CORRELATION ANALYSIS

CORRELATION ANALYSIS

Presented by :
Moon,

Assistant Professor,
Department of Statistics,
Patna Women's College

INTRODUCTION

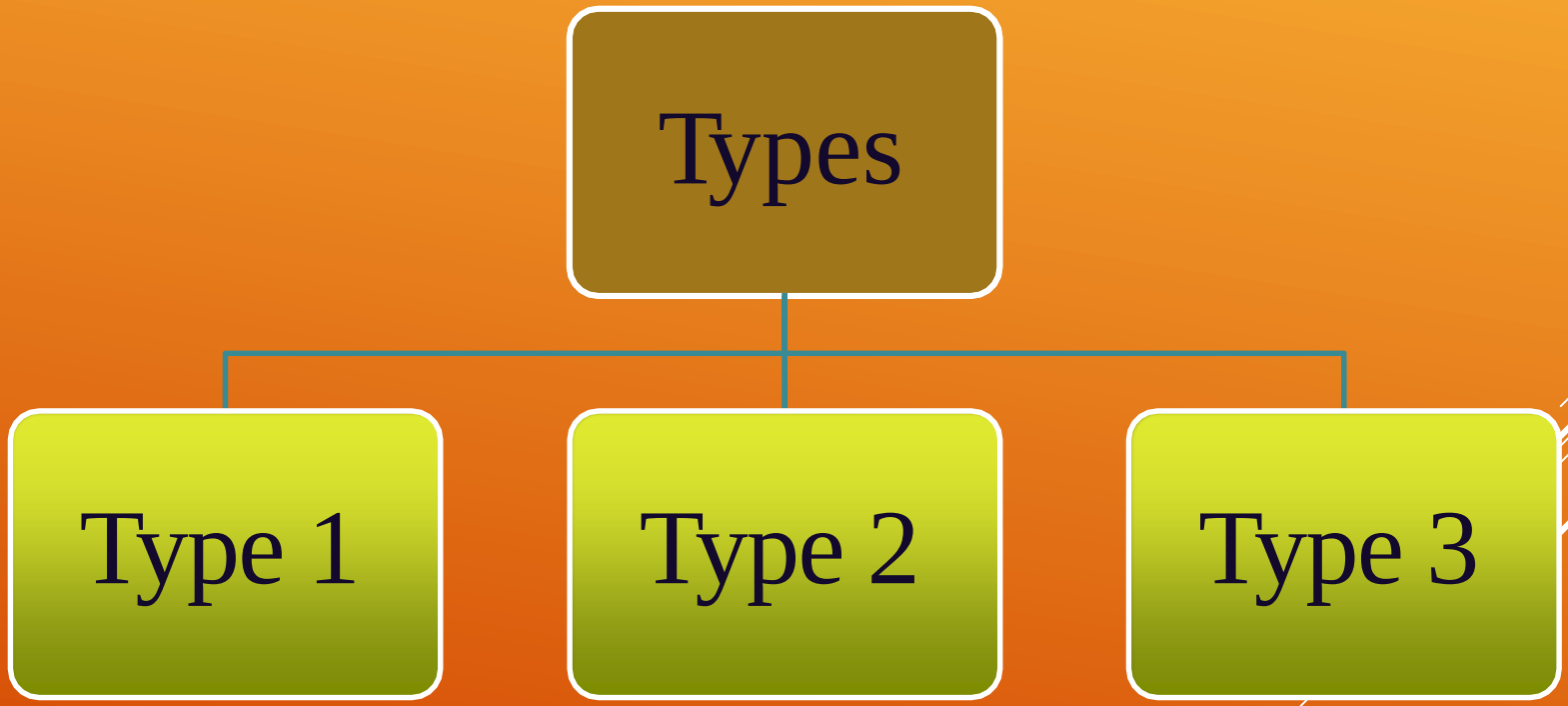
- Correlation a LINEAR association between two random variables
- Correlation analysis show us how to determine both the nature and strength of relationship between two variables
- When variables are dependent on time correlation is applied
- Correlation lies between +1 to -1

