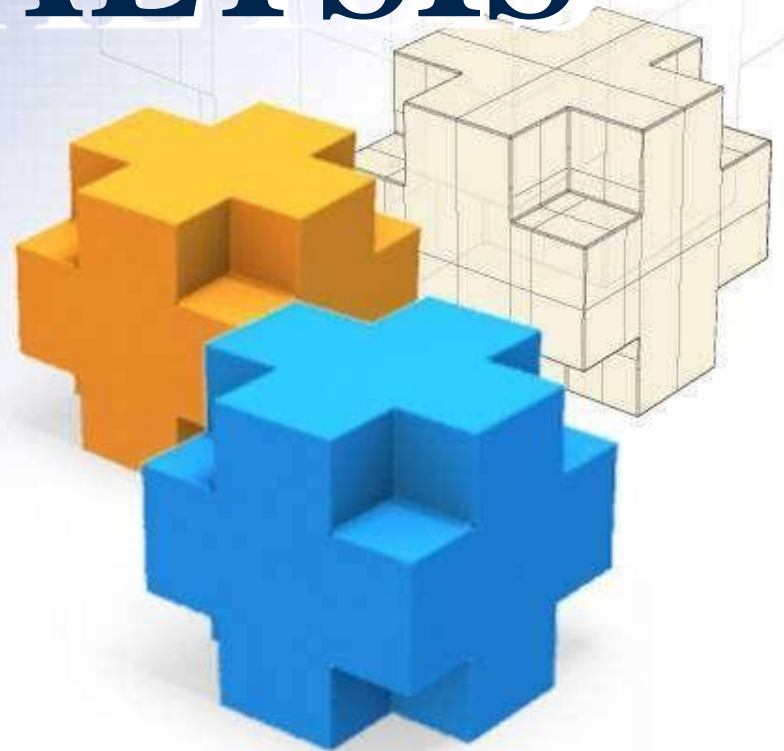




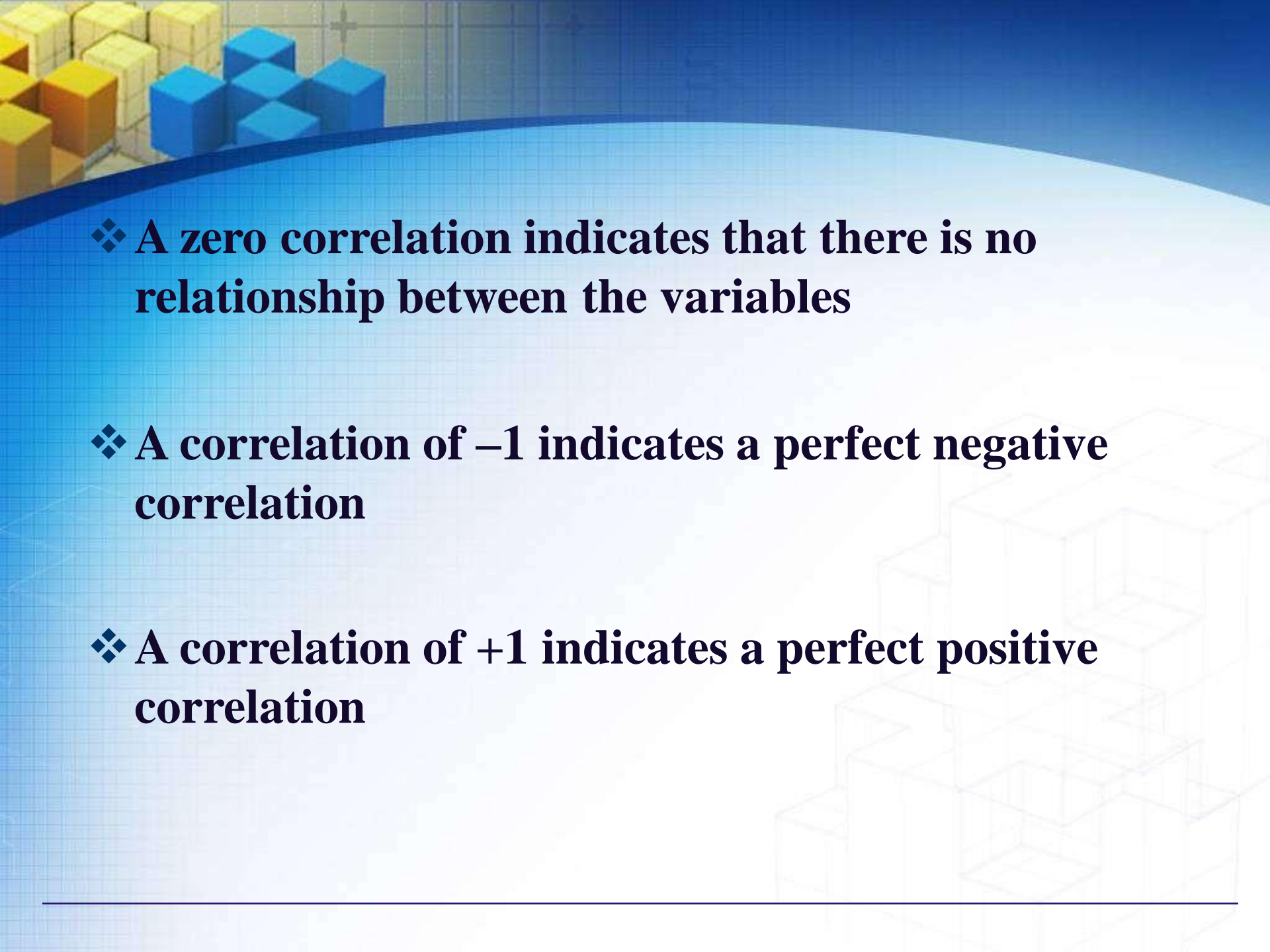
# CORRELATION ANALYSIS





# Introduction

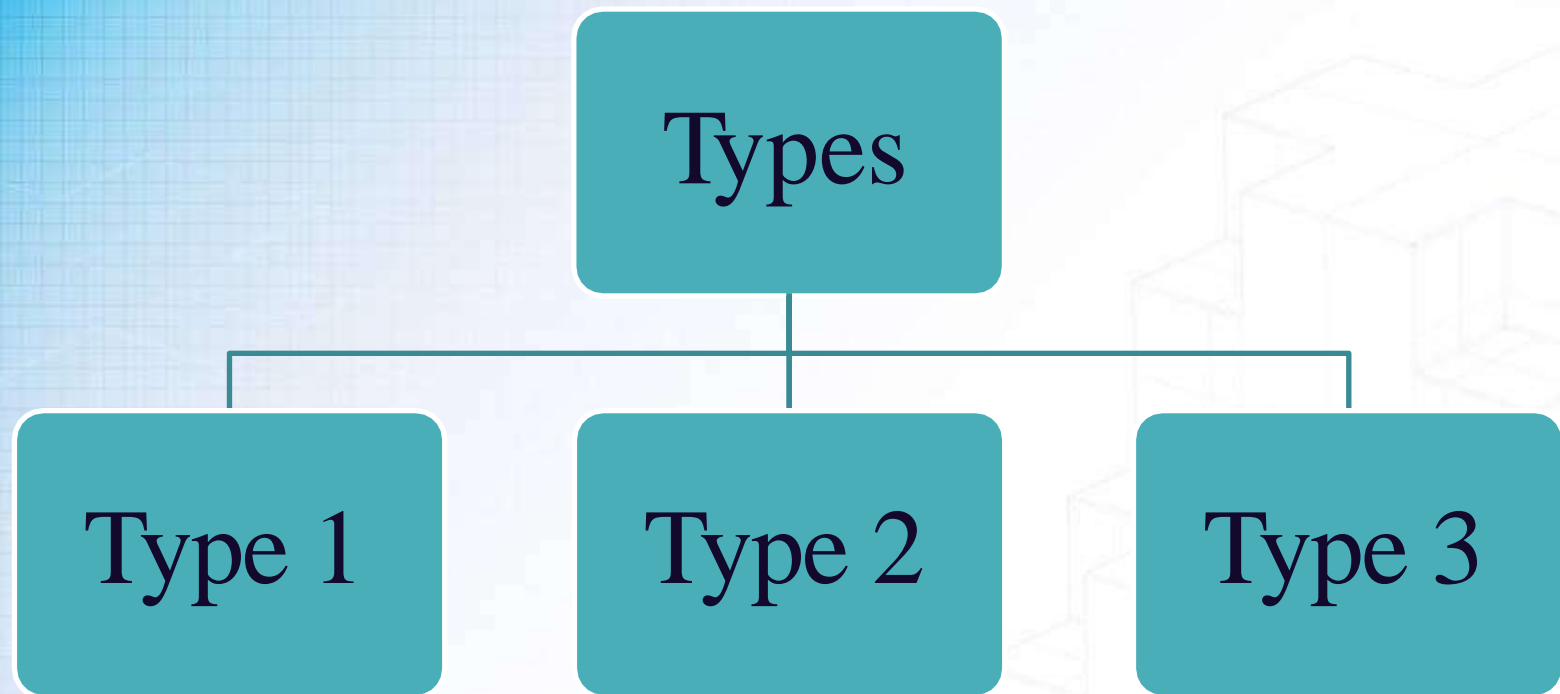
- ❖ **Correlation a LINEAR association between two random variables**
  - ❖ **Correlation analysis show us how to determine both the nature and strength of relationship between two variables**
  - ❖ **When variables are dependent on time correlation is applied**
  - ❖ **Correlation lies between +1 to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

