CORRELATION
&
REGRESSION

Flow of Presentation

- Correlation
  - Definition
  - Types of correlation
  - Method of studying correlation (Scatter Diagram, Karl Pearson’s correlation and Spearman’s rank correlation)

- Regression
  - Definition
  - Regression lines
  - Methods to find regression lines (Scatter diagram and Least square method)
  - Multiple Regression

Correlation

- **The degree of relationship** between the variables under consideration **is measure** through the **correlation analysis**.

- The measure of correlation called the correlation coefficient. It is denoted by ‘r’.

- The degree of relationship is expressed by coefficient which range from correlation **(-1 ≤ r ≤ 1)**

- The correlation analysis enable us to have an idea about the **degree & direction** of the relationship between the two variables under study.

Correlation & Causation

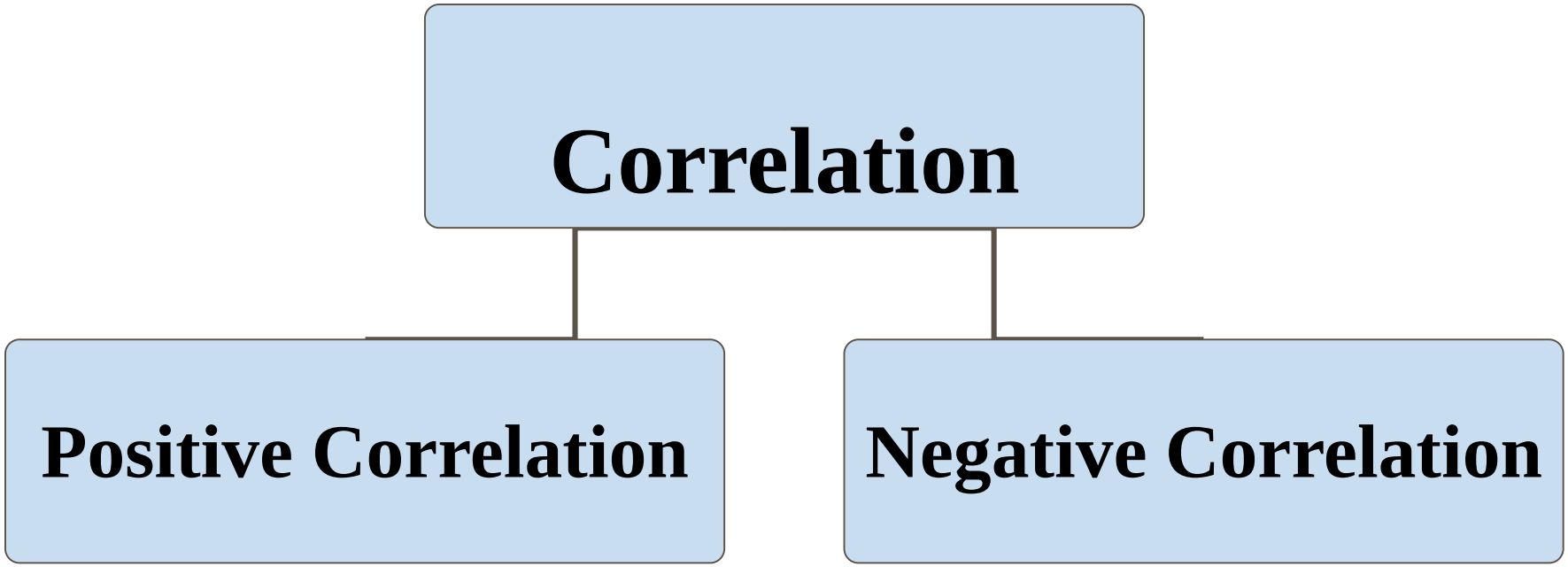
- Causation means **cause & effect** relation. Change in one variable accompanied by others is called causation

- Correlation **denotes the interdependency among the variables** for correlating two phenomenon having causation

