# 1. Introduction

Welcome to the series of E-learning modules on practical problems on product moment correlation coefficient. Here we find the value of product moment correlation coefficient for both raw and tabulated data. We also interpret the value of the coefficient using probable error and coefficient of determination.

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

- Find the value of product moment correlation coefficient – for both raw and tabulated data
- Interpret data using probable error and coefficient of determination

In general the formula for calculating Product Moment or Karl Pearson's coefficient of correlations is as follows: r is equal to covariance of X, Y divided by square root of variance of X into variance of Y is equal to covariance of X, and Y divided by standard deviation of X into standard deviation of Y.
Which is same as, summation x minus x bar into y minus y bar divided by square root of summation x minus x bar whole square into summation y minus y bar whole square.
For practical purpose, we simplify the above equation and use the following formulae

For raw data,
r is equal to n into summation x into y minus summation x into summation y whole divided by square root of n into summation x square minus summation x the whole square into n into summation y square minus summation y the whole square.
And for tabulated data r is equal to N into summation f into x into y minus summation f into x into summation f into y whole divided by square root of N into summation f into x square minus summation f into x the whole square into N into summation f into y square minus summation f into y the whole square.

To interpret the value of r, we can either use probable error or coefficient of determination.
If we use probable error, it is given by,
PE is equal to zero point six, seven, four, five into one minus r square divided by square root of n

- If the value of r is less than the probable error, there is no evidence for correlation
- If the value of r is more than six times of the probable error, it is significant correlation

But it can be used only when n is very large. Otherwise we go for coefficient of determination, which is given by r square.

# 2. Exercises (Part 1)

Exercise 1
In a bivariate data on X and Y, Variance of X is equal to forty nine, variance of Y is equal to nine and covariance of X and Y is equal to minus seventeen point five. Find the coefficient of correlation between X and Y.
Solution :
The coefficient of correlation is
r is equal to covariance of x ad y divided by square root of variance of X into variance of Y is equal to minus seventeen point five divided by square root of forty nine into nine is equal to zero point eight, three, three.

Exercise 2
Variables X and Y are perfectly negatively correlated. Also standard deviation of X is equal to four point six and standard deviation of Y is equal to zero point four, seven. Find Covariance of X, Y
Solution :
Since X and Y are perfectly negatively correlated, r is equal to minus one
r is equal to covariance of x ad y divided by standard deviation of X into standard deviation of Y
Implies covariance of X Y is equal to r into standard deviation of X into standard deviation of Y.
Is equal to minus one into four point six into zero point four, seven is equal to minus two point one, six, two.

Exercise 3
Now consider the following exercises.
Calculate the Karl Pearson's coefficient of correlation between the following two series regarding the age of Husband and wife in years.
The table gives the ages of 11 couples.

**Figure 1**

| Husband's Age | 24 | 27 | 28 | 28 | 29 | 30 | 32 | 33 | 35 | 35 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Wife's Age | 18 | 20 | 22 | 25 | 22 | 28 | 28 | 30 | 27 | 30 | 22 |

Solution :
Let x denote the age of Husband and y denote the age of wife

Now let us consider the following table.

**Figure 2**

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 24 | 18 | 576 | 324 | 432 |
| 27 | 20 | 729 | 400 | 540 |
| 28 | 22 | 784 | 484 | 616 |
| 28 | 25 | 784 | 625 | 700 |
| 29 | 22 | 841 | 484 | 638 |
| 30 | 28 | 900 | 784 | 840 |
| 32 | 28 | 1024 | 784 | 896 |
| 33 | 30 | 1089 | 900 | 990 |
| 35 | 27 | 1225 | 729 | 945 |
| 35 | 30 | 1225 | 900 | 1050 |
| 40 | 22 | 1600 | 484 | 880 |
| **341** | **272** | **10777** | **6898** | **8527** |

The first two columns are written as it is in the question.
The third column is obtained by the squaring the entries in the first column and fourth column are obtained by squaring the entries in the second column. The last column is obtained by multiplying the entries in the first and second columns.
The last row which is written in bold gives the sum of each column.

The coefficient of correlation between x and y is,
r is equal to n into summation x into y minus summation x into summation y whole divided by square root of n into summation x square minus summation x the whole square into n into summation y square minus summation y the whole square.