- 
- ❖ **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





# Type1

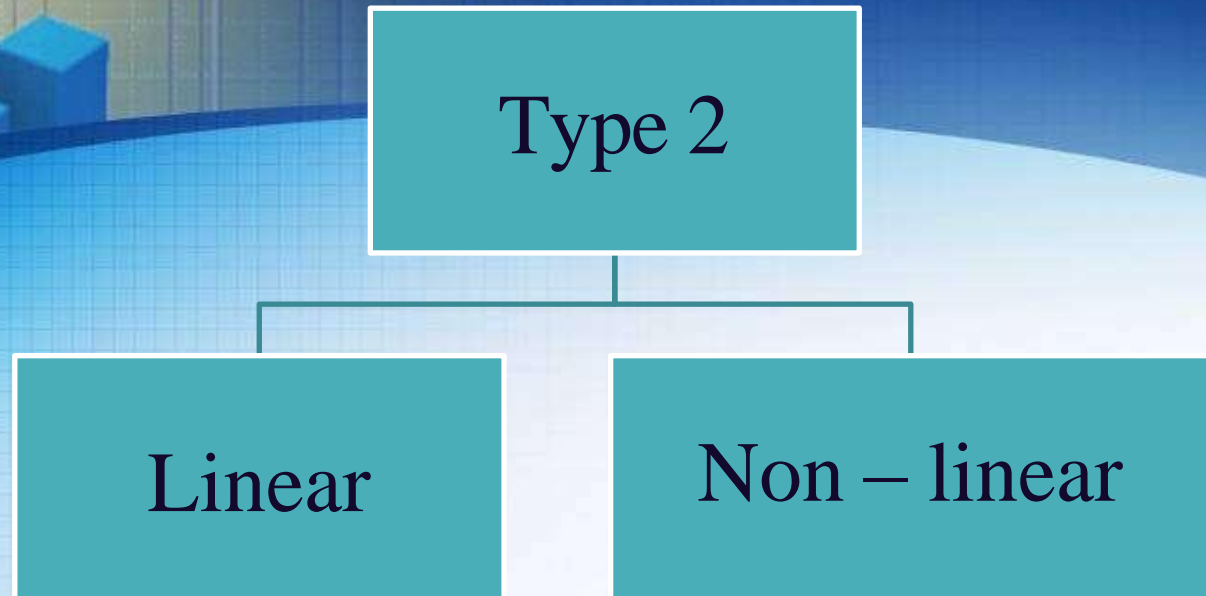
Positive

Negative

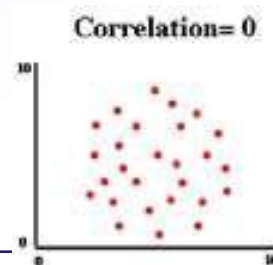
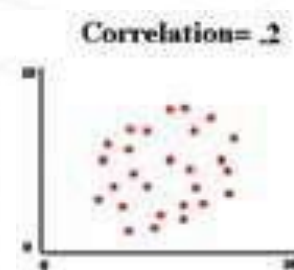
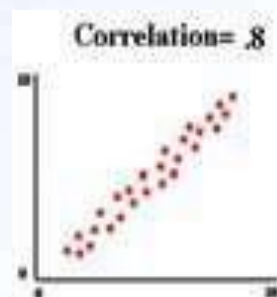
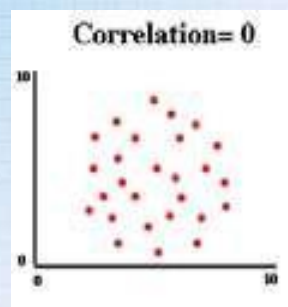
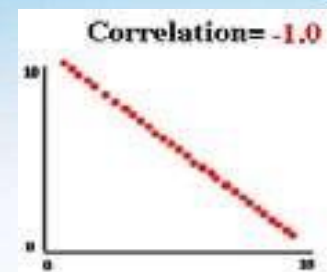
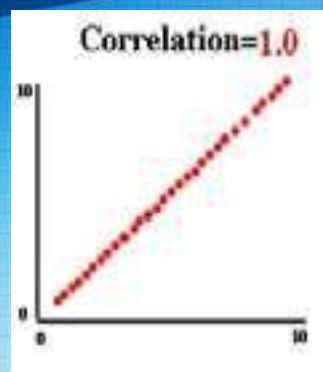
No

Perfect

