



# Correlation

By


Amritansh



# Correlation

- ▶ Finding the relationship between two quantitative variables without being able to infer causal relationships
- ▶ Correlation is a statistical technique used to determine the degree to which two variables are related

$$X = Y$$



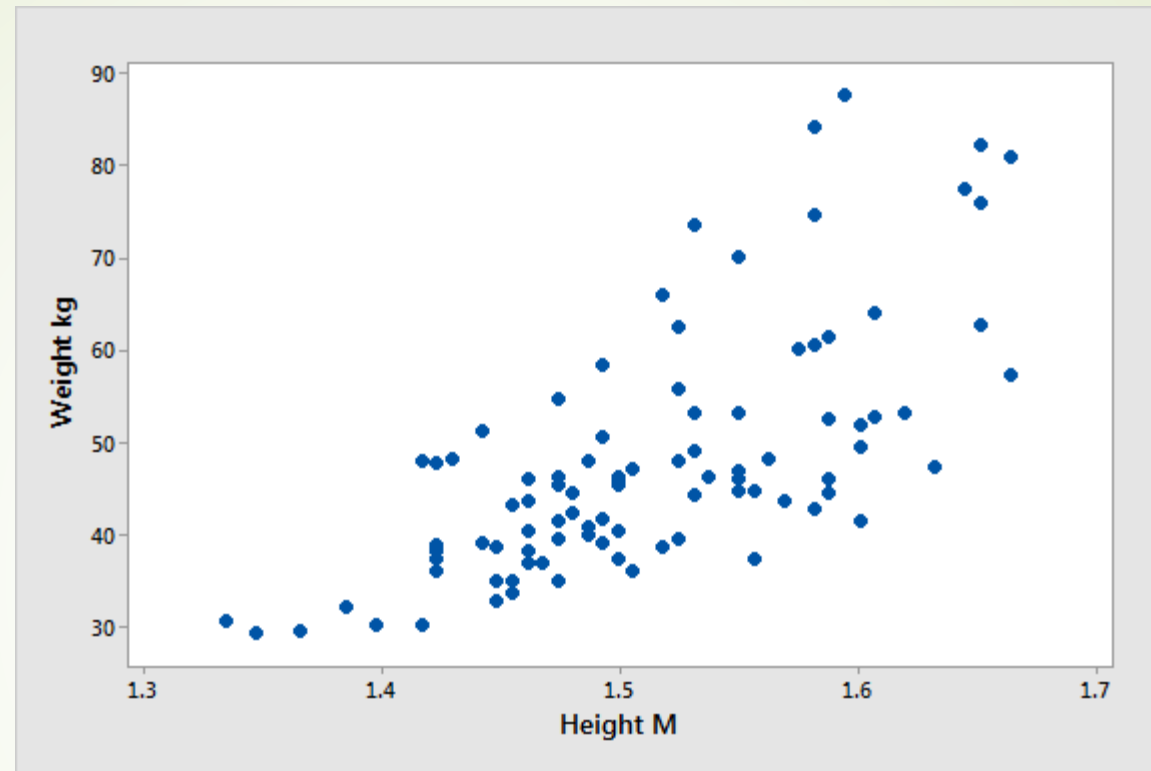
For example, height and weight are related; taller people tend to be heavier than shorter people.

The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one.

Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc.

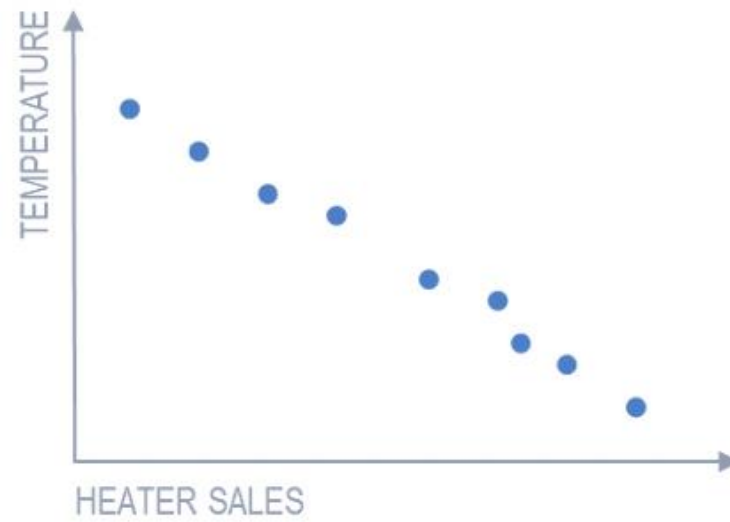
Correlation can tell you just how much of the variation in peoples' weights is related to their heights.



