Statistics for Computing

(CSC 502 0.0)

MSc in Computer Science

Prof. (Dr.) R.M. KAPILA RATHNAYAKA

B.Sc. Special (Math. & Stat.) (Ruhuna), M.Sc. (Industrial Mathematics) (USJ),
M.Sc. (Stat.) (WHUT, China),
Ph.D. (Applied Statistics, WHUT)

Introduction to Correlation

Chapter 04

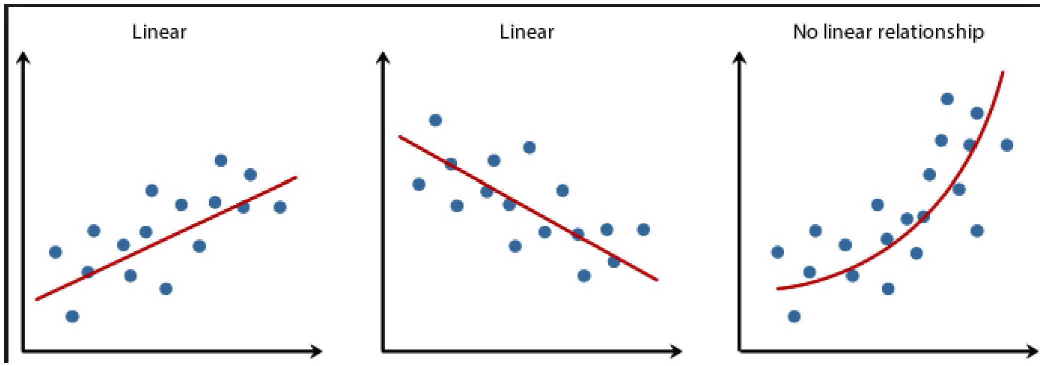
Correlation and Regression Analysis

- Correlation and regression are the two most commonly used techniques for **investigating the relationship** between quantitative variables.

- Correlation is used to give the relationship between the variables whereas linear regression uses an equation to express this relationship.

- In this section we will first discuss correlation analysis, which is used to quantify the association between two continuous variables
  - ✓between an independent and a dependent variable
  - ✓ between two independent variables

Correlation Definition



- If an increase (or decrease) in one variable causes a corresponding increase (or decrease) in another then the two variables are said to be directly correlated.
- Similarly, if an increase in one causes a decrease in another or vice versa, then the variables are said to be indirectly correlated.
- If a change in an independent variable does not cause a change in the dependent variable then they are uncorrelated.

Method 01: Karl Pearson’s Coefficient of Correlation

- Correlation is a *bivariate analysis* that measures the *strength of association between two variables* and the *direction of the relationship*.

- Usually, in statistics, we measure four types of correlations:
  - Pearson correlation,
  - Kendall rank correlation,
  - Spearman correlation
  - Point-Biserial correlation

Method 01: Karl Pearson’s Coefficient of Correlation

- **Assumptions**
- For the Pearson *r* correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve).

- Other assumptions include linearity and homoscedasticity.

- Linearity assumes a straight line relationship between each of the two variables and homoscedasticity assumes

Karl Pearson's Coefficient of Correlation

- **Definition:**

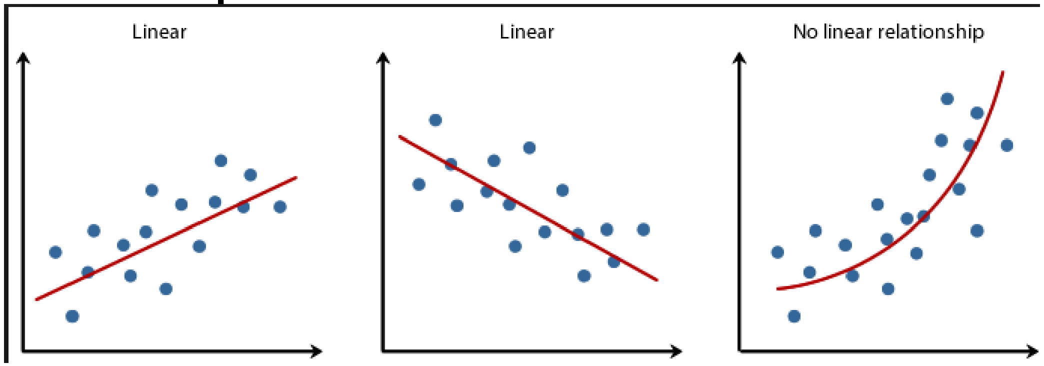
Karl Pearson's Coefficient of Correlation is widely used mathematical method wherein the numerical expression is used to calculate ***the degree*** and ***direction*** of the relationship between linear related variables.