From the table in the last slide,
n is equal to eleven, summation x is equal to three hundred and forty one, summation y is equal to two hundred and seventy two, summation x square is equal to ten thousand seven hundred and seventy seven, summation y square is equal to 6 thousand eight hundred and ninety eight and summation x into y is equal to eight thousand five hundred and twenty seven.
Substituting these values in the above expression we get,
r is equal to eleven into eight thousand five hundred and twenty seven minus three hundred and forty one into two hundred and seventy two whole divided by square root of eleven into ten thousand seven hundred and seventy seven minus three hundred and forty one the whole square into eleven into 6 thousand eight hundred and ninety eight minus two hundred and seventy two the whole square.
Is equal to zero point five zero four.

To interpret r, we use coefficient of determination.
Coefficient of determination is equal to r square is equal to zero point five zero four square is equal to zero point two, five, four.
Hence only twenty five point four percent of the variation in one variable is explained by the other. Hence there is weak correlation between the two variables.

# 3. Exercises (Part 2)

Exercise 4:
The following table gives the distribution of total population and those who are wholly or partially blind among them.

**Figure 3**

| Age | No. of persons(in '000) | Blind |
|---|---|---|
| 0-10 | 100 | 55 |
| 10-20 | 60 | 40 |
| 20-30 | 40 | 40 |
| 30-40 | 36 | 40 |
| 40-50 | 24 | 36 |
| 50-60 | 11 | 22 |
| 60-70 | 6 | 18 |
| 70-80 | 3 | 15 |

Find out if there is any relation between age and blindness.

Solution :
As population of each age group is different, to facilitate comparison it is required to make the number of blinds per equal population and let it be per one lakh.
That is for the first age group, zero to ten, number of persons is hundred thousand and number of blind people is fifty five. Therefore number of blind people per lakh is given by,
Fifty five divided by one lakh into one lakh is equal to fifty five.

In the second age group, ten to twenty, number of persons is sixty thousand and number of blind people is forty. Therefore the number blinds per lakh is given by,
Forty divided by sixty thousand into one lakh is equal to sixty seven.
Like this we calculate for all the age groups and let it be denoted by ‘y’ and since ages are given in terms of class intervals, we take mid values of the class intervals as x.

Now let us construct the following table .

**Figure 4**

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 5 | 55 | 25 | 3025 | 275 |
| 15 | 67 | 225 | 4489 | 1005 |
| 25 | 100 | 625 | 10000 | 2500 |
| 35 | 111 | 1225 | 12321 | 3885 |
| 45 | 150 | 2025 | 22500 | 6750 |
| 55 | 200 | 3025 | 40000 | 11000 |
| 65 | 300 | 4225 | 90000 | 19500 |
| 75 | 500 | 5625 | 250000 | 37500 |
| **320** | **1483** | **17000** | **432335** | **82415** |

The first column denotes the mid values of the age groups. The second column is calculated as number of blinds per one lakh people as explained above.
The third column is obtained by the squaring the entries in the first column and fourth column are obtained by squaring the entries in the second column. The last column is obtained by multiplying the entries in the first and second columns.
The last row which is written in bold gives the sum of each column.

Therefore the coefficient of correlation between x and y is,
r is equal to n into summation x into y minus summation x into summation y whole divided by square root of n into summation x square minus summation x the whole square into n into summation y square minus summation y the whole square.

From the table in the last slide,
n is equal to eight, summation x is equal to three hundred and twenty, summation y is equal to one thousand four hundred and eighty three, summation x square is equal to seventeen thousand, summation y square is equal to four lakh thirty two thousand three hundred and thirty five and summation x into y is equal to eighty two thousand four hundred and fifteen.
Therefore r is equal to eight into eighty two thousand four hundred and fifteen minus three hundred and twenty into one thousand four hundred and eight three whole divided by square root of eight into seventeen thousand minus 320 the whole square into eight into four lakh thirty two thousand three hundred and thirty five minus one thousand four hundred and eight three the whole square.

Which is equal to zero point eight, nine eight.

To interpret r, we use coefficient of determination.
Coefficient of determination, r square is equal to zero point eight, nine, eight square is equal to zero point eight zero seven. Hence eighty point seven percent of the variation in one variable is explained by the other. Hence there is a strong correlation between the two variables. That is, as age increases blindness also increases.

# 4. Exercises (Part 3)

Exercise 5
From the following data find out whether there is any relationship between density of population and death rate.