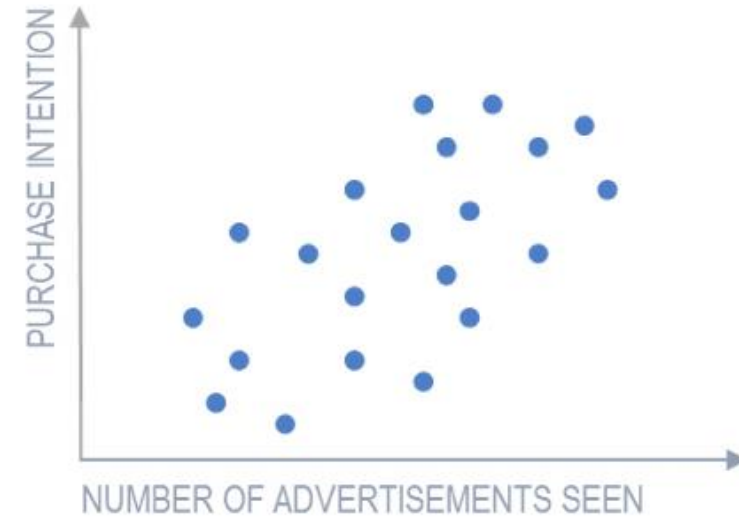


Do you see any Correlation ?

**Strong negative correlation**

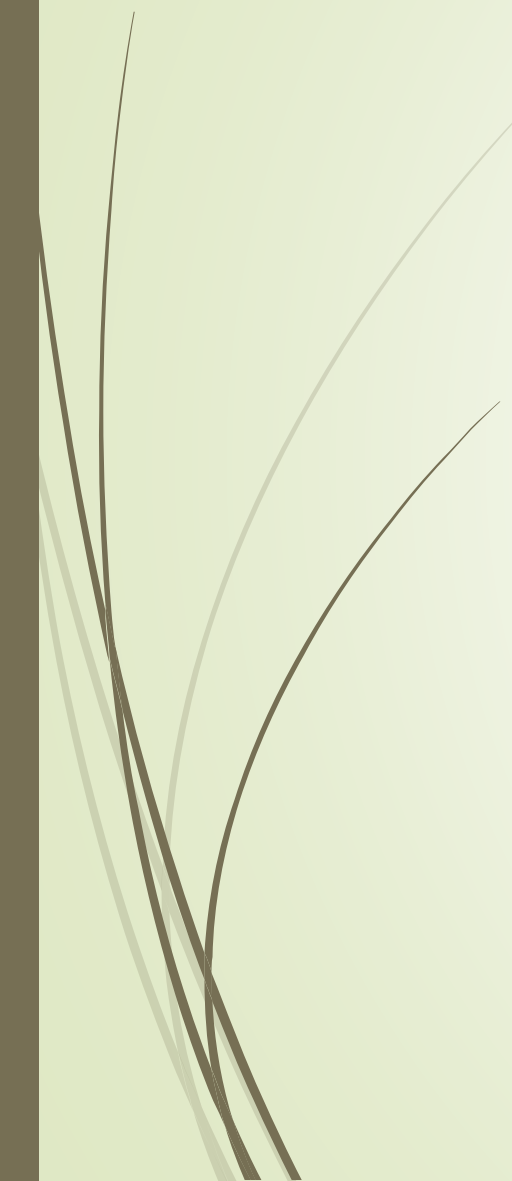


**Weak positive correlation**





# Pearson's Correlation Coefficient

- The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by  $r$
  - Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient,  $r$ , indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit)
  - The Pearson correlation coefficient,  $r$ , can take a range of values from +1 to -1
- 


