Statistics

Correlation

1. Introduction

Welcome to the series of E-learning modules on Correlation.
By the end of this session, you will be able to:

- Define Correlation
- Explain the correlation and causation
- Explain the interpretation of the value of correlation coefficient
- Explain the properties of correlation coefficient
- Explain the importance of correlation coefficient
- List the merits and demerits of the correlation coefficient

A disadvantage of the covariance as a measure of relationship between X and Y is that it depends on the units of X and Y.
Correlation is the degree to which two or more quantities are linearly associated. In a two-dimensional plot, the degree of correlation between the values on the two axes is quantified by the correlation coefficient.

It is the scale-less measure of the relationship between X and Y. It is obtained by dividing the covariance by the product of the two standard deviations. That is, Rho X Y is equal to covariance of X, Y divided by standard deviation of X into Standard deviation of Y Is equal to sigma X, Y divided by sigma X into sigma Y

The quantity row X Y is called the *linear correlation coefficient*, which measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is also referred to as the Pearson product moment correlation coefficient as it was formulated by Karl Pearson. The correlation assumes value between minus 1 and plus 1.

Correlation analysis may answer some questions like,
How long will someone live?

- Fluctuation in the stock market.
- Chances of someone becoming a criminal.
- Chances of a surgery prolonging a cancer patient's life.
- Probability of a depressed person committing suicide.
- Chances of an employee becoming productive.
- Probability of a football team playing first down on the next game.
- Chances of success or failure of a marriage.

2. Correlation and Causation

Now, let us discuss about the correlation and causation.
Causation means 'cause and effect' relationship. No discussion of correlation would be complete without a discussion of causation. It is possible for two variables to be related (correlated), but not have one variable cause another.

For example, suppose there is a high correlation between the number of popsicles sold and the number of drowning deaths. Does that mean that one should not eat popsicles before one swims? Not necessarily. Both of the above variables are related to a common variable, the heat of the day.

The hotter the temperature, the more popsicles sold and also more people swim. Thus, leading to more deaths due to drowning. This is an example of correlation without causation. Much of the early evidence that cigarette smoking causes cancer was correlational. It may be that people who smoke are more nervous and nervous people are more susceptible to cancer. It may also be that smoking does indeed cause cancer. The cigarette companies made the former argument, while some doctors made the latter. In this case, the relationship is causal and therefore do not smoke.

Sociologists are very much concerned with the question of correlation and causation because much of their data is correlational. Sociologists have developed a branch of correlational analysis, which is called path analysis that is precisely to determine causation from correlations (Blalock, 1971). Before a correlation may imply causation, certain requirements must be met.

These requirements include:
1. The causal variable must temporally precede the variable it causes, and
2. Certain relationships between the causal variable and other variables must be met.

If a high correlation was found between the age of the teacher and the students' grades, it does not explain that older teachers are more experienced, teach better, and give higher grades. Neither does it tells us that older teachers are soft touches, don't care, and give higher grades. The correlation means that older teachers give higher grades; younger teachers give lower grades. It does not explain why it is the case. Hence, any correlation between two variables without causation is called non-sense correlation or spurious correlation.

3. Interpretation of the Value of Correlation Coefficient

Now, let us discuss how to interpret the correlation coefficient.
The following points are the accepted guidelines for interpreting the correlation coefficient:

- Zero indicates no linear relationship.
- Plus 1 indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
- Minus 1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
- Values between zero and zero point 3 (0 and minus zero point 3) indicate a weak positive (negative) linear relationship via a shaky linear rule.
- Values between zero point 3 and zero point 7 (minus zero point 3 and minus zero point 7) indicate a moderate positive (negative) linear relationship via a fuzzy-firm linear rule.
- Values between zero point 7 and 1 point zero (minus zero point 7 and minus 1 point zero) indicate a strong positive (negative) linear relationship via a firm linear rule.
- The value of r square is typically taken as “the percent of variation in one variable explained by the other variable,” or “the percent of variation shared between the two variables.”
- Linearity Assumption: The correlation coefficient requires that the underlying relationship between the two variables under consideration is linear. If the relationship is known to be linear, or the observed pattern between the two variables appears to be linear, then the correlation

coefficient provides a reliable measure of the strength of the linear relationship. If the relationship is known to be nonlinear, or the observed pattern appears to be nonlinear, then the correlation coefficient is not useful, or at least questionable.