**Figure 5**

| Zones | Area(Sq. Km.) | Population | No. of Deaths. |
|---|---|---|---|
| 1 | 200 | 40,000 | 480 |
| 2 | 150 | 75,000 | 1200 |
| 3 | 120 | 72,000 | 1080 |
| 4 | 80 | 20,000 | 280 |

Solution
In this question we have to find out the relationship between the density of population and death rate, which are not given directly. First of all we shall obtain the density of population and death rate.
To find these we use the following formulae.
Density of population is equal to Population divided by Area. And
Death rate is equal to number of deaths divided by population into one thousand.
Now let us denote density of population as x and death rate as y.

To calculate correlation coefficient we construct the following table.

**Figure 6**

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 200 | 12 | 40000 | 144 | 2400 |
| 500 | 16 | 25000 | 256 | 8000 |
| 600 | 15 | 36000 | 225 | 9000 |
| 250 | 14 | 62500 | 196 | 3500 |
| **1550** | **57** | **712500** | **821** | **22900** |

The first column gives the density of the population which is calculated by dividing population by area.
The second column gives the death rate which is obtained by dividing the number of deaths by population and then multiplying by one thousand.
The third column is obtained by the squaring the entries in the first column and fourth column are obtained by squaring the entries in the second column. The last column is obtained by multiplying the entries in the first and second columns.

The last row which is written in bold gives the sum of each column.

Therefore the coefficient of correlation between x and y is,
r is equal to n into summation x into y minus summation x into summation y whole divided by square root of n into summation x square minus summation x the whole square into n into summation y square minus summation y the whole square.

From the table in the last slide,
n is equal to four, summation x is equal to one thousand five hundred and fifty, summation y is equal to fifty seven, summation x square is equal to seven lakh twelve thousand five hundred, summation y square is equal to eight hundred and twenty one and summation x into y is equal to twenty two thousand nine hundred.

Therefore r is equal to four into twenty two thousand nine hundred minus one thousand five hundred and fifty into fifty seven whole divided by square root of four into seven lakh twelve thousand five hundred minus one thousand five hundred and fifty the whole square into four into eight hundred and twenty one minus fifty seven the whole square.
Is equal to zero point eight, two, one

To interpret r, we use coefficient of determination.
Coefficient of determination is equal to r square is equal to zero point eight, two, one square is equal to zero point six, seven, four.

Hence sixty seven point four percent of the variation in one variable is explained by the other. Hence there is correlation between the two variables. That is, there is positive association between density of population and death rate.

Exercise 6
A student while calculating the coefficient of correlation between two variates X and Y from twenty five pairs of observations obtained the following constants.
n is equal to twenty five, summation x into y is equal to five hundred and eighth, summation x is equal to one hundred twenty five, summation y is equal to one hundred, summation x square is equal to six hundred and fifty; summation y square is equal to four hundred and sixty.
It was, however detected later on at the time of checking that he had copied down the two pairs of observations, (x,y) is equal to (eight, twelve) and (six, eight) as (six, fourteen) and (eight, six). Obtain the correct value of the coefficient of correlation.

Solution
First we correct the given totals by subtracting the wrong entries and adding the correct entries.
Correct value of summation x into y is equal to five hundred and eight minus (six into fourteen) minus (eight into six) plus (eight into twelve) plus (six into eight) is equal to five hundred and twenty
Correct value of summation x is equal to one hundred and twenty five minus six minus eight plus eight plus six is equal to one hundred and twenty five
Correct value of summation y is equal to one hundred minus fourteen minus six plus twelve plus eight is equal to one hundred
Correct value of summation x square is equal to six hundred and fifty minus six square minus

eight square  plus eight square  plus six square is equal to six hundred and fifty
Correct value of summation y square is equal to four hundred and sixty minus fourteen square minus six square plus twelve square plus eight square is equal to four hundred and thirty six

Therefore the correct value of coefficient of correlation is given by,
r is equal to n into summation x into y minus summation x into summation y whole divided by square root of n into summation x square minus summation x the whole square into n into summation y square minus summation y the whole square.
By substituting the corrected values,
r is equal to twenty five into five hundred and twenty minus one hundred and twenty five into hundred whole divided by square root of twenty five into six hundred and fifty minus one hundred and twenty five the whole square into twenty five into summation four hundred and thirty six minus hundred the whole square.
Is equal to zero point six, seven.

# 5. Exercises (Part 4)

Exercise 7
A student calculates the value of r as zero point seven when the number of observations used is twenty five and concludes that r is highly significant. Is he correct?

Solution:
For testing the significance we shall calculate the probable error as follows.
PE is equal to zero point six, seven, four, five into one minus r square divided by square root of n
Is equal to zero point six, seven, four, five into one minus zero point seven square divided by square root of twenty five
Is equal to zero point zero six, eight, eight
Six times Probable Error is equal to six into (zero point zero six, eight, eight) is equal to zero point four, one, two, eight, which is less than r. Therefore the student is correct.