1. **Karl Pearson's Coefficient of Correlation-** evaluates the linear relationship between two continuous variables

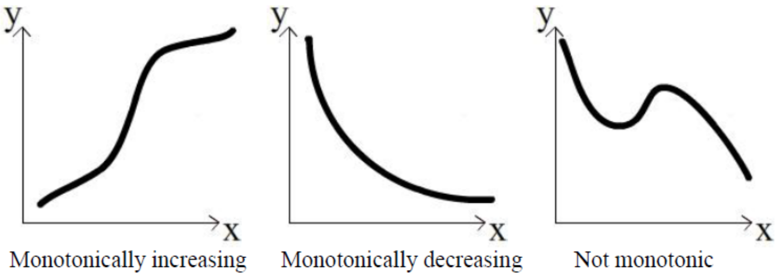
Correlation Definition

- Correlation can be defined as a measurement that is used to quantify the relationship between variables.

1. **Karl Pearson's Coefficient of Correlation-** evaluates the linear relationship between two continuous variables



2. **Spearman rank-order correlation-** evaluates the monotonic relationship between two continuous or ordinal variables



Method 01: Karl Pearson’s Coefficient of Correlation

- The coefficient of correlation is denoted by symbol ***‘r’.***

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

Correlation Analysis

- Correlation is one of the most common and most useful statistics

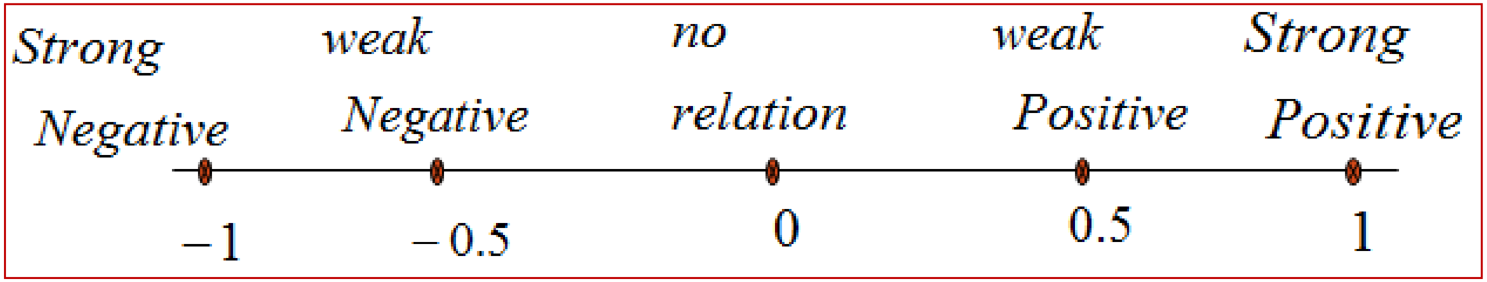
- It is a term that refers to the strength of a relationship between two variables ***(single number that describes the degree of relationship between two variables)***.

- A strong, or high, correlation means that two or more variables have a strong relationship with each other.

- A weak or low correlation means that the variables are hardly related.

