1. Introduction

Welcome to the series of E-learning modules on product moment correlation coefficient, characteristics, assumptions, merits, demerits and its properties.

By the end of this session, you will be able to:

- Explain about the product moment correlation coefficient
- Explain the characteristics and assumptions of product moment correlation coefficient
- Explain their merits and demerits
- Explain the properties

In the last module, we have identified whether there is a relationship or not between the variables using scatter diagram. These diagrams only give ideas about the relation, but do not measure the degree of correlation.

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson, a British Biometrician, developed a formula called correlation coefficient, which is based on moments. Hence, it is called product moment correlation coefficient or Karl Pearson's correlation coefficient.
Correlation coefficient between two random variables X and Y is usually denoted by r of X Y or r X Y. It is a numerical measure of linear relationship between them and is defined as r X Y is equal to covariance of X Y divided by sigma X into sigma Y.

If Xi Yi is the bivariate distribution, where i is equal to 1, 2 up to n then,
Covariance of X, Y is equal to expectation of X minus E of X into Y minus E of Y
Is equal to 1 by n into summation x i minus x bar into y i minus y bar
Sigma X square is equal to expectation of X minus E of X whole square
Is equal to 1 by n into summation x i minus x bar whole square.
Sigma Y square is equal to expectation of Y minus E of Y whole square
Is equal to 1 by n into summation y i minus y bar whole square.
And the summation ranges over i from 1 to n.

Another convenient form of the formula for correlation is as follows:
r X Y is equal to summation x i minus x bar into y i minus y bar divided by square root of summation x i minus x bar whole square into summation y i minus y bar whole square.

On simplification we get,
r X Y is equal to n into summation x i into y i minus summation x i into summation y i divided by square root of n into summation x i square minus summation x i the whole square into n into summation y i square minus summation y i the whole square. We use this formula for raw data. Here n is number of observations.

Suppose if we have tabulated data, then r X Y is equal to n into summation f i into x i into y i minus summation fi into xi into summation fi into yi divided by square root of n into summation f i into x i square minus summation f i x i whole square into n into summation f i into y i square minus summation f i into y i whole square. Here n is equal to total frequency.

2. Characteristics of Product Moment Correlation Coefficient

Following are the main characteristics of product moment correlation coefficient.

- Based on arithmetic mean and standard deviation – The formula is based upon arithmetic mean and standard deviation. The products of the corresponding values of the two series that is, co-variance is divided by the product of standard deviations of the two series to determine the formula

- Determines the direction of relationship - Karl Pearson's method establishes the direction of relationship of variables, namely positive or negative

- Establishes the size of relationship – This method also shows the size of relationship between variables of the two series. It ranges between plus one and minus one. Plus one means perfect positive correlation and minus one means perfect negative relationship. In case the value is zero, then it means no relationship between the variables

- Ideal measure – This method is considered to be an ideal method of calculation of correlation coefficient. It is because of the covariance which is most reliable as a standard statistical tool

Coefficient of correlation of Product Moment is based on the following assumptions.

- The relationship is linear – We assume that there is a linear relationship between the two variables. That means if the two variables are plotted, we get a straight line

- Normal distribution – A large number of independent causes are operating in both the variables correlated so as to produce a normal distribution.

- Related in a Casual Function – This means that the forces operating on each of the variable series are not independent of each other but are related in a causal fashion. In other words, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between:

a) The series of heights and incomes of individuals over a period of time
b) The series of marriage rate and the rate of agricultural production in a country over a period of time.
c) The series relating to the size of the shoe and intelligence of a group of individuals.

