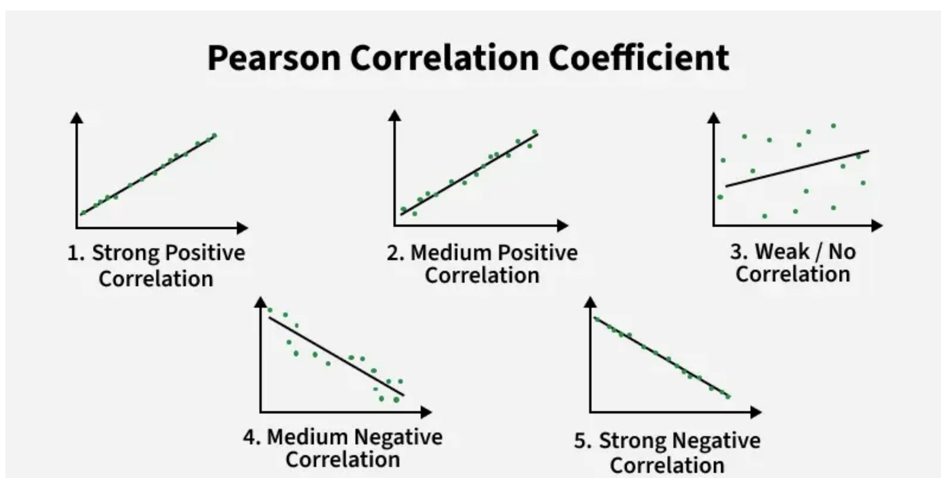
Pearson Correlation Coefficient

Pearson Correlation Coefficient (PCC) is used for measuring the strength and direction of a linear relationship between two variables. It is important in fields like data science, finance, healthcare, and social sciences, where understanding relationships between different factors is important. Quantifying the degree to which two variables are related helps analysts, researchers, and decision-makers to find patterns that make predictions and inform decisions.

Value of PCC(denoted by r) ranges from -1 to 1:

- 1 shows a perfect positive correlation where both variables increase or decrease together at a constant rate.
- -1 shows a perfect negative correlation where one variable increases as the other decreases proportionally.
- 0 shows no linear relationship means changes in one variable do not predict changes in the other.



Pearson Correlation Coefficient

> ***Pearson Correlation Coefficient** (**r**) is known by many names such as: Bivariate correlation, Pearson's r,*

> *Correlation coefficient, and Pearson product-moment correlation coefficient (PPMCC)*

The PCC is a type of descriptive statistic, meaning data is presented in a meaningful way using tables, graphs, etc.

Assumptions of Pearson Correlation Coefficient

1. **Linear Relationship**: Pearson's correlation assumes a linear relationship between the two variables. Non-linear relationships may not be accurately captured.
2. **Normality**: The variables should follow a normal distribution. This helps ensure that the correlation is meaningful and not affected by skewed data.
3. **Homoscedasticity**: The variability in one variable should remain consistent across all values of the other variable. This means the spread of data should be uniform.
4. **Interval or Ratio Scale**: Pearson's correlation is appropriate for interval or ratio data that are continuous and have consistent, meaningful numerical differences.
5. **Independence**: Observations should be independent of each other which means one data point should not influence another. Dependent data points can distort the correlation.

Pearson's Correlation Coefficient Formula and Interpretation

The formula for Pearson's correlation coefficient is shown below:

$$r = [n\sum x2 - (\sum x)2][n\sum y2 - (\sum y)2]n(\sum xy) - (\sum x)(\sum y)$$

r = [n ∑ x2 − (∑ x)2][n ∑ y2 − (∑ y)2]n(∑ xy) − (∑ x)(∑ y)

Where:

- n = number of data points
- x and y = the two variables being compared
- ∑ denotes summation

This formula calculates the covariance of the variables and normalizes it by their standard deviations. It provides a scale-independent measure means it is unaffected by the unit of measurement.

Pearson Correlation Coefficient Table

Pearson Correlation Coefficient (r) Range	Type of Correlation	Interpretation
$0 < r \leq 1$	Positive	Both variables increase or decrease together.
$r = 0$	None	No relationship exists between the changes in both variables.
$-1 \leq r < 0$	Negative	As one variable increases, the other decreases.

Example:

- r = 0.85 suggests a strong positive correlation such as more study time leading to better test scores.
- r = -0.75 shows a strong negative correlation like the inverse relationship between outdoor temperature and heating costs.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative

Less than –.5	Strong	Negative

Key Properties of Pearson's Correlation Coefficient

1. Directionality: The sign of r shows the direction of the relationship

- Positive r means that as one variable increases, the other increases in a similar manner.
- Negative r means that as one variable increases, the other decreases.

