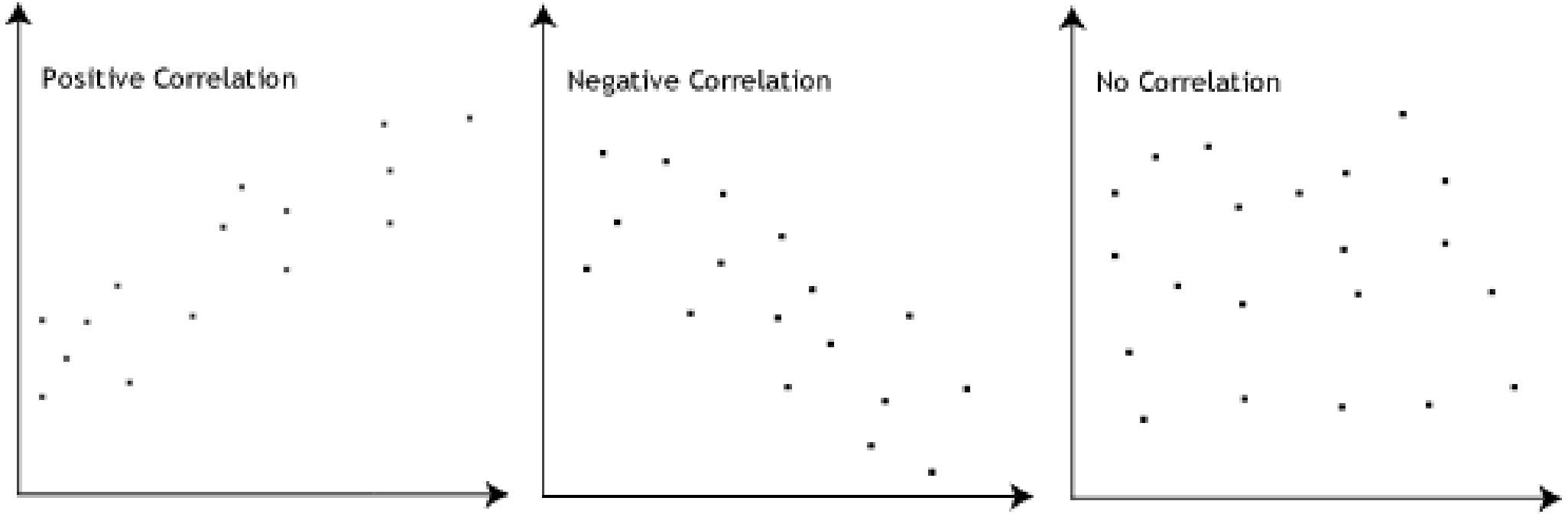
- Correlation coefficients are on a -1 to 1 scale.
- On this scale -1 indicates a perfect negative relationship. High values of one variable are associated with low values of the other.
- Likewise, a correlation of +1 describes a perfect positive relationship. High values of one variable are associated with high values of the other.
- 0 indicates no relationship. High values of one variable co-occur as often with high and low values of the other.
- There is no independent and no dependent variable in a correlation. It's a bivariate descriptive statistic.
- The most common correlation coefficient is the Pearson correlation coefficient.
- Pearson's coefficient assumes that both variables are normally distributed. This requires they be truly continuous and unbounded

What is the difference between association and correlation?

- Association refers to the general relationship between two random variables while the correlation refers to a more or less a linear relationship between the random variables.
- • Association is a concept, but correlation is a measure of association and mathematical tools are provided to measure the magnitude of the correlation.
- • Pearson’s product moment correlation coefficient establishes the presence of a linear relationship and determines the nature of the relationship (whether they are proportional or inversely proportional).
- • Rank correlation coefficients are used to determine the nature of the relationship only, excluding the linearity of the relation (it may or may not be linear, but it will tell whether the variables increase together, decrease together or one increases while the other decreases or vice versa).