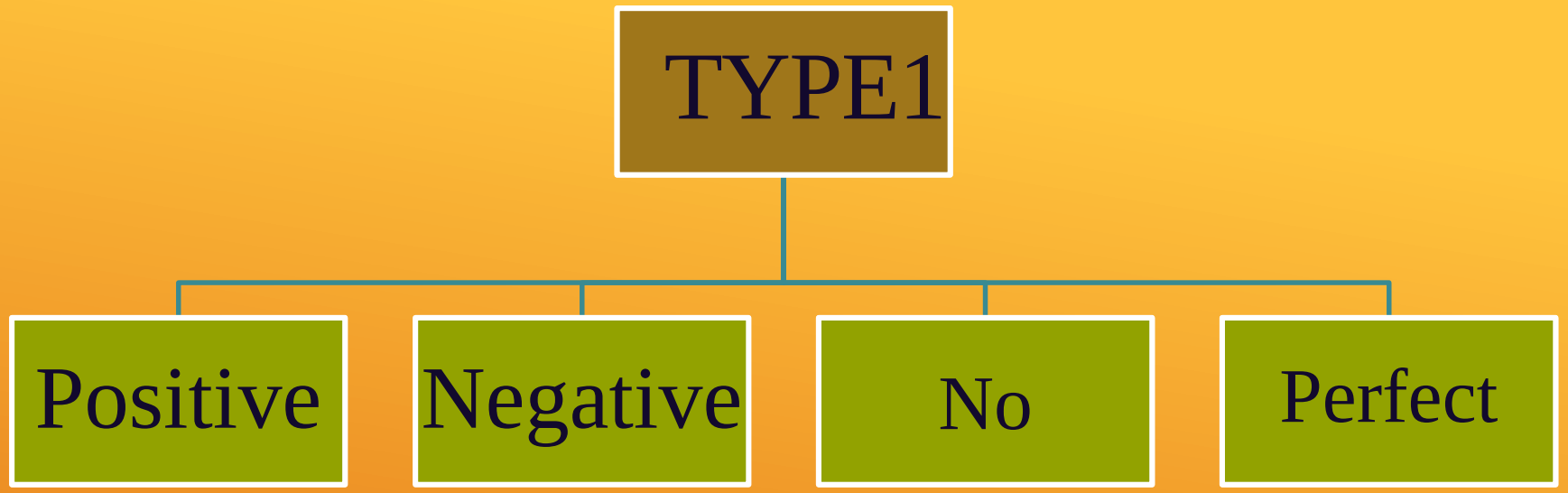
INTRODUCTION

- A zero correlation indicates that there is no relationship between the variables
- A correlation of –1 indicates a perfect negative correlation
- A correlation of +1 indicates a perfect positive correlation

