22/9/21

STATISTICS FOR ENGINEERS
MAT2001 – FALL 2021-2022

Dr. G. Hannah Grace
Division of Mathematics
School of Advanced Sciences
VIT Chennai

SYLLABUS

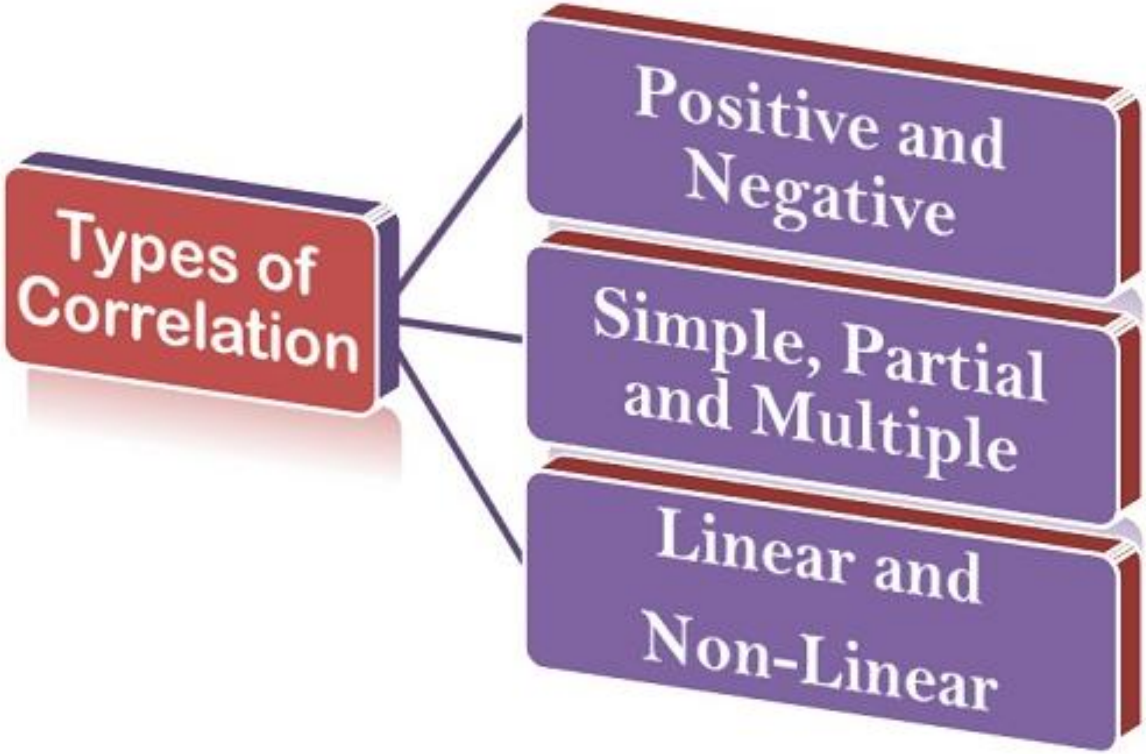
MODULE 1_Part II

Basic Statistical Tools for Analysis

Summary Statistics, Correlation and Regression, Concept of R^2 and Adjusted R^2 and Partial and Multiple Correlation, Fitting of simple and Multiple Linear regression, Explanation and Assumptions of Regression Diagnostics

Definition of Correlation

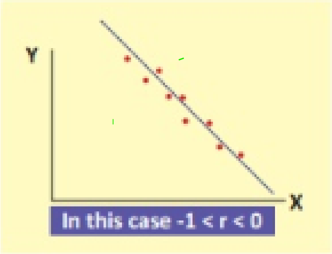
- Correlation is a statistical measure which helps in analyzing the interdependence of two or more variables.
- **A.M. Tuttle** defines correlation as:
- ***"An analysis of the co-variation of two or more variables is usually called correlation"***
- **Ya-kun-chou** defines correlation as:
- ***"The attempts to determine the degree of relationship between variables".***



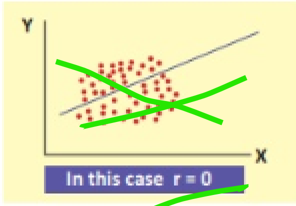
Types of Correlation
Positive and Negative
Simple, Partial and Multiple
Linear and Non-Linear

- **Positive Correlation**

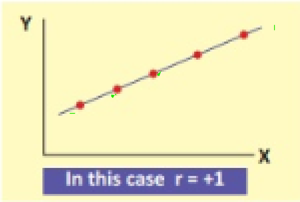
- A correlation in the **same direction** is called a positive correlation. If one variable increases the other also increases and when one variable decreases the other also decreases. For **example**, the length of an iron bar will increase as the temperature increases.