2. Magnitude: The magnitude of r shows the strength of the relationship

- Values closer to ±1 represent a stronger relationship.
- Values closer to 0 signify a weaker or no relationship between the variables.

3. No Causation: Pearson's correlation measures association but does not imply causation. Even if two variables are strongly correlated, it doesn't mean that one causes the other to change.

4. Symmetry: Pearson's correlation is symmetric which mean the correlation between X and Y is the same as that between Y and X. In other words, r(X, Y) = r(Y, X).

5. Invariance: Pearson's correlation remains unchanged under linear transformations of the data. This means scaling (multiplying by a constant) or shifting (adding a constant) the data does not affect the correlation.

Types of Pearson Correlation Coefficient

Pearson Correlation Coefficient can be adjusted in different ways to handle specific data characteristics:

1. Adjusted Correlation Coefficient: It modifies the standard Pearson correlation to account for sample size and bias useful in small datasets. It provides a more accurate estimation of the population correlation by reducing overestimations from limited data.

2. Weighted Correlation Coefficient: It assigns weights to data points based on their importance or reliability and helps in giving more influence to more accurate or significant observations. This ensures that less reliable data doesn't distort the correlation.

3. Reflective Correlation Coefficient: Used in structural equation modeling (SEM) it measures the relationship between observed variables and latent constructs (unmeasured variables). It's useful in analyzing complex relationships in fields like psychology and sociology.

4. Scaled Correlation Coefficient: It normalizes the correlation to a specific range or magnitude which allows easier comparison across different datasets and studies. It standardizes correlation values for consistency in interpretation.

5. Pearson's Distance: It measures how much the correlation between two variables deviates from perfect correlation. It quantifies the dissimilarity between data points and provides insights into the relationship strength between variables.

6. Circular Correlation Coefficient: Used for circular data (e.g. angles or directions) it is used for the cyclical nature of the data which helps in making it ideal for analyzing variables like wind direction or time of day where traditional correlation methods don't apply.

7. Partial Correlation: It evaluates the direct relationship between two variables while controlling for the influence of other variables. It isolates the unique association between variables and helps in making it ideal for understanding true relationships in the presence of confounders.

Steps to Find the Correlation Coefficient

> *Steps to find the correlation coefficient with Pearson's correlation coefficient formula:*
>
> ***Step 1: Prepare the Data****: Create a table with columns for X (variable 1), Y (variable 2), XY (product of X and Y), X^2 (square of X) and Y^2 (square of Y).*
>
> ***Step 2: Multiply X and Y****: Multiply each value of X and Y to fill the XY column.*
>
> ***Step 3: Square the X Values:*** *Square each value in the X column and fill the X^2 column.*
>
> ***Step 4: Square the Y Values:*** *Square each value in the Y column and fill the Y^2 column.*
>
> ***Step 5: Sum the Columns****: Add up all the values in each column* $(\Sigma X, \Sigma Y, \Sigma XY, \Sigma X^2, \Sigma Y^2)$*(ΣX, ΣY , ΣXY , ΣX², ΣY ²). The summation symbol (ΣΣ) represents the sum of all the values in the respective columns.*
>
> ***Step 6: Apply the Formula****: Now use the Pearson correlation formula:*
>
> $r = [n\sum x2 - (\sum x)2][n\sum y2 - (\sum y)2]n(\sum xy) - (\sum x)(\sum y)$
> *r = [n ∑ x2 − (∑ x)2][n ∑ y2 − (∑ y)2]n(∑ xy) − (∑ x)(∑ y)*
>
> ***Step 7: Interpret the Result****: After calculating r check whether the correlation is positive or negative based on the sign of r:*
>
> - *r > 0 shows a positive correlation.*

> - *$r < 0$ shows a negative correlation.*
> - *$r = 0$ shows no linear correlation.*

Pearson Correlation Coefficient Examples

Let's see some real examples for better understanding:

Example 1: There is some correlation coefficient that was given to tell whether the variables are positive or negative? **0.69, 0.42, -0.23, -0.99**

Solution:

> *The given correlation coefficient is as follows:*
>
> *0.69, 0.42, -0.23, -0.99*
>
> *Tell whether the relationship is negative or positive*
>
> - ***0.69:*** *The relationship between the variables is a strong positive relationship*
> - ***0.42:*** *The relationship between the variables is a strong positive relationship*
> - ***-0.23:*** *The relationship between the variables is a weak negative relationship*
> - ***-0.99:*** *The relationship between the variables is a very strong negative relationship*

Example 2: Calculate the correlation coefficient for the following data by the help of Pearson's correlation coefficient formula: **X = 21, 31, 25, 40, 47, 38 and Y = 70, 55, 60, 78, 66, 80**

Solution:

Given variables are, X = 21, 31, 25, 40, 47, 38, and Y = 70, 55, 60, 78, 66, 80

To find the correlation coefficient of the following variables Firstly a table is to be constructed as follows, to get the values required in the formula also add all the values in the columns to get the values used in the formula.