- ❖ 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**
-







# Type 3

```
graph TD; A[Type 3] --> B[Simple]; A --> C[Multiple]; A --> D[Partial];
```

Simple

Multiple

Partial

- ❖ **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 Studying Correlation

❖ **Scatter Diagram Method**

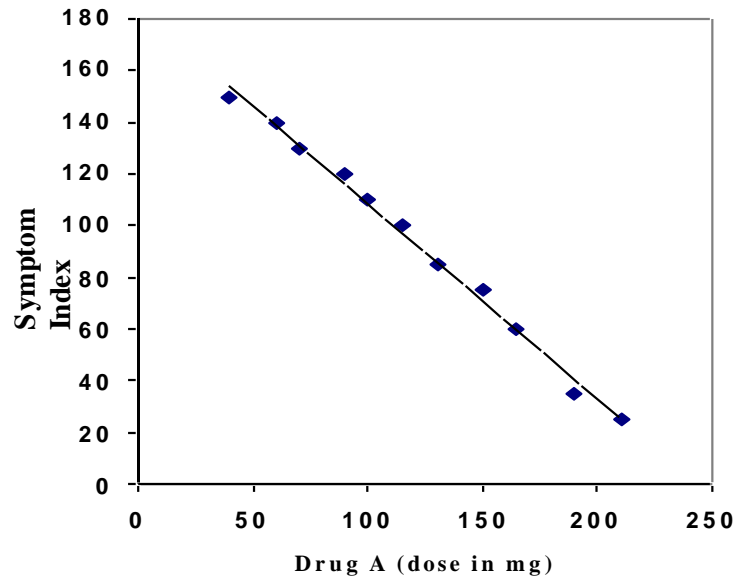
❖ **Karl Pearson Coefficient Correlation of Method**

