Coefficient of Correlation

***Population Correlation Coefficient:***

1. The measure of joint or mutual variation in a bivariate population with two variables *x* and *y*, is called *‘covariance of x and y’*:



1. In order to make comparison, the covariance must be standardised by dividing (*x* – *μx*) and (*y* – *μy*) by their SDs *σx* and *σy* respectively. This expression is called *‘coefficient of correlation’*; the *‘population coefficient of correlation’* is denoted by *‘ρ’* (rho):



***Sample Correlation Coefficient:***

1. The sample covariance of *x* and *y*, *Sxy*, measures the tendency for *x* and *y* to increase or decrease together in the sample:



1. The *‘sample coefficient of correlation’* is denoted by *‘r’*. It is also known as *‘Karl Pearson’s product moment coefficient of correlation’*. The coefficient of correlation always lies between –1 and +1 respectively, i.e., *–1 ≤ r ≤ +1:*





1. (a) If *r* = –1, all the points on the scatter diagram lie on the regression line of negative slope. It is called a *‘perfect negative correlation’*.

(b) If *r* = 1, all the points on the scatter diagram lie on the regression line of positive slope. It is called a *‘perfect positive correlation’*.

(c) If *r* = 0, all the points on the scatter diagram are spread throughout the diagram indicating no correlation between *x* and *y*.

“Correlation coefficient is a measure of the closeness of linear relationship between the two variables.”

X

Y

*r* = 1

(a) Perfect positive (or direct) correlation

0

X

Y

0 < *r* ≤ 1

(b) Positive Correlation

0

–1 ≤ *r* < 0

X

Y

(d) Negative Correlation

0

X

Y

*r* = 0

(c) No correlation

0

X

Y

(e) Perfect negative (or inverse) correlation

*r* = –1

0

***Correlation Coefficient and Regression Coefficient:***

1. The two regression coefficients *b* and *d* of the two regression lines can also be stated as follows:



1. Since , therefore, *Sxy = r ∙ Sx ∙ Sy.*
2. The regression coefficients *b* and are related to correlation coefficient *r* by:



or

or



Where 





***Properties of Coefficient of Correlation:***

1. The correlation coefficient is symmetrical with respect to *x* and *y*, i.e., *rxy = ryx*
2. The correlation coefficient is the geometric mean of the two regression coefficients, i.e.: .
3. The correlation coefficient is a pure number and does not depend upon the units employed. For e.g., if the correlation coefficient between the heights and weights of students is computed as 0.98, it will be expressed simply as 0.98 (neither as 0.98 inches nor 0.98 pounds).
4. The correlation coefficient is independent of origin and unit of measurement. By this we mean that if we take deviations of *x* and *y* from some suitable origins or transform *x* and *y* into *u* and *v* respectively, it will not affect the correlation coefficient. Symbolically:

*rxy = ruv*

1. The correlation coefficient lies between –1 and +1, i.e., it cannot be less than –1 and greater than +1:

*–1 ≤ r ≤ +1*

**Example:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 3 | 1 | 1 | 2 | 4 | 2 | 3 | 5 | 2 | 3 |
| **y** | 2 | 4 | 3 | 2 | 1 | 2 | 1 | 3 | 2 | 1 |

*Required:*

1. Covariance of x and y,
2. Standard deviation of x and y,
3. Coefficient of correlation, and
4. Scatter diagram.

**Solution:**

**(a) Covariance of x and y:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***x*** | ***y*** | ***x – μx*** | ***x – μy*** | ***(x – μx)( x – μy)*** | ***(x – μx)2*** | ***(x – μy)2*** |
| 3 | 2 | 0.4 | –0.1 | –0.04 | 9 | 4 |
| 1 | 4 | –1.6 | 1.9 | –3.04 | 1 | 16 |
| 1 | 3 | –1.6 | 0.9 | –1.44 | 1 | 9 |
| 2 | 2 | –0.6 | –0.1 | 0.06 | 4 | 4 |
| 4 | 1 | 1.4 | –1.1 | –1.54 | 16 | 1 |
| 2 | 2 | –0.6 | –0.1 | 0.06 | 4 | 4 |
| 3 | 1 | 0.4 | –1.1 | –0.44 | 9 | 1 |
| 5 | 3 | 2.4 | 0.9 | 2.16 | 25 | 9 |
| 2 | 2 | –0.6 | –0.1 | 0.06 | 4 | 4 |
| 3 | 1 | 0.4 | –1.1 | –0.44 | 9 | 1 |
| 26 | 21 |  |  | –4.6 | 82 | 53 |