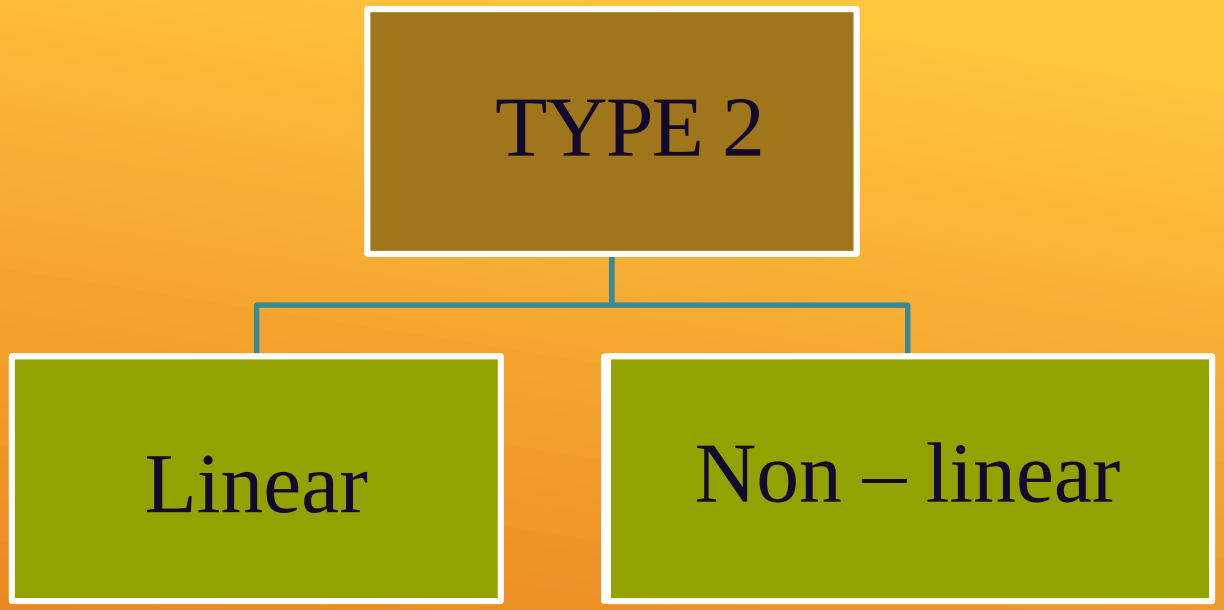
TYPES OF CORRELATION

- **There are three types of correlation**

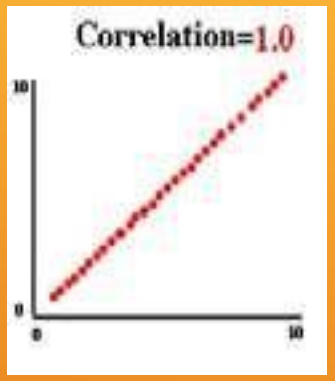


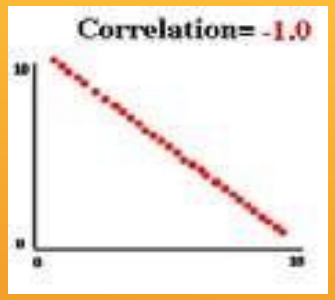


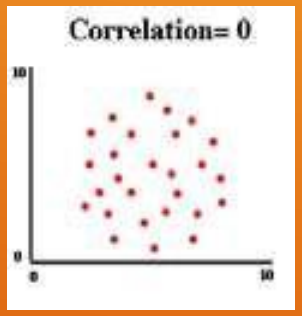
- If two related variables are such that when one increases (decreases), the other also increases (decreases).
- If two variables are such that when one increases (decreases), the other decreases (increases)
- If both the variables are independent

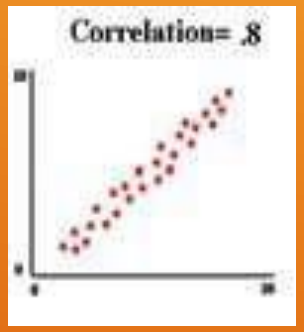


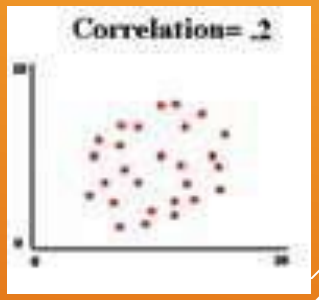
- When plotted on a graph it tends to be a perfect line
- When plotted on a graph it is not a straight line