❖ **Spearman's Rank Correlation Method**

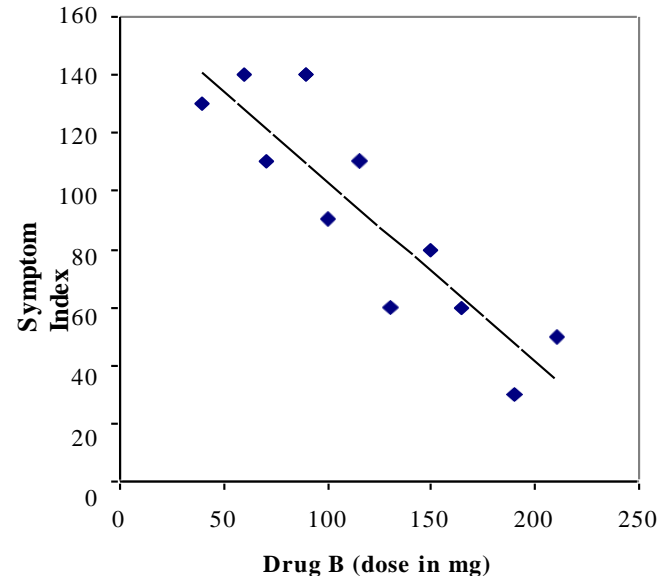
---

# 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 REGRESSION line ( $Y = bX + a$ )**



# Coefficient of Correlation

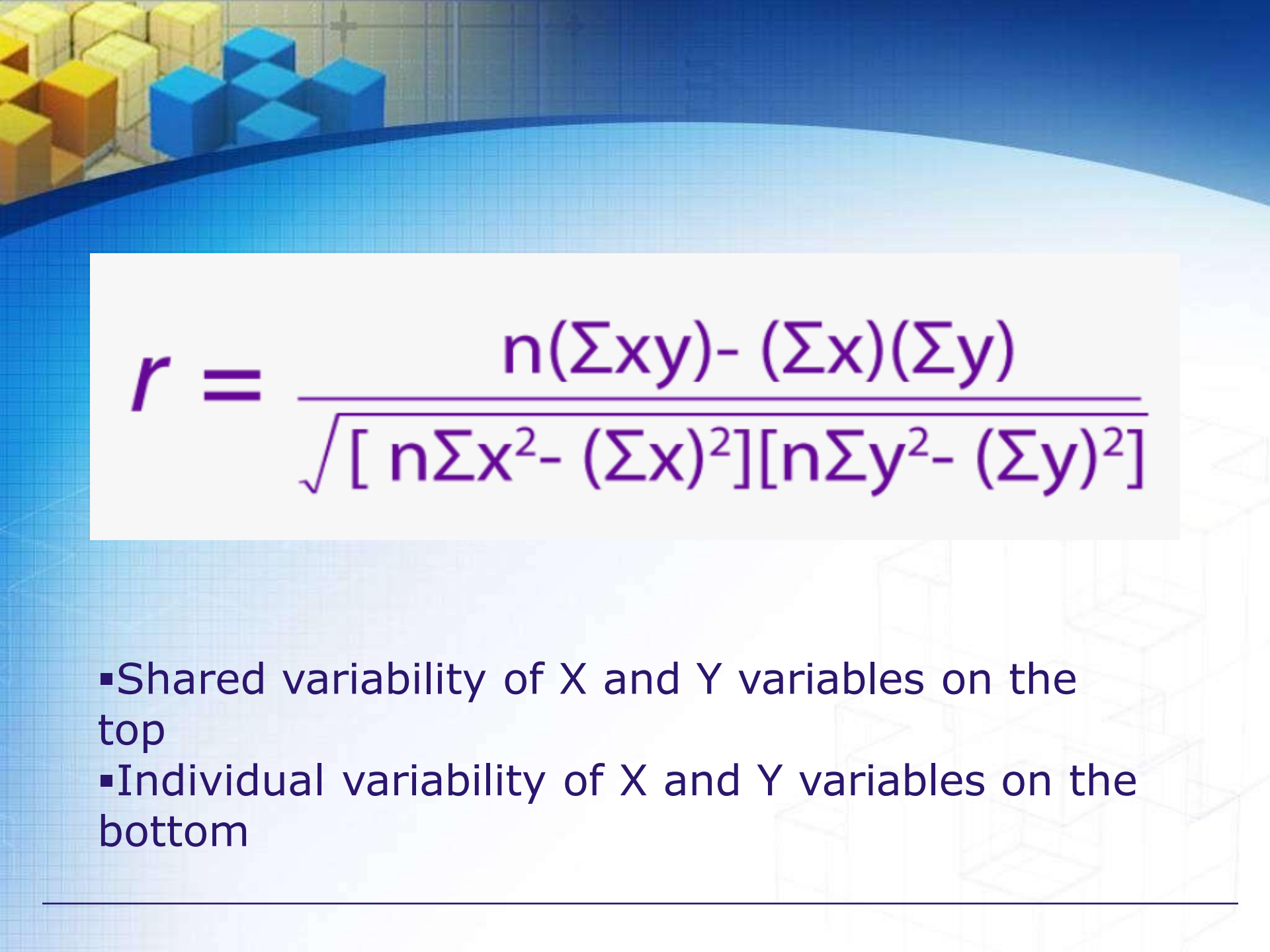
- ❖ A measure of the strength of the linear relationship between two variables that is defined in terms of the (sample) covariance of the variables divided by their (sample) standard deviations
  - ❖ Represented by “ $r$ ”
  - ❖  $r$  lies between  $+1$  to  $-1$
  - ❖ Magnitude and Direction
-