- Causation implies correlation but reverse is not true

Types of Correlation

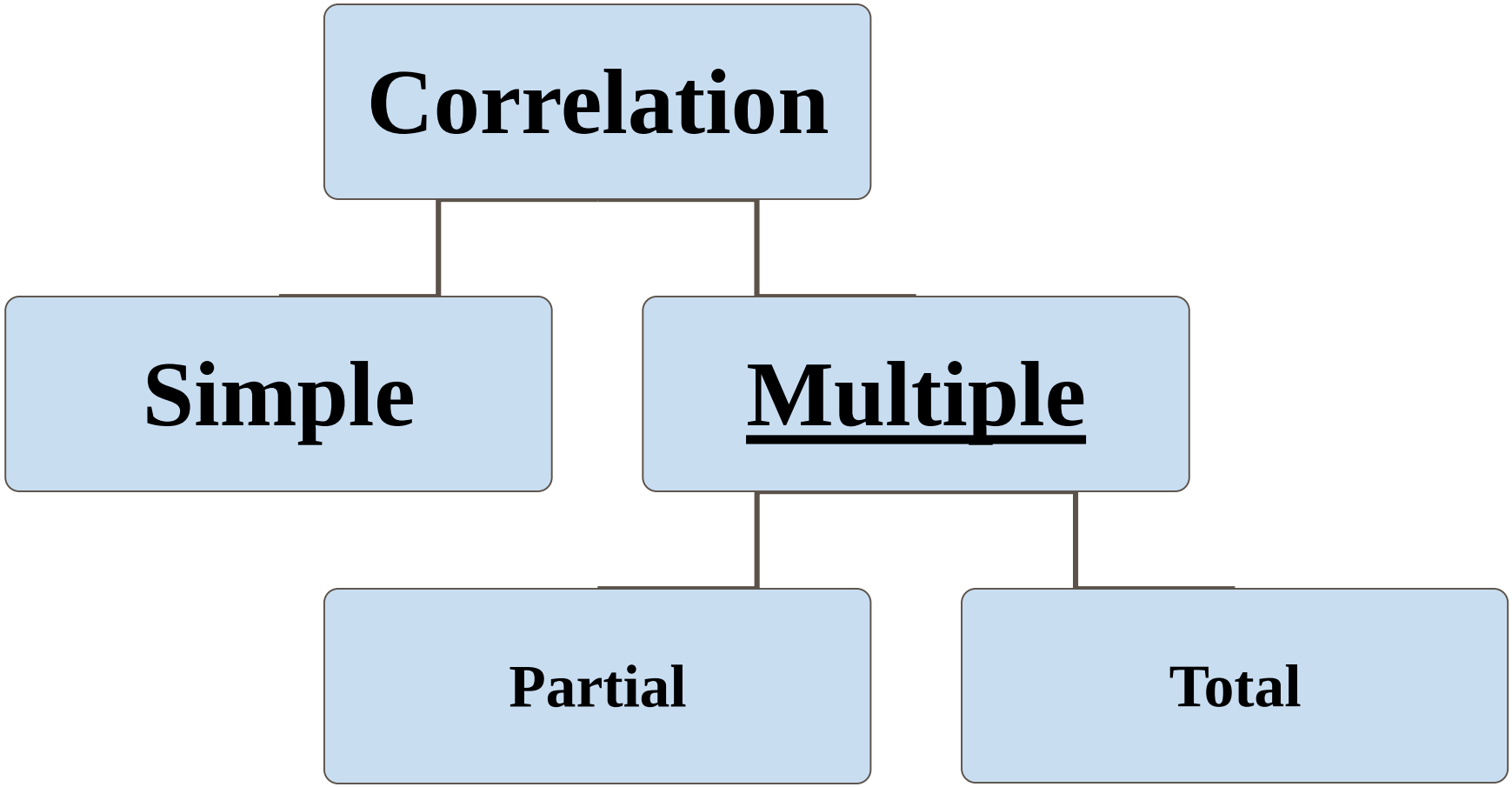
Type I: Direction of the Correlation



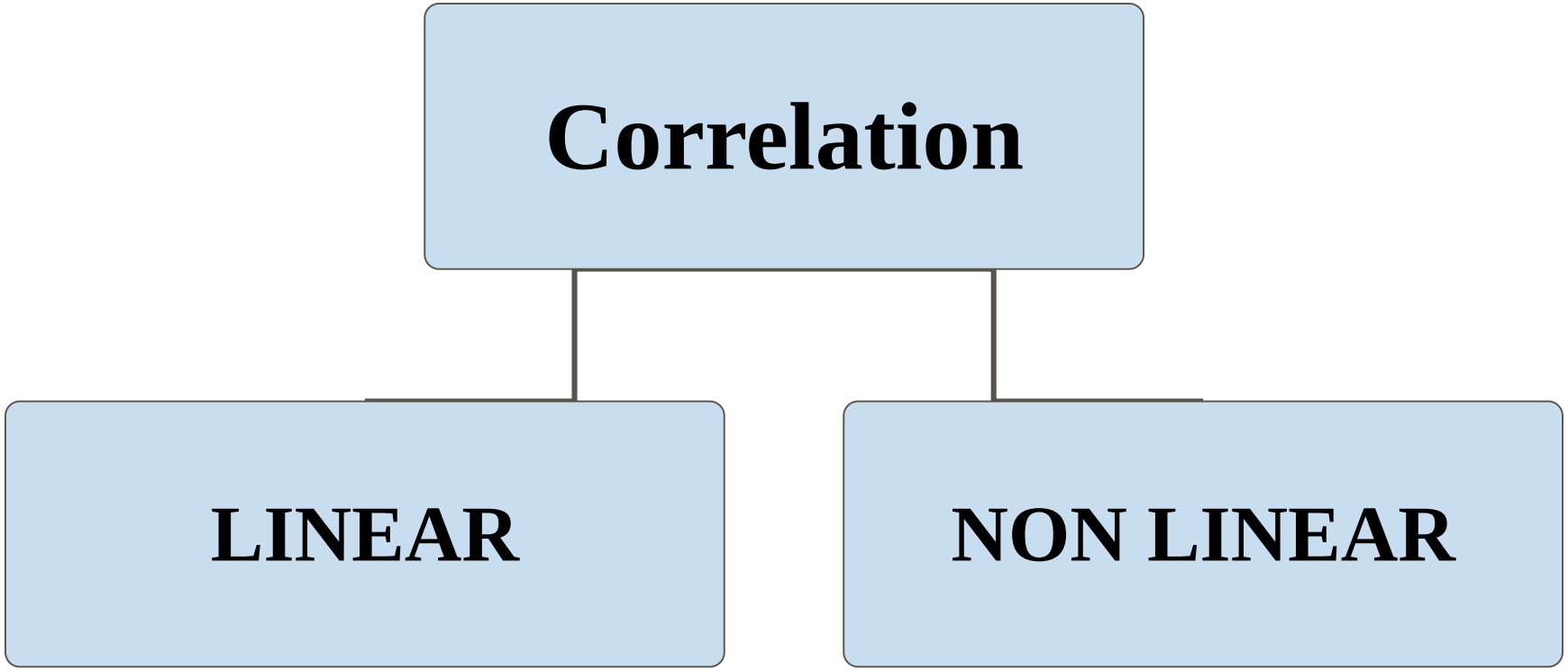
Examples:

- **Positive relationships**
  - water consumption and temperature
  - study time and grades
- **Negative relationships**
  - alcohol consumption and driving ability
  - Price & crops