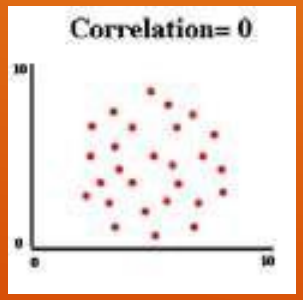

Correlation=1.0
10
0
0
10

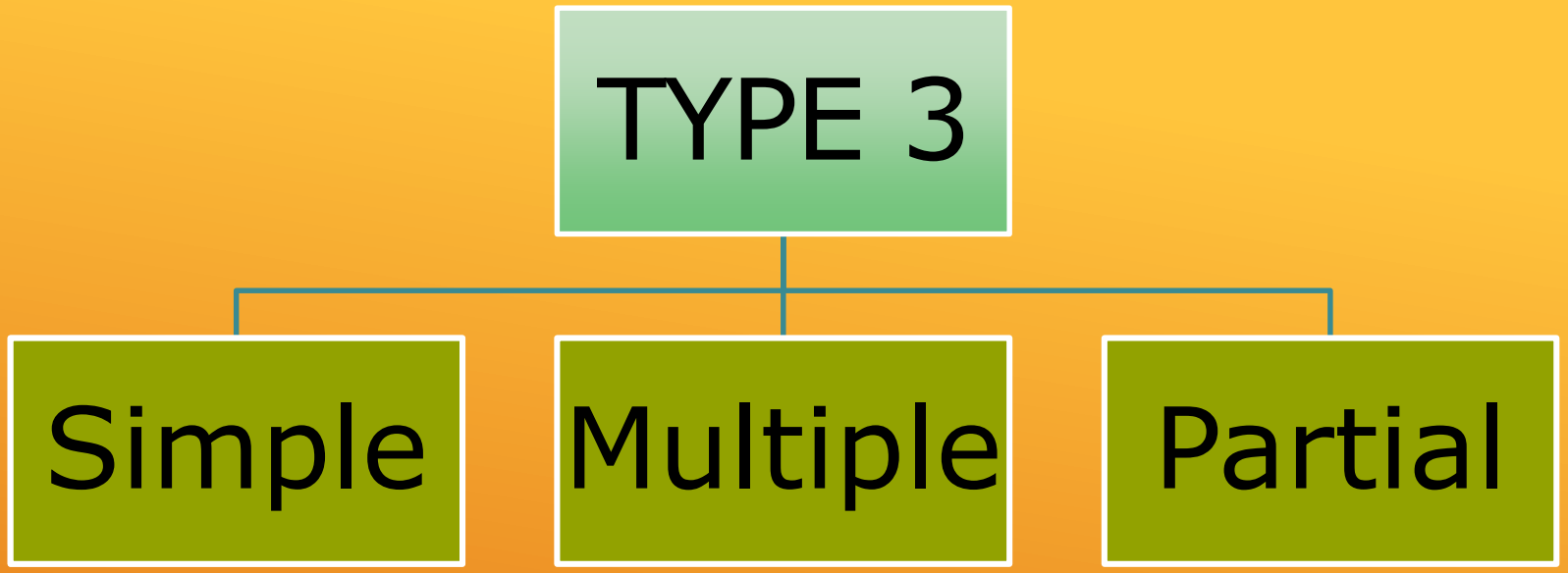

Correlation= -1.0
10
0
0
10


Correlation= 0
10
0
0
10


Correlation= .8
10
0
0
10


Correlation= .2
10
0
0
10


Correlation= 0
10
0
0
10



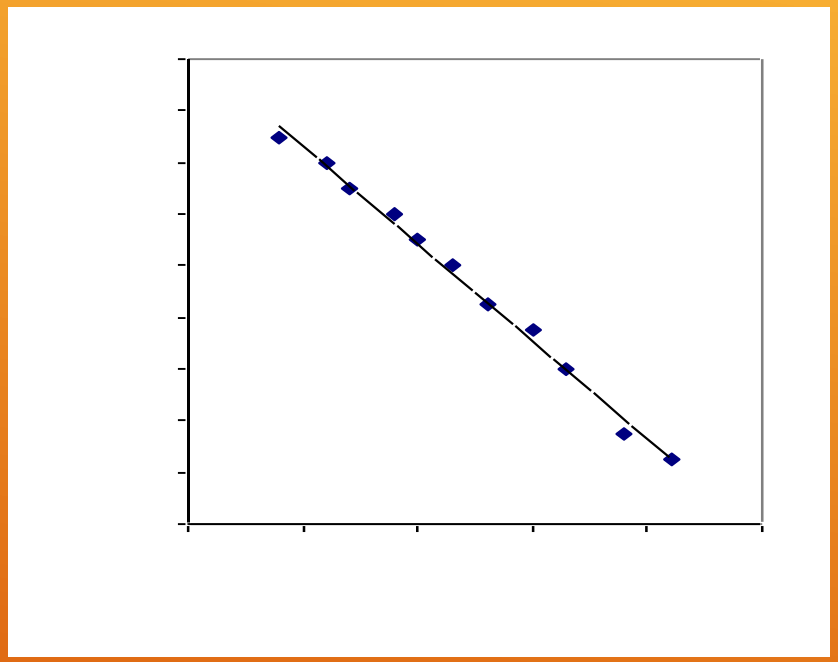
- Two independent and one dependent variable
- One dependent and more than one independent variables
- One dependent variable and more than one independent variable but only one independent variable is considered and other independent variables are considered constant

