Karl Pearson's Coefficient Correlation

Karl Pearson's Correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linear related variables.

It is used when the data are normally distributed. And hence it is a parametric technique.

While calculating the Pearson's Correlation Coefficient, we make / follow some assumptions :

M T W T F S S
Date __/__/__

→ There is a linear relationship (or any linear component of the relation-ship) between the two variables.

→ We keep Outliers either to a minimum or remove them entirely.

→ The Karl Pearson's product-moment Correlation Coefficient is a measure of the strength of a linear association between two variables and is denoted by r, X and y being the two variables.

There are many situations in our daily life where we know from experience, the direct association between certain variables but we can't put a certain measure to it. For example, you know that the chances of you going out to watch a newly released movie is directly associated with the number of friends who go with you because the more the merrier.

Linearity and homoscedasticity are important assumptions of Karl Pearson's Correlation. Linearity assumes a straight line relationship between each of the two variables and homoscedasticity assumes that data is equally distributed about the regression line.

It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s and for which the mathematical formula was derived and published by Auguste Bravais in 1844.

The coefficient of correlation measures not only the magnitude of correlation but also tells the direction.

Such as, $r = -0.67$ which shows correlation is negative because the sign is "−" and the magnitude is 0.67.

Assumptions:

Independent of case: Cases should be independent to each other.

Linear relationship: Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.

Homoscedasticity: the residuals scatterplot should be roughly rectangular-shaped.

Properties:

Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..

Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.

Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).

High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.

Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.

Low degree: When the value lies below ± .29, then it is said to be a small correlation.

No correlation: When the value is zero.

- When a correlation coefficient is (1) that means every increase in one variable, there is a positive increase in other fixed proportion. For instance, shoe sizes change according to the length of the foot and are (almost) perfect correlation.
- When a correlation coefficient is (-1) that means every positive increase in one variable, there is a negative decrease in other fixed proportion. For instance,with the decrease in the quantity of gas in a gas tank, it shows (almost) a perfect correlation with speed.
- When a correlation coefficient is (0) for every increase, it means there is no positive or negative increase and the two variables are not related.