X
21
31
25
40
47
38
∑x = 202

$\sum xy = 13937$, $\sum x = 202$, $\sum y = 409$, $\sum x^2 = 7280$, $\sum y^2 = 28265$

Put n =6 all the values in the Pearson's correlation coefficient formula:-

R= $n(\sum xy) - (\sum x)(\sum y)/\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2}$

$n(\sum xy) - (\sum x)(\sum y)/\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2}$

R= 6(13937) - (202)(409) / √ [6(7280)-(202)²][6(28265)-(409)²]

R= 1004 / √[2876][2909]

R=1004 / 2892.452938

R=-0.3471

The correlation coefficient is -0.3471

Bivariate Correlation

Pearson's Correlation Coefficient is used to measure bivariate correlation which refers to the relationship between two variables. It helps in assessing both the strength and direction of their linear relationship:

- **Positive correlation:** As one variable increases, the other tends to increase as well.
- **Negative correlation:** As one variable increases, the other tends to decrease.
- **Zero correlation:** A value of 0 shows no linear relationship between the two variables.

This helps researchers understand how two variables move together and whether changes in one are associated with changes in the other.

Correlation Matrix

When analyzing datasets with multiple variables, the Pearson correlation coefficient is used to construct a correlation matrix. This is a square table summarizing the correlation coefficients between all possible pairs of variables in the dataset.

- By examining the correlation matrix, researchers can quickly identify which variables have strong positive, negative or no correlation with each other.
- This helps in understanding the overall structure of the data and identifying relationships that may warrant further analysis or investigation.

You can also refer to more related articles:

> - *Correlation Coefficient Formula*
> - *Karl Pearson's Coefficient of Correlation*

Understanding Pearson’s Correlation Coefficient provides a solid foundation for finding relationships in data which helps researchers and analysts to make informed, data-driven decisions.

> ***Note****: It might be helpful to mention that Pearson's "r" only measures linear relationships and does not capture non-linear relationships. For non-linear data, other correlation measures like Spearman's rank correlation should be considered.*

Practice Problem Based On Pearson Correlation Coefficient

Question 1: Given the following data:X = [21, 31, 25, 40, 47, 38] and Y = [70, 55, 60, 78, 66, 80]. Calculate the Pearson Correlation Coefficient (r) for the data and interpret the strength and direction of the correlation.

Question 2: A Pearson correlation coefficient of r = -0.95 is calculated between the number of hours a person exercises per week and their blood pressure levels. What does this value suggest about the relationship between the two variables? Is it positive or negative, and how strong is the correlation?

Question 3: Consider the following data:Hours Studied (X) = [1, 2, 3, 4, 5]andTest Scores (Y) = [50, 55, 65, 70, 80].Calculate the Pearson correlation coefficient for the data**.**

Question 4: A researcher finds a Pearson correlation coefficient of r = 0.85 between the time spent on social media (X) and stress levels (Y). How would you interpret the strength and direction of the relationship, and what does this tell you about the association between these two variables

Question 5:

1. Given a Pearson correlation coefficient of r = 0.85 between the amount of time students spent studying and their score on a math test, interpret the strength and direction of the relationship.

2. Consider the following small dataset representing hours studied (X) and test scores (Y):

Hours Studied (*X*)	**Test Score (*Y*)**
1	50
2	55
3	65
4	70
5	80

Calculate the Pearson correlation coefficient (r) for the data.

Answer:-

> 1. *r ≈ 0.38, moderate positive correlation*
> 2. *very strong negative correlation*
> 3. *r ≈ 0.803, strong positive correlation*
> 4. *strong positive correlation*
> 5. *r= 0.993, very strong positive correlation*

Suggested Quiz

3 Questions

When should Spearman’s Rank Correlation be used instead of Pearson’s correlation?

- A

  When the relationship between variables is strictly linear

- B

  When the data contains outliers or follows a non-normal distribution

- C

  When the data is already standardized

- D

  When analyzing categorical data

Which method is used to calculate the **linear correlation** between two continuous variables?

- A
- B
- C
- D

When the Pearson correlation coefficient is **-1**, it means:

- A
- B

  Strongest negative linear relationship

- C

  Strongest positive relationship

- D

Quiz Completed Successfully

Your Score : 2/3

Accuracy : 0%

Login to View Explanation

1/3 < Previous Next >