- **Negative Correlation**
- Correlation in **the opposite direction** is called a negative correlation. Here if one variable increases the other decreases and vice versa. For **example**, the volume of gas will decrease as the pressure increases, or the demand for a particular commodity increases as the price of such commodity decreases.



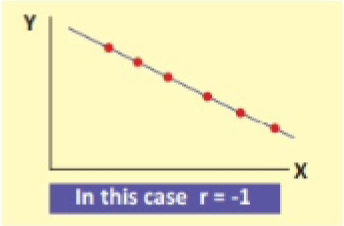
- **<u>No Correlation or Zero Correlation</u>**
- If there is no relationship between the two variables such that the value of one variable changes and the other variable remains constant, it is called no or zero correlation



- **<u>Perfect Positive Correlation</u>**

r = 1

- If the values of *x* and *y* increase or decrease ***proportionately*** then they are said to have perfect positive correlation.



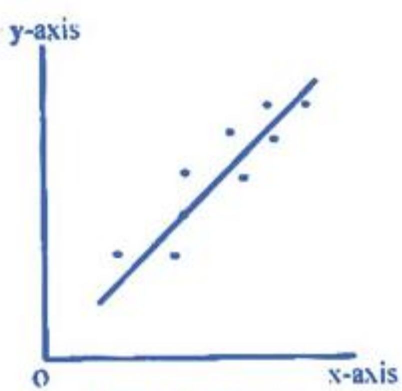
- **<u>Perfect Negative Correlation</u>**
- If *x* increases and *y* decreases ***proportionately*** or if *x* decreases and *y* increases ***proportionately***, then they are said to have perfect negative correlation.

II) Simple, multiple and Partial Correlation

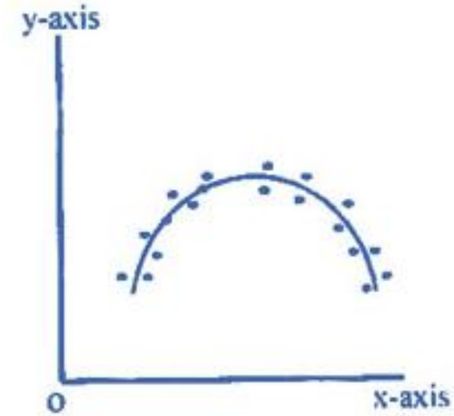
- **Simple Correlation**: Association between only two variables
- Eg. Number of computers and organization efficiency.
- Income and expenditure
- **Multiple Correlation**: When 3 or more variables are studied simultaneously.
- Eg. Rainfall, production of rice, price of rice are studied simultaneously.
- **Partial correlation**: When two or more variables are considered for analysis but only two influencing variables are studied and rest influencing variables are kept constant.
- Eg. Correlation is done with demand, supply and income where income is kept constant.

Linear and non linear correlation

- **Linear Correlation**: If the change in amount of one variable tends to make change in amount of other variable bearing constant changing ratio, it is said to be linear correlation.
- Eg. When the amount of output in a factory is doubled by doubling the number of workers
- **Non Linear Correlation(curvilinear)**: If the change in amount of one variable tends to make changes in amount of other variables but not bearing constant ratios it is said to be non linear correlation.



