Correlation Analysis (Part 1)

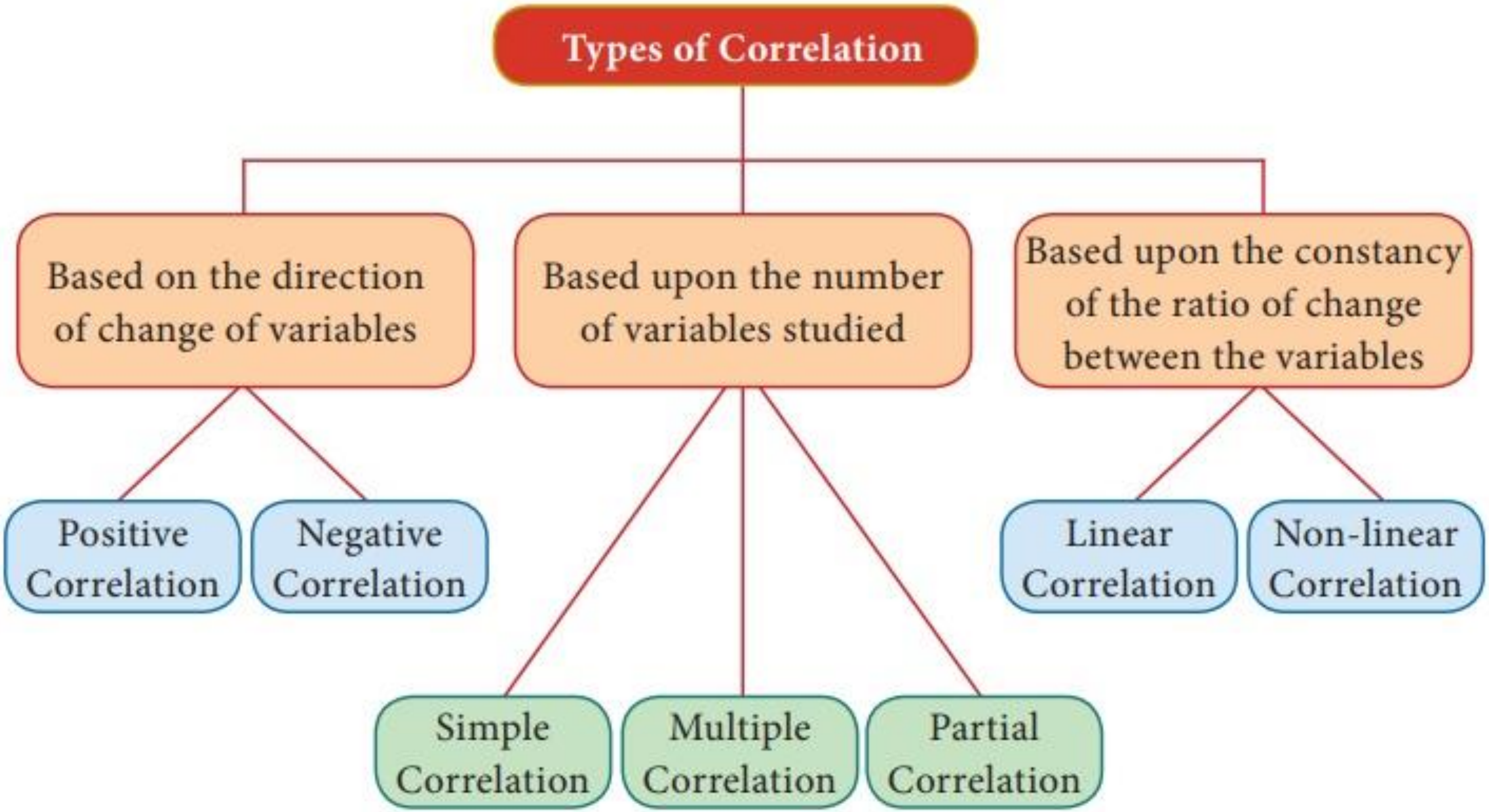
Munmun Akter

Lecturer

NFE, DIU

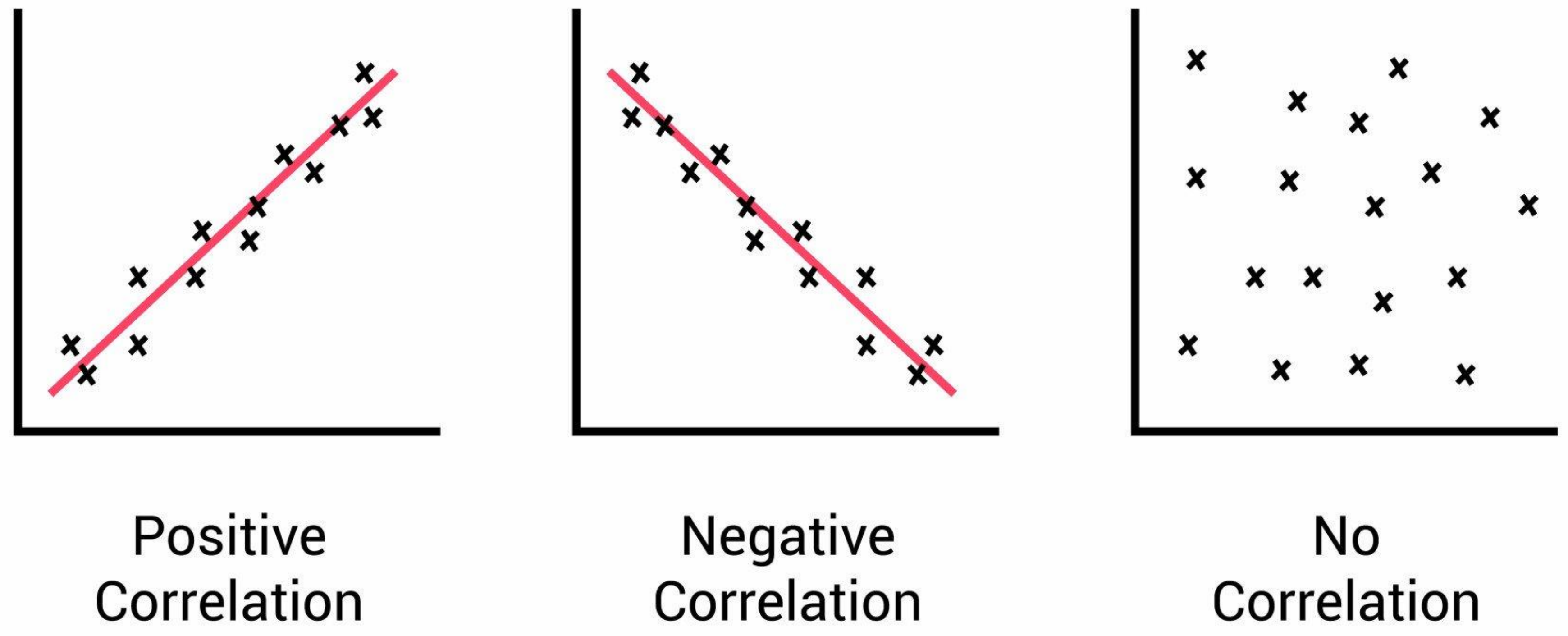
Correlation

- It is a statistical measure that describes the relationship between two or more variables.
- It helps us understand how changes in one variable are associated with changes in another variable.
- The key idea behind correlation is to determine the degree to which two variables tend to move together or apart.


Types of Correlation
Based on the direction of change of variables
Based upon the number of variables studied
Based upon the constancy of the ratio of change between the variables
Positive Correlation
Negative Correlation
Simple Correlation
Multiple Correlation
Partial Correlation
Linear Correlation
Non-linear Correlation

Positive Correlation: This implies that when one variable increases, the other one tends to increase as well, and when one variable decreases, the other one tends to decrease.

Negative Correlation: On the other hand, when two variables have a negative correlation, they move in opposite directions.



1. Simple Correlation or Bivariate correlation:

- refers to the statistical relationship between **two** variables only.
- It measures the **strength** and **direction** of the linear association between two **quantitative** variables.
- The most common method to quantify the simple correlation between two variables is by using the **Pearson correlation coefficient (denoted as "r").**
- The Pearson correlation coefficient ranges from **-1 to +1**, where +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Steps to calculate the Pearson correlation coefficient (r):

1. Calculate the mean ($\bar{x}$) of variable x and the mean ($\bar{y}$) of variable y.