Now, let us discuss the necessity to know the strength of the relationship.
It is important to know the strength of relationship because it allows you to make predictions. The main point of knowing the strength is that a strong relationship allows you to make much more accurate predictions than a weak relationship. This ability to make accurate predictions is critical in many professional settings. Psychologists, medical professionals, business executives, stock brokers, military leaders, law enforcement agents are all interested in being able to make predictions. The concept of correlation provides a tool that helps people to make predictions with some amount of accuracy.

4. Properties and Importance of Correlation Coefficient

Following are the properties of coefficient of correlation:

1. Correlation coefficient is independent of unit of measurement of the variables.
2. If correlation is present, then coefficient of correlation would lie between plus or minus 1. If correlation is absent, then it is denoted by zero. That is minus 1 less than or equal to r less than or equal to 1.
3. Correlation coefficient is independent of change of origin and scale.
4. If X and Y are random variables and a, b, c, d are any numbers provided that a is not equal to zero, c is not equal to zero then

Coefficient of correlation between a into X plus b and c into Y plus d is equal to a into c divided by mod a into c into r of X, Y. (these properties are proved in earlier modules)

The variables X and Y are connected by the equation a into X plus b into Y plus c is equal to zero. Show that the correlation between them is minus 1 if the signs of a and b are alike and plus 1 if they are different.
To prove this result, consider, a into X plus b into Y plus c is equal to zero. Taking expectation, we get,
a into Expectation of X plus b into Expectation of Y plus c is equal to zero Implies, a into X minus expectation of X plus b into Y minus Expectation of Y is equal to zero Implies, X minus Expectation of X is equal to minus b divided by a into Y minus Expectation of Y
Therefore, Covariance of X, Y is equal Expectation of X minus Expectation of X into Y minus Expectation of Y Is equal to minus b divided by a into Y minus Expectation of Y the whole square Is equal to minus b divided by a into sigma Y square.

Sigma X square is equal to Expectation of X minus Expectation of X the whole square Is equal to b square divided by a square into sigma Y square

Therefore, r is equal to covariance of X, Y divided by sigma X into sigma Y is equal to minus b divided by a into sigma Y square divided by square root of sigma Y square into square root of b square divided by a square into sigma Y square is equal to minus b divided by a into sigma y square divided by modulus of b divided by a into sigma y square is equal to plus 1 if b and a are of opposite signs and minus 1 if b and a are of same signs.

If X and Y are uncorrelated random variables, and r of a into X plus b into Y and b into X plus a into Y is equal to 1 plus 2 into a into b divided by a square plus b square. Find r of (X, Y), the coefficient of correlation between X and Y.

Let us prove this result as follows.
Since X and Y are standardised random variables, we have Expectation of X is equal to Expectation of Y is equal to zero.

Variance of X is equal to Variance of Y is equal to 1 implies, Expectation of X square is equal to Expectation of Y square is equal to 1.
And covariance of X Y is equal to Expectation of X into Y Implies, Expectation of X into Y is equal to r of X, Y into sigma X into sigma Y is equal to r of X, Y.

Also, we can write r of a into X plus b into Y and b into X plus a into Y Is equal to Expectation of a into X plus b into Y into b into X plus a into Y minus Expectation of a into X plus b into Y into Expectation of b into X plus a into Y Whole divided by Variance of a into X plus b into Y into Variance of b into X plus a into Y the whole power half.

On multiplying the first two expression and substituting the values of Expectation for next two expressions and using the property of variance in the denominator, we get, Expectation of a into b into X square plus a square into X into Y plus b square into Y into X plus a into b into Y square minus zero whole divided by a square into variance of X plus b square variance of Y plus 2 into a into b into Covariance of X, Y into b square into Variance of X plus a square into variance of Y plus 2 into b into a into covariance of x, Y whole power half Substituting for the values of expectation and variance, we get, a into b into 1 plus a square into r of X, Y plus b square into r of X, Y plus a into b into 1 whole divided by a square plus b square plus 2 into a into b into r of X, Y into b square plus a square plus 2 into b into a into r of X, Y whole power half Is equal to 2 into a into b plus a square plus b square into r of X, Y whole divided by a square plus b square plus 2 into a into b into r of X, Y.