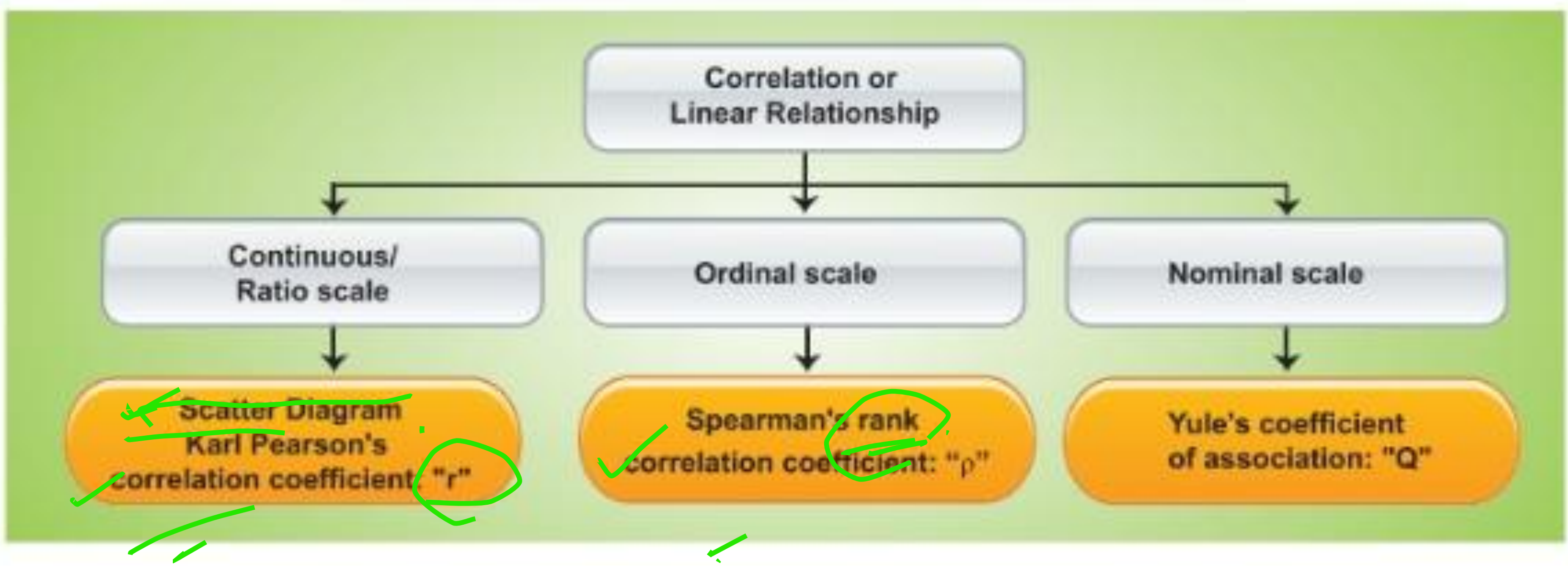
Linear Correlation



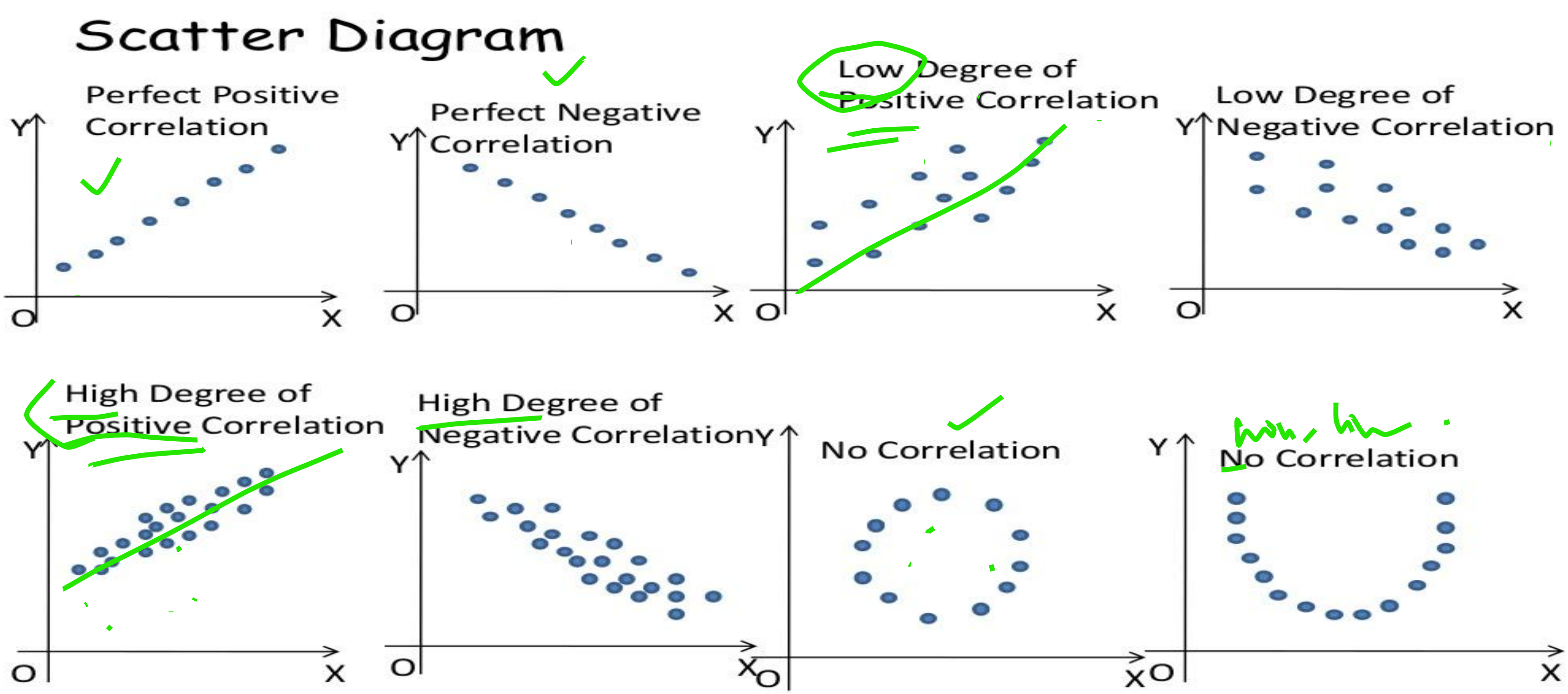
Non Linear Correlation

- **Correlation Analysis**
- The purpose of correlation analysis is to find the existence of linear relationship between the variables. However, the method of calculating correlation coefficient depends on the types of measurement scale, namely, ratio scale or ordinal scale or nominal scale.

Statistical tool selection



SCATTER DIAGRAM

- A scatter diagram is the simplest way of the diagrammatic representation of bivariate data. The direction of flow of points shows the type of correlation that exists between the two given variables.
- 

KARL PEARSON'S CORRELATION COEFFICIENT

- When there exists some relationship between two measurable variables, we compute the degree of relationship using the correlation coefficient.
- **Co-variance**
- Let (X,Y) be a bivariable normal random variable where $V(X)$ and $V(Y)$ exists. Then, covariance between X and Y is defined as

xy

E(x) → mean(x)

E(y) → mean(y)

- $\mathbf{cov(X,Y) = E(XY) - E(X)E(Y)}$

mean(xy)

- If (x_i,y_i), $i=1,2, \ldots, n$ is a set of n realisations of (X,Y), then the sample covariance between X and Y can be calculated from

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})$$

Karl Pearson's coefficient of correlation

- When X and Y are linearly related and (X,Y) has a bivariate normal distribution, the co-efficient of correlation between X and Y is defined as

$$\gamma(X,Y) = \frac{cov(X,Y)}{\sigma_x \, \sigma_y}$$

$\sigma_x \rightarrow SD(x)$

$\sigma_y \rightarrow SD(y)$