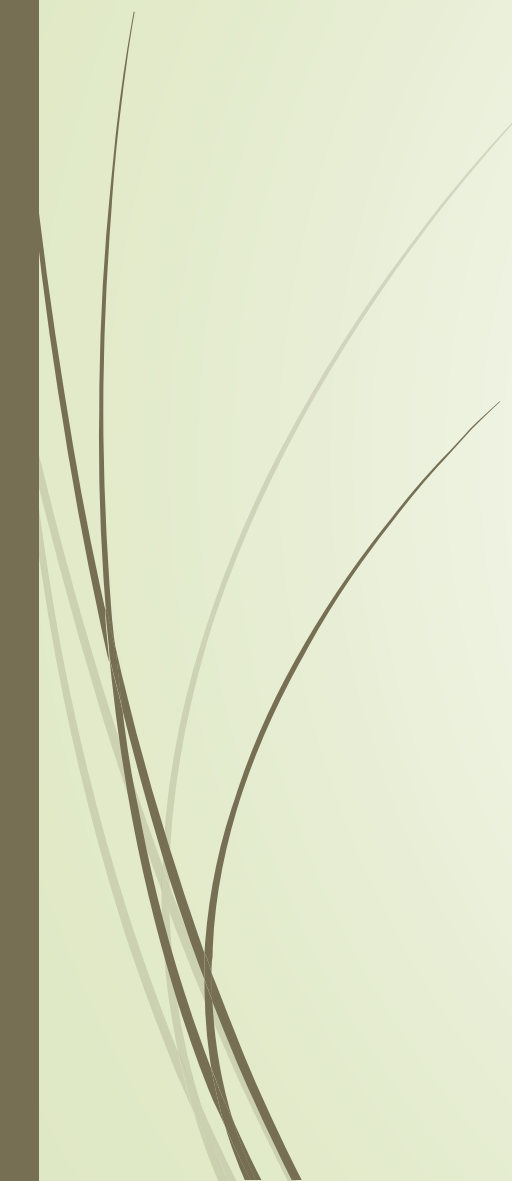


# Assumptions

- There are five assumptions that are made with respect to Pearson's correlation:
  - The variables must be either interval or ratio measurements
  - The variables must be approximately normally distributed
  - There is a linear relationship between the two variables
  - Outliers are either kept to a minimum or are removed entirely
  - There is homoscedasticity of the data

Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables

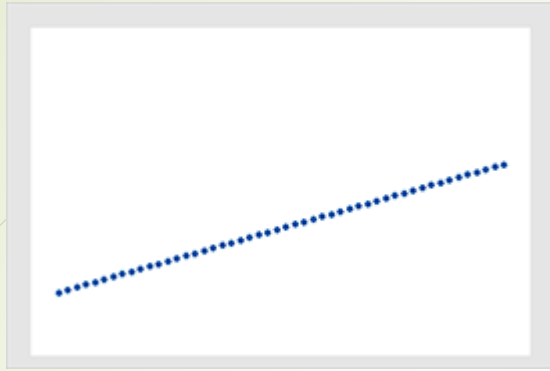


- 
- 
- Pearson's correlation coefficient is represented by the Greek letter rho ( $\rho$ ) for the population parameter and  $r$  for a sample statistic. This coefficient is a single number that measures both the strength and direction of the linear relationship between two continuous variables



- 
- ▀ Values can range from -1 to +1.