Pearson product-moment correlation

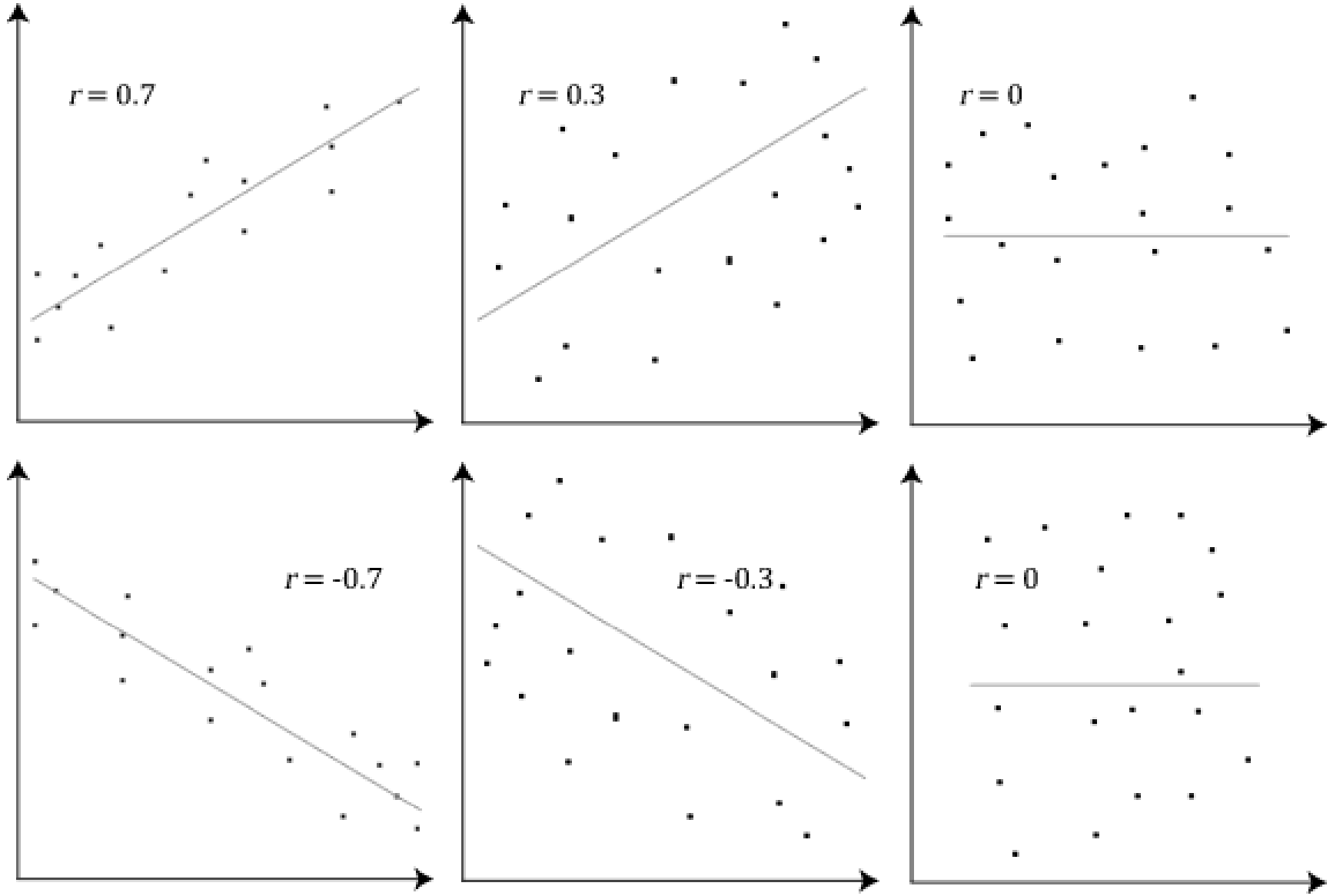
- The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit)



How can we determine the strength of association based on the Pearson correlation coefficient?

- The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit.

DIFFERENT RELATIONSHIPS AND THEIR CORRELATION COEFFICIENTS ARE SHOWN IN THE DIAGRAM BELOW:



MERITS

The following are the chief points of merit that go in favour of the Karl Pearson's method of correlation:

- This method not only indicates the presence, or absence of correlation between any two variables but also, determines the exact extent, or degree to which they are correlated.
- Under this method, we can also ascertain the direction of the correlation i.e. whether the correlation between the two variables is positive, or negative.
- This method enables us in estimating the value of a dependent variable with reference to a particular value of an independent variable through regression equations.
- This method has a lot of algebraic properties for which the calculation of co-efficient of correlation, and a host of other related factors viz. co-efficient of determination, are made easy.