Given that r of a into X plus b into Y and b into x plus a into Y is equal to 1 plus 2 into a into b whole divided by a square plus b square.

Hence, equating right hand sides of both the expressions, we get, 1 plus 2 into a into b divided by a square plus b square Is equal to 2 into a into b plus a square plus b square into r of X, Y whole divided by a square plus b square plus 2 into a into b into r of X, Y Implies, 1 plus 2 into a into b into a square plus b square plus 2 into a into b into r of X Y is equal to a square plus b square into 2 into a into b plus a square plus b square into r of X, Y Implies, 1 plus 2 into a into b into a square plus b square plus 2 into a into b into r of X, Y into 1 plus 2 into a into b Is equal to a square plus b square the whole square into r of X, Y plus 2 into a into b into a square plus b square.
Implies, a power 4 plus b power 4 plus 2 into a square into b square minus 2 into a into b minus 4 into a square into b square into r of X, Y is equal to a square plus b square.
Implies, r of X, Y is equal to a square plus b square divided by a square minus b square the whole square minus 2 into a into b.

Following is the importance of correlation analysis:

1. Most of the variables in economics and business show relationship. For example, price and supply, income and expenditure etc. The correlation analysis helps the investigator to know the degree and direction of such relationship between variables. These days, correlation analysis finds application in various fields including the field of life science.
2. Once the correlation is established between the two variables, regression analysis helps us to estimate value of dependent variable for the given value of independent variable.

3. Correlation analysis together with regression analysis helps us to understand the behavior of various social and economic variables.
4. The effect of correlation is to reduce the range of uncertainty in our predictions.

5. Merits and Demerits of Correlation Coefficient

Following are the merits of correlation coefficient:

1. It gives a precise quantitative value indicating the degree of relationship existing between the two variables.
2. It measures the direction as well as relationship between the two variables.
3. Further, in regression analysis it is used for estimating the value of dependent variable from the known value of the independent variable.

Following are demerits of correlation coefficient:

1. Extreme items affect the value of the coefficient of correlation.
2. Its computational method is difficult as compared to other methods.
3. It assumes the linear relationship between the two variables, whether such relationship exist or not.

Let the random variable have marginal density f of (x) is equal to 1, where minus half less than x less than half And the conditional density of Y be f of y given x is equal to 1, where x less than y less than x plus 1 and minus half less than x less than zero Is equal to 1, where minus x less than y less than 1 minus x and zero less than x less than half. Show that variables X and Y are uncorrelated.

Let us solve the above problem as follows:

We have Expectation of X is equal to integral from minus half to half x into f of x dx Is equal to integral from minus half to half x into 1 dx is equal to zero. If f of x, y is the joint probability density function of X and Y then, f of x, y is equal to f of x into f of y given X is equal to f of y given x, since f of x is equal to 1.

Now, let us find expectation of X, Y. Expectation of X Y is equal to integral from minus half to zero, integral from x to x plus 1, x into y into dx, dy plus integral from zero to half, integral from minus x to 1 minus x, x into y dx, dy Is equal to integral from minus half to zero x into integral from x to x plus 1 y dy dx plus integral from zero to half x into integral from minus x to 1 minus x, y dy dx Is equal to integral from minus half to zero, x into y square divided by 2, ranges from x to x plus 1 dx plus integral from zero to half, x into y square divided by 2, ranges from minus x to 1 minus x dx.

Is equal to half into integral from minus half to zero, x into 2 into x plus 1 dx plus half into integral from zero to half x into 1 minus 2 into x dx is equal to zero.

Therefore, covariance of X, Y is equal to Expectation of X into Y minus Expectation of X into Expectation of Y is equal to zero Implies, r of X, Y is equal to zero. Hence, the variables are uncorrelated.

What we have learned in this session, where in we have understood about:

- The definition of Correlation
- The correlation and causation
- The interpretation of the value of correlation coefficient
- The properties of correlation coefficient
- The importance of correlation coefficient
- The merits and demerits of correlation coefficient