2. For each data point, subtract the mean of the corresponding variable (x or y) from the data point.
3. Square each of the differences obtained in step 2.
4. Multiply the differences obtained for each data point (x and y) and sum up these products.
5. Take the square root of the sum of the squared differences for variable x and the square root of the sum of the squared differences for variable y.
6. Divide the sum of the products (from step 4) by the product of the square roots (from step 5).

X	Y	X^2	Y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum X = 70$	$\sum Y = 63$	$\sum X^2 = 728$	$\sum Y^2 = 651$	$\sum XY = 676$

Table 9.2

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \times \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{7(676) - (70)(63)}{\sqrt{7(728) - (70)^2} \times \sqrt{7(651) - (63)^2}}$$

$$= \frac{322}{339.48}$$

$$r = +0.95$$

Now calculate Pearson Correlation Coefficient(r) from the following table:

x	y	x^2	y^2	xy
16	11	256	121	176
15	18	225	324	270
12	10	144	100	120
10	20	100	400	200
8	17	64	289	136
$\sum x = 61$	$\sum y = 76$	$\sum x^2 = 789$	$\sum y^2 = 1234$	$\sum xy = 902$

Interpretation of correlation

How to interpret correlation coefficients:

1. Strength of the Relationship
2. Direction of the Relationship
3. Magnitude of the Correlation

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Example interpretations:

- If the correlation coefficient is +0.85, it indicates a strong positive linear relationship between the two variables.
- If the relationship is statistically significant ($p < 0.05$), you can say that there is strong evidence to suggest that as one variable increases, the other variable tends to increase as well.
- Correlation does not indicate causation or capture non-linear associations. So, it should be mentioned.


Thank you!