DEMERITS

This method also suffers from the following demerits:

- It is comparatively difficult to calculate as its computation involves intricate algebraic methods of calculations.
- It is very much affected by the values of the extreme items.
- It is based on a large number of assumptions viz. linear relationship, cause and effect relationship etc. which may not always hold good.
- It is very much likely to be misinterpreted particularly in case of homogeneous data.
- In comparison to the other methods, it takes much time to arrive at the results.
- It is subject to probable error which its propounder himself admits, and therefore, it is always advisable to compute it probable error while interpreting its results.

Spearman's Rank Correlation

- Spearman's rank correlation measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.
- The formula for Spearman's rank coefficient is: $$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2-1)}$$
- $\boldsymbol{\rho}$ = Spearman's rank correlation coefficient
- di = Difference between the two ranks of each observation
- n = Number of observations
- The Spearman Rank Correlation can take a value from +1 to -1 where,
- A value of +1 means a perfect association of rank
- A value of 0 means that there is no association between ranks
- A value of -1 means a perfect negative association of rank

ADVANTAGES OF SPEARMAN RANK CORRELATION COEFFICIENT:

- This method is much simpler to carry out and understand compared to Karl Pearsons method of correlation. It gives the same result as the Karl Pearson method if none of the values/ranks are repeated.
- This method can be used to carry out correlation analysis for variables that are not numerical. We can study the relationships between qualitative variables such as beauty, intelligence, honesty, efficiency, and so on.
- Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure that is distribution-free (or non-parametric). A distribution-free measure is one that does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (i.e., distribution-free) since no strict assumptions are made about the form of the population from which the sample observations are drawn.
- Spearman's formula is the only formula to be used for finding the correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially.
- It can also be used when actual numerical data is given. The sample data of values of two variables are converted into ranks either in ascending order or descending order for calculating the degree of correlation between two variables.

Disadvantages of Spearman Rank Correlation Coefficient

- Values of both variables are assumed to be describing a linear relationship rather than a non-linear relationship.
- A large computational time is required when the number of pairs of values of two variables exceeds 30. In such cases, assigning ranks to each of the numerical values is a very time-consuming and tedious process.
- This method cannot be applied to measure the association between two variables whose distribution is given in the form of a grouped frequency distribution

COMPARISON OF PEARSON AND SPEARMAN COEFFICIENTS

- The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.
- 2. One more difference is that Pearson works with raw data values of the variables whereas Spearman works with rank-ordered variables.

What is a monotonic relationship?

- A monotonic relationship is a relationship that does one of the following:
- (1) as the value of one variable increases, so does the value of the other variable, OR,
- (2) as the value of one variable increases, the other variable value decreases.
- BUT, not exactly at a constant rate whereas in a linear relationship the rate of increase/decrease is constant.

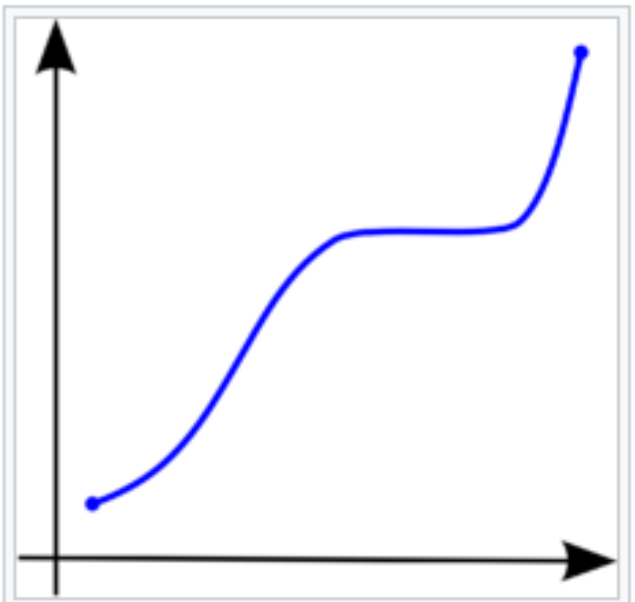


Figure 1. A monotonically increasing function.