Should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

3. Merits & Limitations of Product Moment correlation coefficient

The Product Moment correlation coefficient has the following merits:

- Counts all values: It takes into account all values of the given data of x and y. Therefore, it is based on all observations of the series
- More Practical and Popular: Product moment correlation coefficient ‘r’ is considered to be more practical method as compared to other mathematical methods us
- ed for ‘r’. It is also very popular and commonly used method.
- Numerical Measurement of ‘r’: It provides numerical measurement of coefficient of correlation
- Measures Degree and Direction: This method measures both degree and direction of the correlation between the variables at a time
- Facilitates Comparison: Product Moment coefficient of correlation is a pure number independent of units. Therefore, the comparison between the series can be done easily
- Algebraic Treatment Possible: Product Moment coefficient of correlation techniques can easily be applied for higher algebraic treatment

The use of coefficient of correlation has certain limitations which are as under:

- Linear relationship: Coefficient of correlation assumes linear relationship between the variables regardless of the fact whether that assumption is correct or not
- More time consuming: Compared with some other methods, this method is time consuming
- Affected by extreme items: Another drawback of coefficient of correlation is that it is affected by the extreme items
- Difficult to interpret: It is not easy to interpret the significance of correlation coefficient. It is generally misinterpreted

Two independent variables are uncorrelated but the converse of the above statement is not true.
We prove using the expression of r X Y and the converse is proved by using an example.
If X and Y are independent variables, then Covariance of X Y is equal to zero
Implies, r X Y is equal to covariance of X Y divided by sigma X into sigma Y is equal to zero.
Hence, two independent variables are uncorrelated.

To prove that the converse is not true, that is uncorrelated variables may not be independent as the following example illustrates.

Figure 1

X	Y	XY
-3	9	-27
-2	4	-8
-1	1	-1
1	1	1
2	4	-8
3	9	27
0	**25**	**0**

Consider the following table, where X takes values minus 3, minus 2 and so on. Y takes values 9, 4 and so on. If we find product of these two variables, we get minus 27, minus 8 and so on. Totals of all the columns are denoted by the bold numbers.
Let us find mean X bar is equal to 1 divided by n into summation X is equal to zero.

Covariance of X, Y is equal to 1 divided by n into summation X into Y minus X bar into Y bar is equal to zero.

That is r X Y is equal to covariance of X Y divided by sigma X into sigma Y is equal to zero.

Thus, in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y=X^2$. Hence two uncorrelated variables need not necessarily be independent.

A simple reasoning for this strange conclusion is that r X Y is equal to zero, merely implies the absence of any linear relationship between the variables X and y. There may however, exist some other form of relationship between them. For example, quadratic, cubic or trigonometric.

Even though, in general the converse statement may not hold. But in the following cases, the converse statement that is two uncorrelated variables may be independent holds in the following cases.

If X and Y are jointly normally distributed with row is equal to zero, then they are independent. If row is equal to zero then f of x y is equal to 1 by sigma x into square root of 2 into pi into exponential of minus half into x minus mew x divided by sigma x whole square into 1 by sigma y into square root of 2 into pi into exponential of minus half into y minus mew y divided by sigma y whole square is equal to f 1 of x into f 2 of y.
Hence, X and Y are independent.

1. If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then r of X Y is equal to zero implies X and Y are independent.

We prove the above statement as follows:
Let X take values 1 and zero with positive probabilities p 1 and q 1 respectively and let Y take the values 1 and zero with positive probabilities p 2 and q 2 respectively. Then,
r of X, Y is equal to zero implies, covariance of X Y is equal to zero.

Implies zero is equal to expectation of X into Y minus expectation of X into expectation of Y
Is equal to 1 into probability of X is equal to 1 intersection Y is equal to 1 minus 1 into probability of X is equal to 1 into 1 into probability of Y is equal to 1
Is equal to p of x is equal to 1 intersection Y is equal to 1 minus p 1 into p 2

Implies, probability of X is equal to 1 intersection Y is equal to 1 is equal to p 1 into p 2 is equal to probability of X is equal to 1 into Probability of Y is equal to 1.
Hence, X and Y are independent.

4. Properties of Product Moment Correlation Coefficient- Property 1&2

Now let us state and prove the properties of product moment correlation coefficient.

The first property says that,
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 plus 1.

To prove this property, we consider the expression of correlation coefficient
r X Y is equal to 1 by n into summation x i minus x bar into y i minus y bar divided by square root of 1 by n into summation x i minus x bar whole square into 1 by n into summation y i minus y bar whole square.

That is r X Y square is equal to summation a i into b i whole square divided by summation a i square into summation b i square.
Where a i is equal to x i minus x bar and b i is equal to y i minus y bar.