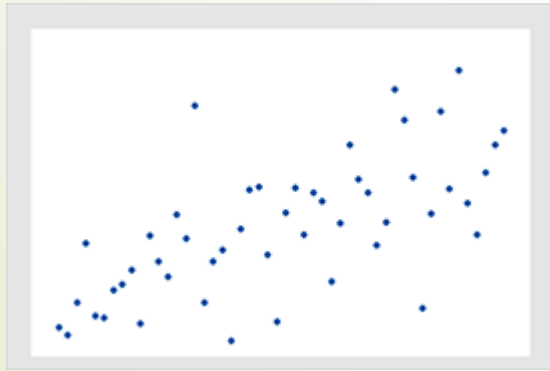
- ▀ **Strength:** The greater the absolute value of the coefficient, the stronger the relationship. The extreme values of -1 and 1 indicate a perfectly linear relationship where a change in one variable is accompanied by a perfectly consistent change in the other. For these relationships, all of the data points fall on a line. In practice, you won't see either type of perfect relationship
- ▀ A coefficient of zero represents no linear relationship. As one variable increases, there is no tendency in the other variable to either increase or decrease
- ▀ When the value is in-between 0 and +1/-1, there is a relationship, but the points don't all fall on a line. As  $r$  approaches -1 or 1, the strength of the relationship increases and the data points tend to fall closer to a line
- ▀ **Direction:** The coefficient sign represents the direction of the relationship. Positive coefficients indicate that when the value of one variable increases, the value of the other variable also tends to increase. Positive relationships produce an upward slope on a scatterplot.
- ▀ Negative coefficients represent cases when the value of one variable increases, the value of the other variable tends to decrease. Negative relationships produce a downward slope.



+1 : A perfect positive correlation



+0.8 : A strong positive correlation



+0.6 : A moderate positive correlation



0 : No correlation



- 1 : A perfect negative correlation



-0.8 : A high negative correlation



-0.6 : A moderate negative correlation

# How to Calculate Pearson Coefficient

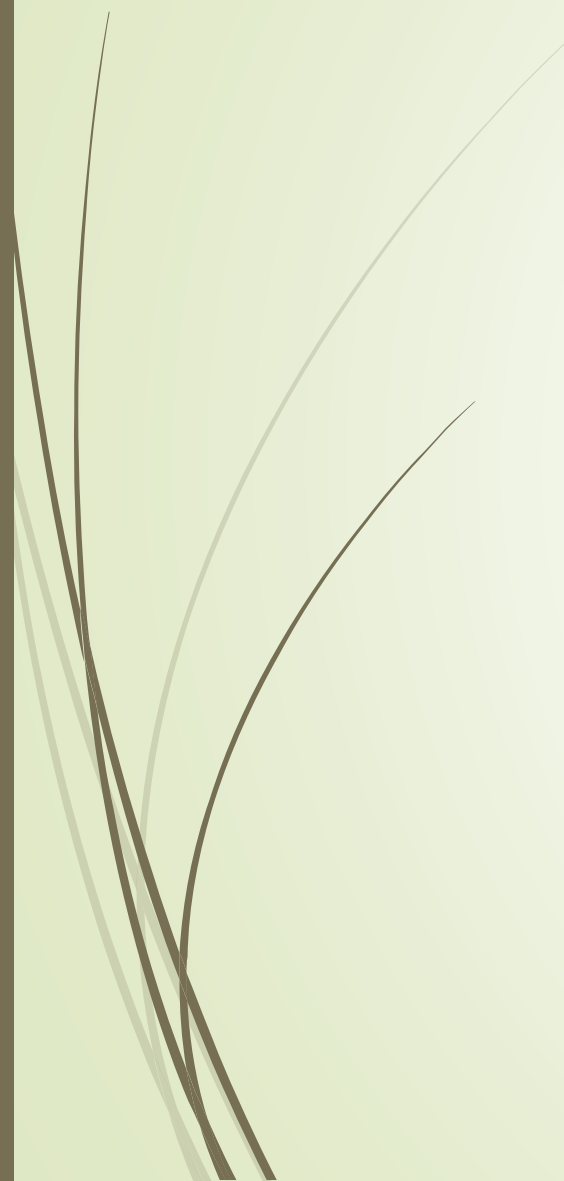

Index	Variable 1 (X)	Variable 2 (Y)	XY	X <sup>2</sup>	Y <sup>2</sup>
1					
2					
3					

Step 1: Make the following table and fill in with all calculations

# How to Calculate Pearson Coefficient

Index	Variable 1 (X)	Variable 2 (Y)	XY	X <sup>2</sup>	Y <sup>2</sup>
1					
2					
3					
$\Sigma$					

Step 2: Do a Summation of all rows for each column individually


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

Step 3: Now plug in all the values obtained previously

Where,  $n$  = no. of pairs of scores

# Question

- 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 . Find the correlation coefficient ( $r$ ) between age and weight

Sl no.	Age (Years)	Weight(Kgs)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13