Type II: Number of variables considered



Type III: Relationship assumed



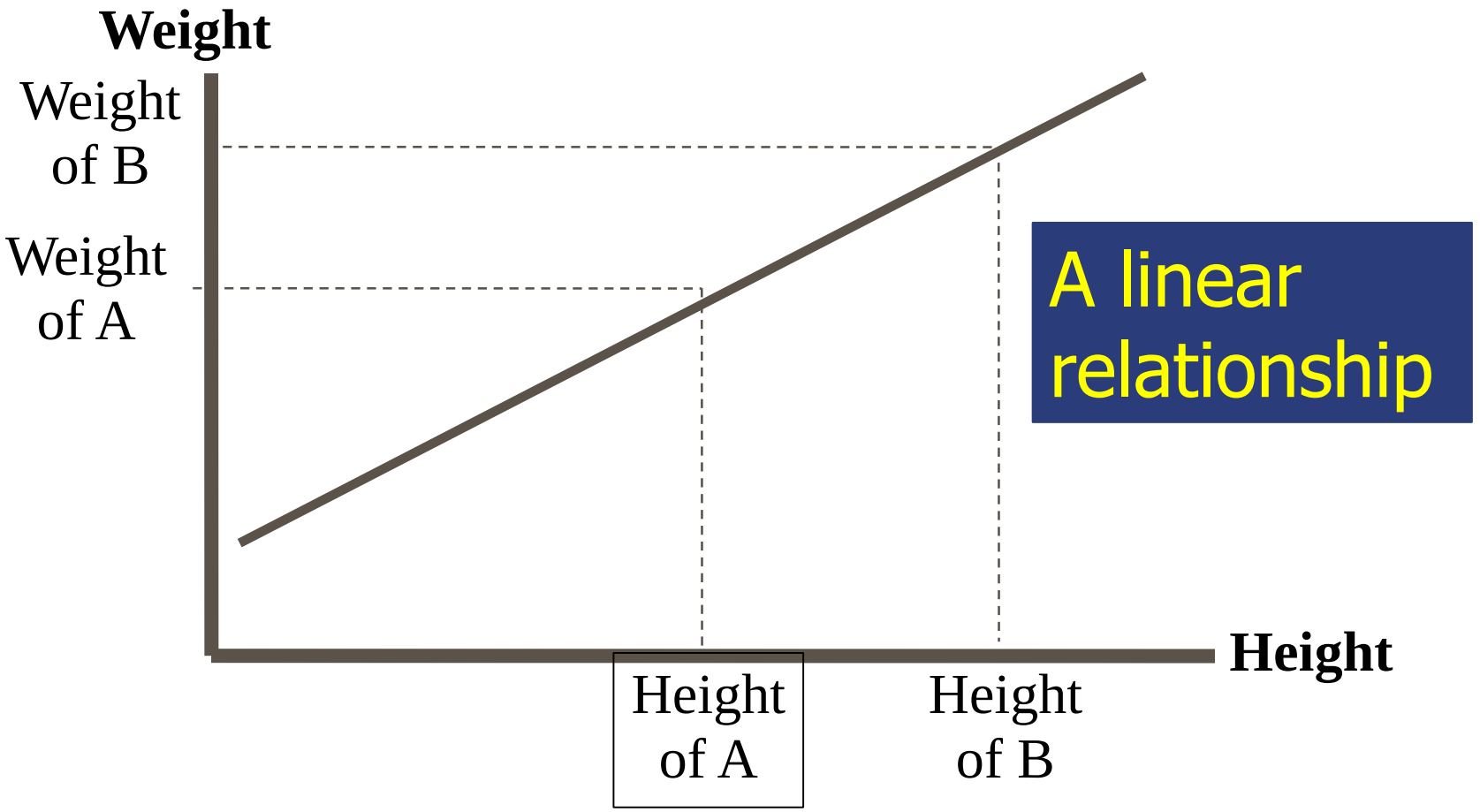
Methods of Studying Correlation

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

Scatter Diagram Method

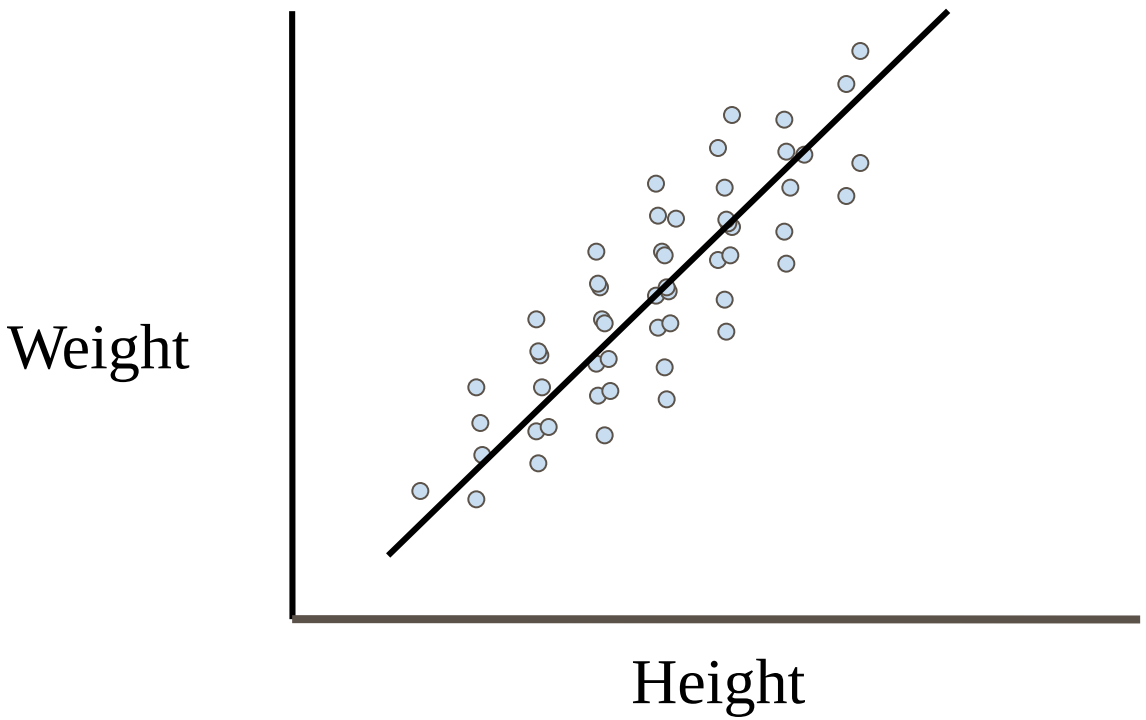
- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate.
- It portrays the relationship between these two variables graphically.