From Schwartz inequality, which states that if a i b i is equal to 1, 2, up to n are real quantities, then summation a i into b i whole square is less than or equal to summation a i square into summation b i square.
The sign of equality holding if a 1 by b 1 is equal to a 2 by b 2 is equal to .... is equal to a n by b n.
Using Schwartz inequality, we get, r X Y is less than or equal to 1. That is, modulus of r X Y is less than or equal to minus 1, implies, minus 1 less than or equal to r is less than or equal to plus 1.
Hence the proof.

Second property says that,
Correlation coefficient is independent of change of origin and scale.

To prove this property, consider the transformation, U is equal to X minus a divided by h and V is equal to y minus b divided by k, so that X is equal to a plus h into U and Y is equal to b plus k into V, where a, b, h and k are constants and h and k are positive.

We shall prove that r X Y is equal to r U V.
Since x is equal to a plus h into U and Y is equal to b plus k into V.

On taking expectations, we get,
Expectation of X is equal to a plus h into expectation of U and expectation of Y is equal to b plus k into expectation of V.

Therefore, X minus E of X is equal to h into U minus Expectation of U and

Y minus Expectation of Y is equal to k into V minus Expectation of V.

Implies covariance of X Y is equal to Expectation of X minus E of X into Y minus E of Y is equal to h into U minus E of U into k into V minus E of V
Is equal to h into K into Expectation of U minus E of U into V minus E of V is equal to h into k into covariance of U V.

Sigma X square is equal to Expectation of X minus E of X the whole square is equal to Expectation of h square into U minus E of U whole square is equal to h square into sigma U square
Implies sigma X is equal to h into sigma U, where h is positive.
Sigma Y square is equal to expectation of Y minus E of Y whole square is equal to k square into V minus E of V whole square is equal to k square into sigma V square.
Implies sigma Y is equal to k into sigma V where k is positive.

Substituting in r of X Y is equal to covariance of X Y divided by sigma X into sigma Y is equal to h into k into covariance of U, V divided by h into k into sigma U into sigma V is equal to covariance of U, V divided by sigma U into sigma V is equal to r of U, V.

5. Properties of Product Moment Correlation Coefficient- Property 3 & Others

The third property says that, If X and Y are random variables and a, b, c, d are any numbers provided, that a is not equal to zero and c is not equal to zero then r of a X plus b c Y plus d is equal to a into c divided by modulus of a into c into r of X, Y.

We prove this property as follows:
Variance of a into X plus b is equal to a square into sigma X square and variance of c into Y plus d is equal to c square into sigma Y square.

Covariance of a X plus b and c into Y plus d is equal to a into c into sigma X Y.

Therefore,
r of a into X plus b c into Y plus d is equal to covariance of a into X plus b c into Y plus d divided by square root of variance of a into x plus b into variance of c into Y pus d

is equal to a into c sigma X Y divided by modulus of a into modulus of c into sigma X into sigma Y is equal to a into c divided by modulus of a into c into r of X, Y
If ac is greater than zero, that is if a and c are of same signs, then a into c divided by modulus of a into c is equal to plus 1.

If ac is less than zero, that is if a and c are of opposite signs, then a into c divided by modulus of a into c is equal to minus 1.

Some of the other properties of product moment correlation coefficient are:

- Coefficient of correlation is based on a suitable measure of variation as it takes into account all items of the variable
- If there is an accidental correlation, in that case the coefficient of correlation might lead to fallacious conclusions. It is known as non-sense or spurious correlation
- Coefficient of correlation works both ways

i.e., $r_{XY} = r_{YX}$

The coefficient of correlation does not prove causation but it is simply a measure of co-variation. It is because variations in X and Y series may be due to,

1. Some common cause
2. Some mutual dependence
3. Some change and
4. Some causation of the subject to be relative.

Coefficient of correlation is independent of the unit of measurement.

Here’s a summary of our learning in this session, where we have understood:

- About the product moment correlation coefficient
- The characteristics and assumptions of product moment correlation coefficient
- Merits and demerits of product moment correlation coefficient
- Properties of product moment correlation coefficient