**(b) Standard deviation of x and y:**



**(c) Coefficient of correlation:**



**(d) Scatter diagram:**

Y

4

3

2

1

0

1 2 3 4 5 X

*ρ* = –0.0698

**Example:**

Calculate:

1. Covariance of x and y,
2. Variances of x and y,
3. Coefficient of correlation, and
4. Coefficient of determination.

For the following sample data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **x** | 1 | 2 | 4 | 6 | 8 | 10 | 14 | 15 | 18 | 20 |
| **y** | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |

**Solution:**

**(a) Covariance of x and y:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **x** | **y** |  |  | **()()** | **()2** | **()2** |
| 1 | 10 | –8.8 | –45 | 396 | 77.44 | 2025 |
| 2 | 20 | –7.8 | –35 | 273 | 60.84 | 1225 |
| 4 | 30 | –5.8 | –25 | 145 | 33.64 | 625 |
| 6 | 40 | –3.8 | –15 | 57 | 14.44 | 225 |
| 8 | 50 | –1.8 | –5 | 9 | 3.24 | 25 |
| 10 | 60 | 0.2 | 5 | 1 | 0.04 | 25 |
| 14 | 70 | 4.2 | 15 | 63 | 17.64 | 225 |
| 15 | 80 | 5.2 | 25 | 130 | 27.04 | 625 |
| 18 | 90 | 8.2 | 35 | 287 | 67.24 | 1225 |
| 20 | 100 | 10.2 | 45 | 459 | 104.04 | 2025 |
| 98 | 550 |  |  | 1820 | 405.6 | 8250 |





**(b) Variances of x and y:**



**(c) Coefficient of correlation:**



**(d) Coefficient of determination:**



r2 = b × d

r2 = 4.48720 × 0.22059 = 0.9898 = 98.98%

***Probable Error:***

1. The probable error is about two-third of the standard error:



1. Assuming *ρ* = 0, the sampling distribution of *r* has standard error:



1. In a standard normal distribution, *z* = ± 0.6745 will contain 50% of the area under curve, symbolically:

## P(–0.6745 ≤ z ≤ 0.6745) = 0.5

1. Thus, the probable error *r* is:

*P.E. = 0.6745 × σr*

or

*P.E. = 0.6745 × *

1. Probabilities of *r* can now be calculated using P.E. as a unit of deviation:

*P(–P.E. ≤ r ≤ P.E.) = 0.5*

*P(–3P.E. ≤ r ≤ 3P.E.) = 0.9544*

***Rank Correlation:***

1. If observations on two variables are given in the form of ranks rather than some numerical measurements, it is possible to compute a coefficient of correlation between ranks of the two variables. This correlation coefficient is called *‘Rank Correlation Coefficient’.*
2. As this formula was presented by Spearman in 1904, it is also known as *‘Spearman’s Rank Correlation Coefficient’*:



Where *di = xi – yi* (the difference between the rankings).

1. In order to test that there is no correlation between the two rankings, critical values of *rs* at *α* = 0.05 are given below:

|  |  |
| --- | --- |
| **Number of ranks (*n*)** | **Critical value (*rs*)** |
| 5 | 1.0 |
| 6 | 0.89 |
| 7 | 0.79 |
| 8 | 0.74 |
| 9 | 0.74 |
| 10 | 0.65 |
| 20 | 0.45 |
| 25 | 0.40 |
| 50 | 0.28 |

**Example:**

Ranks of 9 students in a class in History (x) and Geography (y) are as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Students | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** | **IX** |
| **x** | 1 | 9 | 7 | 4 | 5 | 3 | 8 | 2 | 6 |
| **y** | 4 | 5 | 6 | 3 | 7 | 2 | 8 | 1 | 9 |

Calculate Spearman’s Rank Correlation Coefficient and test its significance.

**Solution:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Students** | ***x*** | ***y*** | ***d = x – y*** | ***d2*** |
| I | 1 | 4 | –3 | 9 |
| II | 9 | 5 | 4 | 16 |
| III | 7 | 6 | 1 | 1 |
| IV | 4 | 3 | 1 | 1 |
| V | 5 | 7 | –2 | 4 |
| VI | 3 | 2 | 1 | 1 |
| VII | 8 | 8 | 0 | 0 |
| VIII | 2 | 1 | 1 | 1 |
| IX | 6 | 9 | –3 | 9 |
| Total | 45 | 45 | 0 | 42 |

 Where *di = xi – yi*

Critical value of *rs* for *n* = 9 and *α* = 0.05 is 0.74

Since 0.65 is less than the critical value of 0.74, *rs* is insignificant.