METHODS OF STUDING CORRELATION

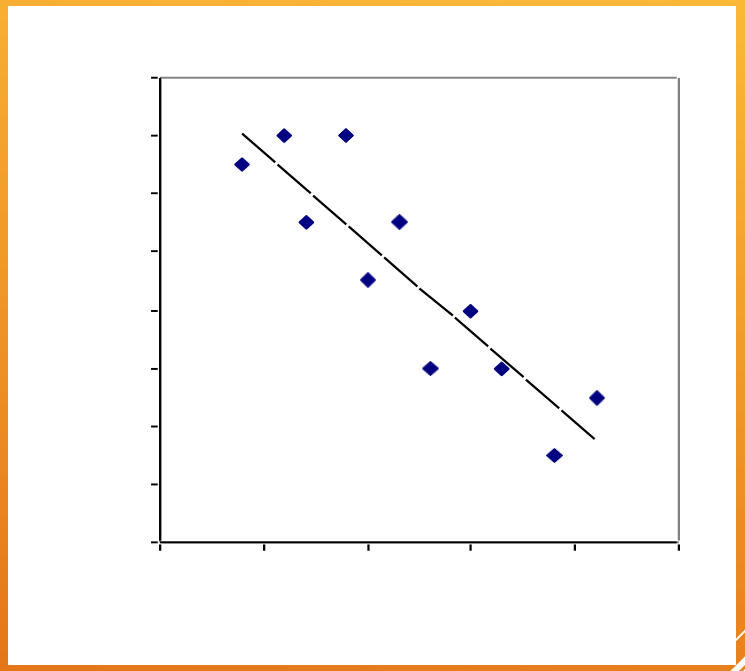
- **Scatter Diagram Method**
- **Karl Pearson Coefficient Correlation of Method**
- **Spearman's Rank Correlation Method**

CORRELATION: LINEAR RELATIONSHIPS

Strong relationship = good linear fit



Very good fit



Moderate fit

Points clustered closely around a line show a strong correlation. The line is a good predictor (good fit) with the data. The more spread out the points, the weaker the correlation, and the less good the fit. The line is a <u>REGRESSSION line</u> (Y = bX + a)

COEFFICIENT OF CORRELATION

- **A measure of the strength of the linear relationship between two variables that is defined in terms of the (sample) covariance of the variables divided by their (sample) standard deviations**
- **Represented by “r”**
- **r lies between +1 to -1**
- **Magnitude and Direction**

- $-1 \leq r \leq +1$

- **The + and – signs are used for positive linear correlations and negative linear correlations, respectively**

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2 - n\sum Y^2 - (\sum Y)^2}}$$

- Shared variability of X and Y variables on the top
- Individual variability of X and Y variables on the bottom

INTERPRETING CORRELATION COEFFICIENT *R*

- **strong correlation:** ***r*** **> .70 or** ***r*** **< –.70**
- **moderate correlation:** ***r*** **is between .30 & .70**

or ***r*** **is between –.30 and –.70**

- **weak correlation:** ***r*** **is between 0 and .30**

or ***r*** **is between 0 and –.30 .**

COEFFICIENT OF DETERMINATION

- **Coefficient of determination lies between 0 to 1**
- **Represented by r^2**
- **The *coefficient of determination* is a measure of how well the regression line represents the data**
- **If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation**
- **The further the line is away from the points, the less it is able to explain**

- r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable

- It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph

- The *coefficient of determination* is the ratio of the explained variation to the total variation

- The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y

- **The C*oefficient of determination* represents the percent of the data that is the closest to the line of best fit**
- **For example, if $r = 0.922$, then $r^2 = 0.850$**
- **Which means that 85% of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation)**
- **The other 15% of the total variation in *y* remains unexplained**

Thank you