Exercise 8
The following is the joint distribution of age of brides and bride-grooms. Calculate the product moment coefficient of correlation.

**Figure 7**

| Age of bride-groom | Age of Bride | | | | |
|---|---|---|---|---|---|
| | 18-20 | 20-22 | 22-24 | 24-26 | 26-28 |
| 20-23 | 7 | 6 | 1 | - | - |
| 23-26 | 3 | 8 | 6 | 4 | 8 |
| 26-29 | 1 | 2 | 3 | 8 | 8 |
| 29-32 | - | 1 | 1 | 1 | 2 |

Solution :
Let X denotes age of bride-groom and Y denotes age of bride.
Since we have been given class intervals, we consider x as mid-point of the class interval corresponding to X and y as the mid-point of the class interval corresponding to Y.
r is equal to N into summation f into x into y minus summation f into x into summation f into y divided by square root of N into summation f into x square minus summation f into x the whole square into N into summation f into y square minus summation f into y the whole square.

First we write the given values in the table as it is. Then we find different columns which are required.

**Figure 8**

| Age of bridge-groom | Age of Bride | | | | | f | x | u | fu | $fu^2$ | fuv |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 18-20 | 20-22 | 22-24 | 24-26 | 26-28 | | | | | | |
| 20-23 | 7 (0) | 6 (0) | 1 (0) | - | - | **14** | 21.5 | 0 | 0 | 0 | 0 |
| 23-26 | 3(0) | 8 (8) | 6 (12) | 4(12) | 8(32) | **29** | 24.5 | 1 | 29 | 29 | 64 |
| 26-29 | 1 (0) | 2(4) | 3(12) | 8(48) | 8(64) | **22** | 27.5 | 2 | 44 | 88 | 12 |
| 29-32 | - | 1(3) | 1(6) | 1(9) | 2(24) | **5** | 30.5 | 3 | 15 | 45 | 42 |
| **f** | **11** | **17** | **11** | **13** | **18** | **N=70** | | | **88** | **162** | **234** |
| y | 19 | 21 | 23 | 25 | 27 | | | | | | |
| v | 0 | 1 | 2 | 3 | 4 | | | | | | |
| fv | 0 | 17 | 22 | 39 | 72 | **150** | | | | | |
| $fv^2$ | 0 | 17 | 44 | 117 | 288 | **466** | | | | | |
| fuv | 0 | 15 | 30 | 69 | 120 | **234** | | | | | |

Here we consider u is equal to (x minus twenty one point five) divided by three and v is equal to (y minus nineteen) divided by two. Hence we construct the following table which gives the necessary totals to substitute in the formula of product moment correlation coefficient.
The numbers written in brackets give the values of (f into u into v) for each cell. By adding these values row wise and column wise we write the totals in the last row and last column. Hence we get total of last row and last column same, that is, two hundred and thirty four.
The numbers written in yellow, give the totals of various rows and columns.

Observe that summation f into u is equal to eighty eight, summation f into u square is equal to one hundred and sixty two, summation f into v is equal to one hundred and fifty, summation f into v square is equal to four hundred and sixty six, and summation f into u into v is equal to two hundred and thirty four.

Further we know that correlation coefficient is independent of change of origin and scale. We can write the formula as,
r is equal to N into summation f into u into v minus summation f into u into summation f into v divided by square root of N into summation f into u square minus summation f into u the whole square into N into summation f into v square minus summation f into v the whole square.

Hence by substituting these values in the formula, we get,
r is equal to seventy into two hundred and thirty four minus eighty eight into one hundred and fifty divided by seventy into one hundred and sixty two minus eighty eight square into seventy into four hundred and sixty six minus one hundred and fifty square is equal to zero point five, two, seven, one.

Since N is large, we can use probable error for interpreting the data.
Probable error PE is equal to zero point six, seven, four, five into one minus r square divided by square root of n
Is equal to zero point six, seven, four, five into one minus zero point five, two, seven, one square divided by square root of seventy .
Is equal to zero point zero eight, six, three.
six times probable error is (six into zero point zero eight, six, three) is equal to zero point five, one, seven, nine, which is less than r. Therefore there is significant correlation between age of bride-groom and bride.

Here's a summary of our learning in this session where we have understood how to:

- Find the value of product moment correlation coefficient – for both raw and tabulated data
- Interpret data using probable error and coefficient of determination