Pearson Correlation is a statistical tool for determining the degree and direction of a linear relationship between two continuous variables. It is denoted by the Pearson correlation coefficient (r)..

## Definition of Pearson Correlation Coefficient (r)

- The coefficient ranges from -1 to +1:
    - **r = +1**: Indicates a perfect positive linear relationship, showing that as one variable increases, the other variable also increases proportionally.
    - **r = -1**: Indicates a perfect negative linear relationship, indicating that as one variable increases, the other variable decreases proportionally.
    - **r = 0**: Indicates no linear relationship between the variables.

## Interpretation of the Coefficient

1. **Strength of the Relationship**:
    - **0.0 < |r| < 0.3**: Weak correlation
    - **0.3 ≤ |r| < 0.7**: Moderate correlation
    - **0.7 ≤ |r| < 1.0**: Strong correlation
2. **Direction of the Relationship**:
    - A **positive r** indicates that the variables move in the same direction (both increase or decrease together).
    - A **negative r** indicates that the variables move in opposite directions (one increases while the other decreases).

    N.B: The data should not have significant outliers, as they can skew the results.

# P-value

The **p-value** is a number that helps you decide whether the result of an experiment or study is meaningful or if it could have happened just by chance.

- A **small p-value** (usually less than 0.05) means the result is very surprising if the null hypothesis were true, so you can confidently say your result is **probably real**.
- A **large p-value** (like 0.2 or 0.5) means the result is not that surprising if the null hypothesis were true, so you don't have enough evidence to claim something real is happening.

**p-value** doesn't say how *true* something is, but rather how likely it is that your results happened by chance.

Like in the case of a coin, just imagine you're flipping a coin, and someone tells you it's a fair coin (50% heads, 50% tails). If you flip the coin 10 times and get 9 heads, you might wonder if the coin is actually fair.

- If you calculate a **p-value** and it's **really small**, like 0.01, it means getting 9 heads out of 10 flips would rarely happen by chance with a fair coin. So, you start to believe the coin might not be fair.
- If the **p-value** is **large**, like 0.5, it means getting 9 heads could easily happen by chance, so there's no strong reason to doubt the coin is fair.

In short, a small p-value gives you confidence that something meaningful is going on, while a large p-value means you don't have enough evidence to say anything for sure.

## Predictor Variable (Independent Variable): similar to x in the case of the equation of a straight line or the slope-intercept form of a linear equation.

- This is the variable that you think **influences or predicts** the outcome.
- It is typically plotted on the **x-axis**.
- Examples: Hours studied, temperature, age, etc.
- You use this variable to predict or explain changes in the target variable.

## Target Variable (Dependent Variable): similar to y in the case of the equation of a straight line or the slope-intercept form of a linear equation.

- This is the variable whose value you are trying to **predict or explain**.
- It is usually plotted on the **y-axis**.
- Examples: Test scores, product sales, rainfall, etc.
- The target variable depends on or is influenced by the predictor variable.

## Target Variable (Dependent Variable):

- **Survived**: This variable indicates whether a passenger survived the sinking of the Titanic or not (coded as 0 for "not survived" and 1 for "survived").
- This is the **outcome** you're trying to predict, so it's your **target** variable.

## Predictor Variables (Independent Variables):

These are characteristics or factors you think may have influenced whether someone survived or not. In a scatter plot, you could analyze relationships with:

- **Age**: Did age affect survival? Younger passengers may have had a higher chance of survival.

- **Fare**: Did the ticket fare a passenger paid correlate with survival? Wealthier passengers in first-class might have had better access to lifeboats.
- **Gender**: Did gender (male or female) influence survival? Women and children were prioritized for lifeboats.

In this example:

- **Survived (0 or 1)** is the **target variable** because it's the outcome you're trying to predict.
- Variables like **age, fare, and gender** are **predictor variables** because they might influence the chance of survival.

## Example Scatter Plot

- If you plot **Fare (x-axis)** against **Survived (y-axis)**, you may see a pattern where passengers who paid higher fares were more likely to survive.