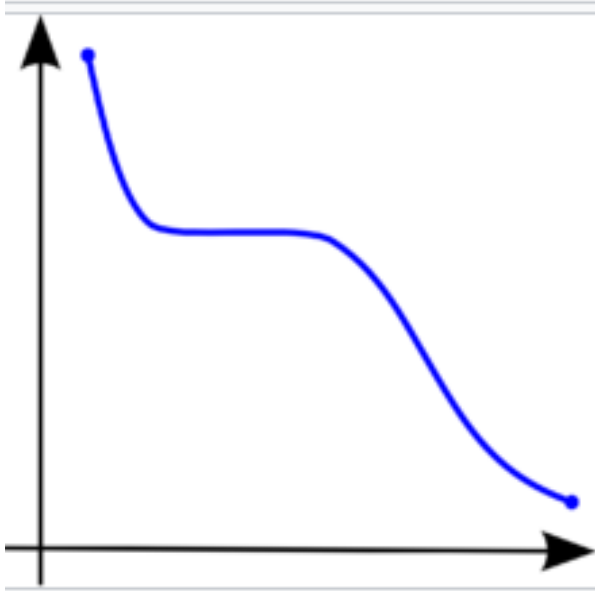


Figure 2. A monotonically decreasing function

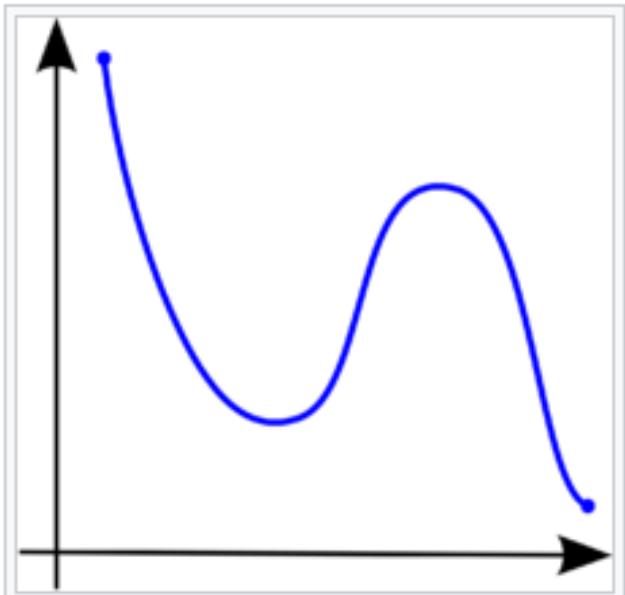


Figure 3. A function that is not monotonic

Rank Coefficient	Karl Pearson's Coefficient
It is suitable when data is given in the qualitative form.	It is a suitable method when data is given in the quantitative form.
It cannot be applied in the case of bivariate frequency distribution.	It is an effective method to determine the correlation in the case of grouped series.
It is not possible to determine the combined coefficient of correlation	If coefficients of correlation and number of items of each subgroup is given then one can determine the combined coefficient of correlation items.
Changing the actual values in the series does not result in a change in the coefficient of correlation.	Changing the actual values in the series results in a change in the coefficient of correlation.
The coefficient of correlation is perfectly positive if both the series have equal corresponding ranks i.e. D = 0 for each.	The coefficient of correlation is perfectly positive if both the series change uniformly i.e. X and Y series are related linearly correlation.
It is difficult to use and understand.	It is easier to use and understand.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation	Regression
Correlation describes as a statistical measure that determines the association or co-relationship between two variables.	Regression depicts how an independent variable serves to be numerically related to any dependent variable.
Its coefficients may range from -1.00 to +1.00.	Both variables serve to be different, One variable is independent, while the other is dependent.
To find the numerical value that defines and shows the relationship between variables.	To estimate the values of random variables based on the values shown by fixed variables.
Its coefficient serves to be independent of any change of Scale or shift in Origin.	Its coefficient shows dependency on the change of Scale but is independent of its shift in Origin.
Its coefficient is mutual and symmetrical.	Its coefficient fails to be symmetrical.
Its correlation serves to be a relative measure.	Its coefficient is generally an absolute figure.
In this, both variables x and y are random variables	In this, x is a random variable while y is a fixed variable. At times, both variables may be like random variables.

Reference:-

- Statistical Methods by N.G.Das
- https://corporatefinanceinstitute.com/resources/data-science/regression-analysis
- https://www.alchemer.com/resources/blog/regression-analysis

THANK YOU

SOMA MUKHOPADHYAY, RKSMVV