A perfect positive correlation



High Degree of positive correlation

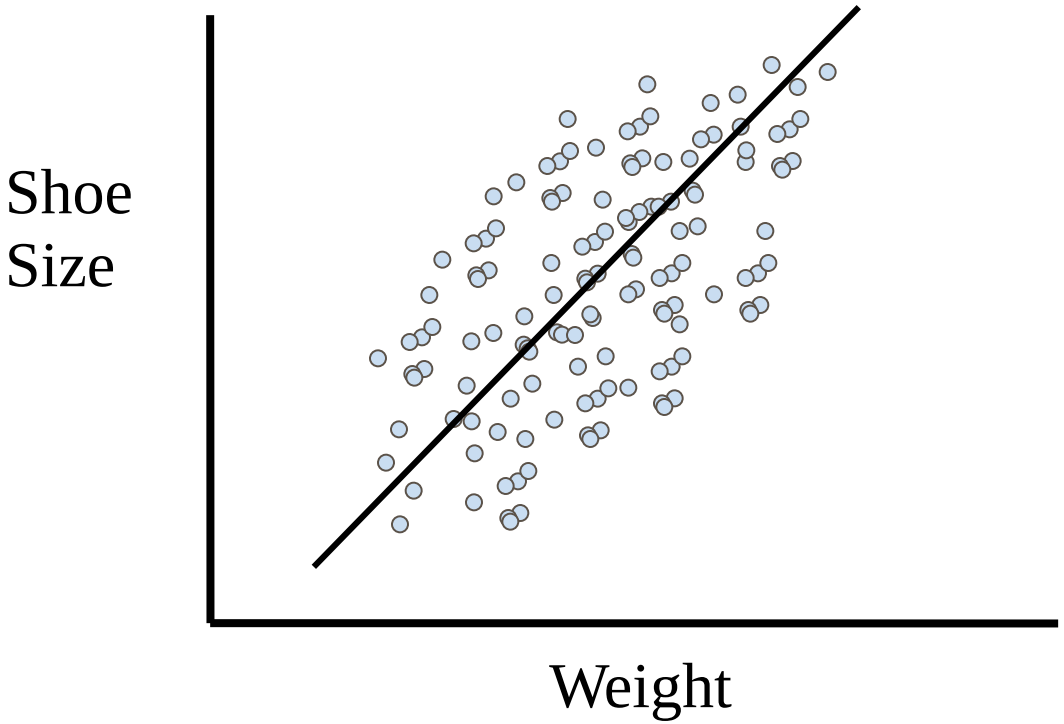
- Positive relationship



r = +.80

Degree of correlation

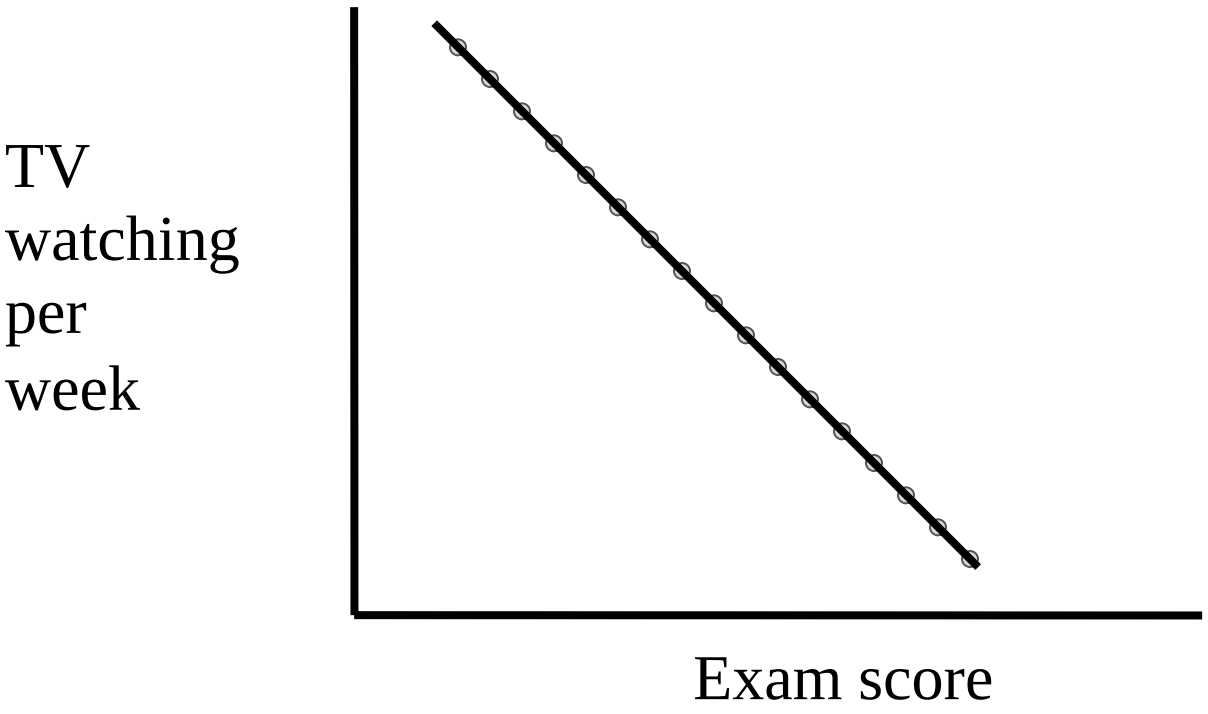