- This is also called as product moment correlation co-efficient which was defined by Karl Pearson.
- Based on a given set of n paired observations (x_i, y_i), i=1,2, ... n the sample correlation co-efficient between X and Y can be calculated from

$V(x,y)$

$$\gamma(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(y_i^2 - \bar{y}^2)}}$$

$\bar{x} = \Sigma x / n$

$\bar{y} = \Sigma y / n$

- or equivalently

$$\gamma(X,Y) = \frac{N\Sigma x_i y_i - \Sigma x_i \; \Sigma y_i}{\sqrt{(n\Sigma x_i^2 - [\Sigma x_i]^2)(N\Sigma y_i^2 - [\Sigma y_i]^2}}$$

$$\gamma(X,Y) = \gamma(U,V) = \frac{N\Sigma UV = \Sigma U \; \Sigma V}{\sqrt{(N\Sigma U^2 - [\Sigma U]^2)(N\Sigma V^2 - [\Sigma V]^2}} \quad where\ U = \frac{X-a}{h}\ and\ V = \frac{Y-b}{k}$$

Interpretations

- The correlation coefficient lies between -1 and+1. *i.e.* $-1 \leq r \leq 1$
- A positive value of 'r' indicates positive correlation.
- A negative value of 'r' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.
- If (0,0.3) week positive correlation
- If (-0.3,0) week negative correlation
- (0.7,1) strong positive correlation
- (1,-0.7) strong negative correlation

graph

Problem 1

Calculate the correlation coefficient *r* for the following data and hence discuss the nature of correlation

$n=5$

x	y	xy	x^2	y^2
1	– 3	– 3	1	9
2	– 1	– 2	4	1
3	0	0	9	0
4	1	4	16	1
5	2	10	25	4
$\sum x = 15$	$\sum y = -1$	$\sum xy = 9$	$\sum x^2 = 55$	$\sum y^2 = 15$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{5(9) - (15)(-1)}{\sqrt{5(55) - 15^2}\sqrt{5(15) - (-1)^2}}$$

$$= \frac{60}{\sqrt{50}\sqrt{74}} \approx 0.986$$

There is a strong positive linear correlation between x and y.