❖  $-1 \leq r \leq +1$

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




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

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

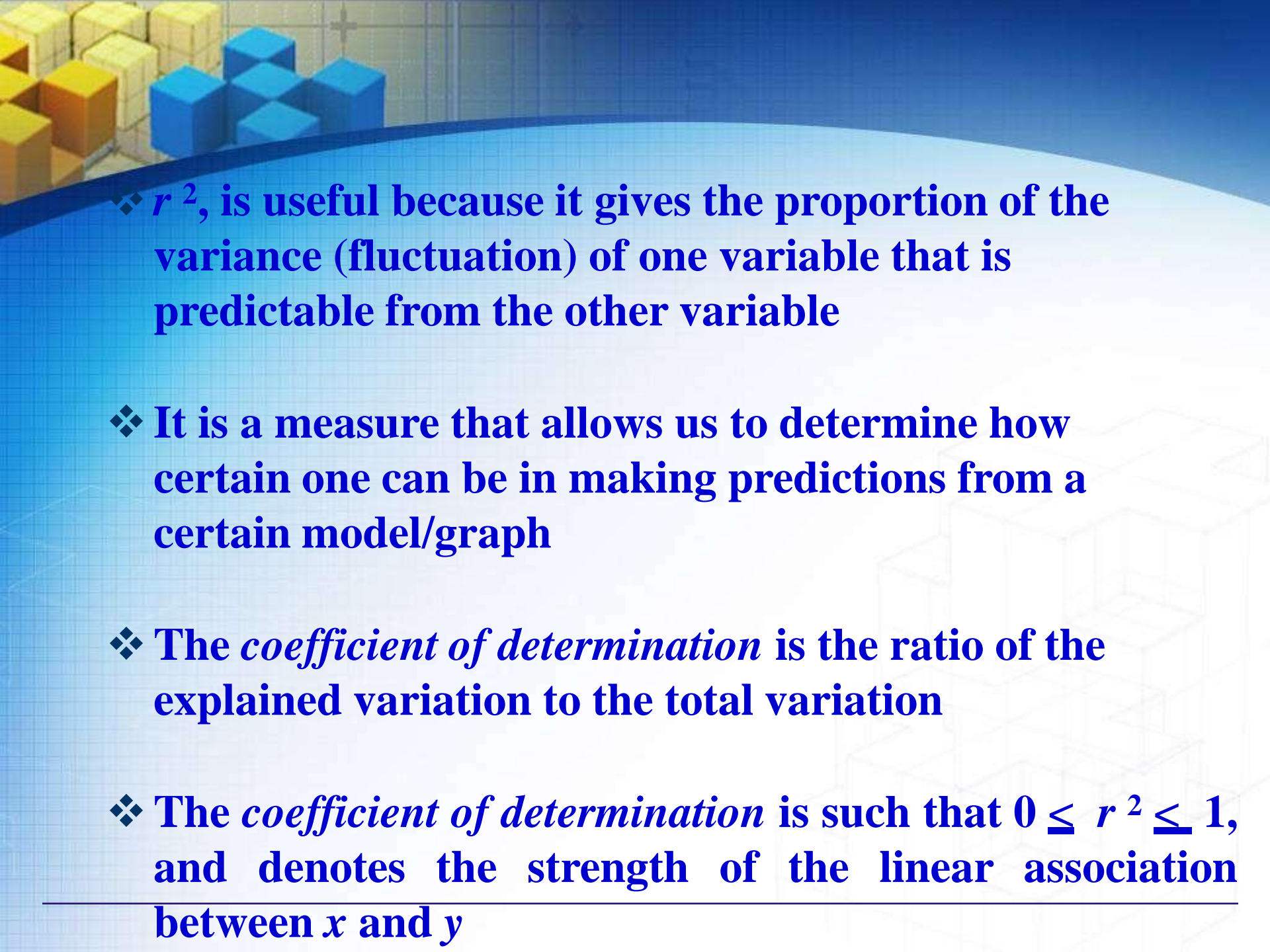


# Interpreting Correlation Coefficient $r$

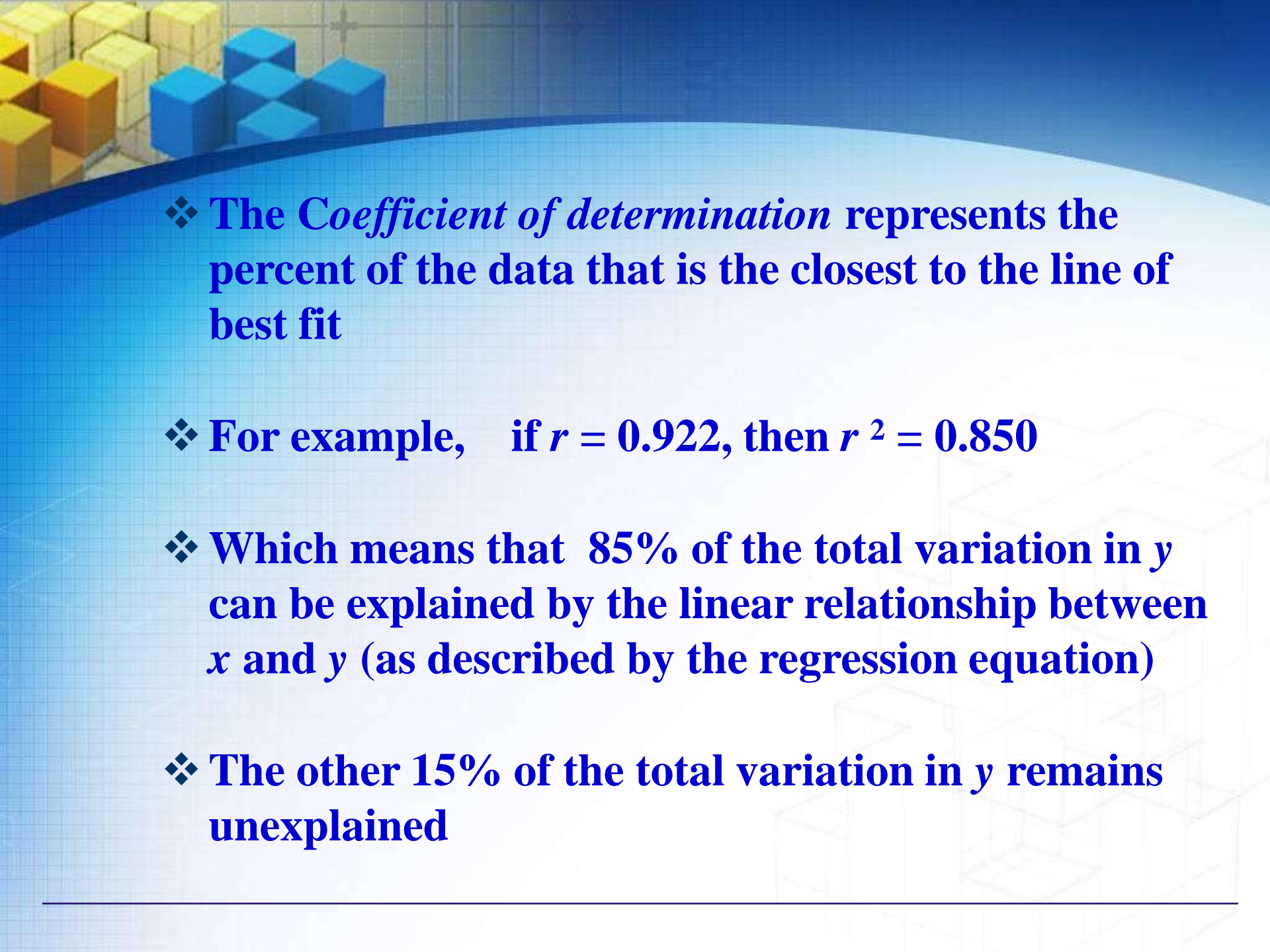
- **strong correlation:**  $r > .70$  or  $r < -.70$
  - **moderate correlation:**  $r$  is between  $.30$  &  $.70$   
or  $r$  is between  $-.30$  and  $-.70$
  - **weak correlation:**  $r$  is between  $0$  and  $.30$   
or  $r$  is between  $0$  and  $-.30$  .
-

# 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 *Coefficient 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**
-



# Spearman's rank coefficient

- ❖ **A method to determine correlation when the data is not available in numerical form and as an alternative the method, the method of rank correlation is used. Thus when the values of the two variables are converted to their ranks, and there from the correlation is obtained, the correlations known as rank correlation.**
-

# Computation of Rank Correlation

## Spearman's rank correlation coefficient

**$\rho$  can be calculated when**

- Actual ranks given
- Ranks are not given but grades are given but not repeated
- Ranks are not given and grades are given and repeated

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$