Properties of Coefficient of Correlation

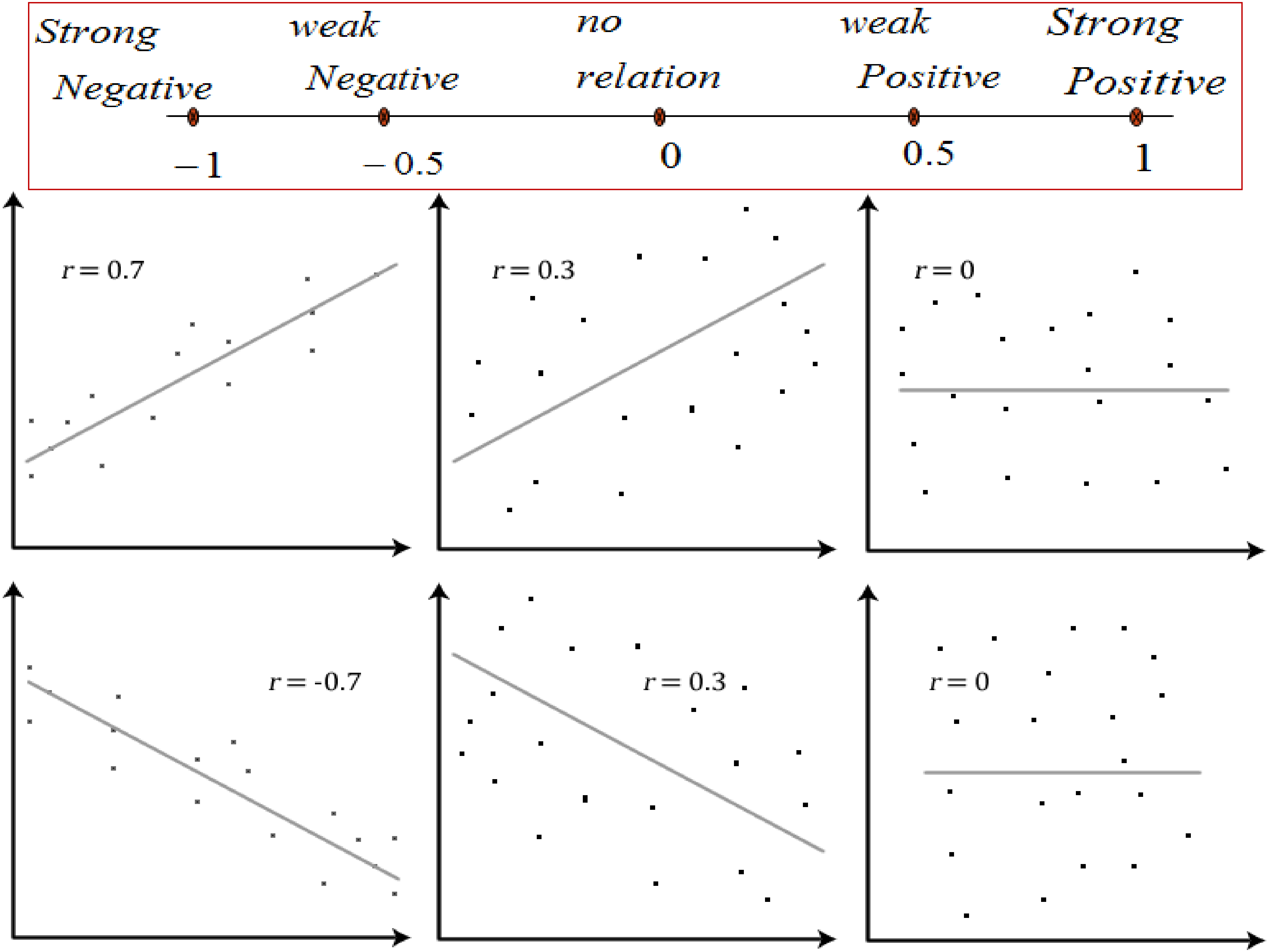
1. The value of the coefficient of correlation (r) always **lies between ±1**.



r=+1, perfect positive correlation

r=-1, perfect negative correlation

r=0, no correlation


Strong Negative
weak Negative
no relation
weak Positive
Strong Positive
– 1
– 0.5
0
0.5
1
r = 0.7
r = 0.3
r = 0
r = -0.7
r = 0.3
r = 0

Example (01) :Calculate the coefficient of correlation between X and Y from the following data.

X	1	2	3	4	5	6	7
Y	2	4	5	3	8	6	7

x	1	2	3	4	5	6	7
y	2	4	5	3	8	6	7

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

$$\sum x = 28 \quad \sum y = 35$$

$$0.79 \Rightarrow \textit{Strong positive}$$

Example (02):

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

Weight (Kg)	Age (years)	serial No
12	7	1
8	6	2
12	8	3
10	5	4
11	6	5
13	9	6

$$\boxed{r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,]\,[\, n\Sigma y^2 - (\Sigma y)^2 \,]}}}$$

$$r = \frac{461 - \dfrac{41 \times 66}{6}}{\sqrt{\left[291 - \dfrac{(41)^2}{6}\right] . \left[742 - \dfrac{(66)^2}{6}\right]}}$$

r = 0.759

strong direct correlation