- **Moderate Positive Correlation**



r = + 0.4

Degree of correlation

- **Perfect Negative Correlation**

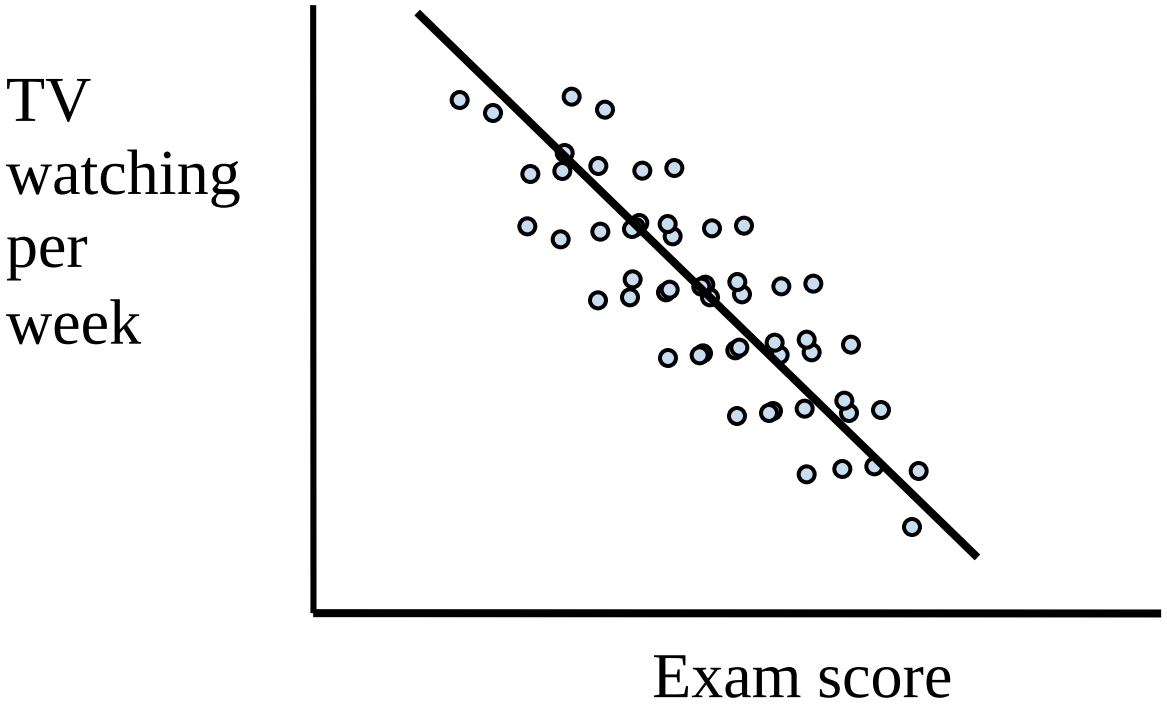


r = -1.0

Degree of correlation

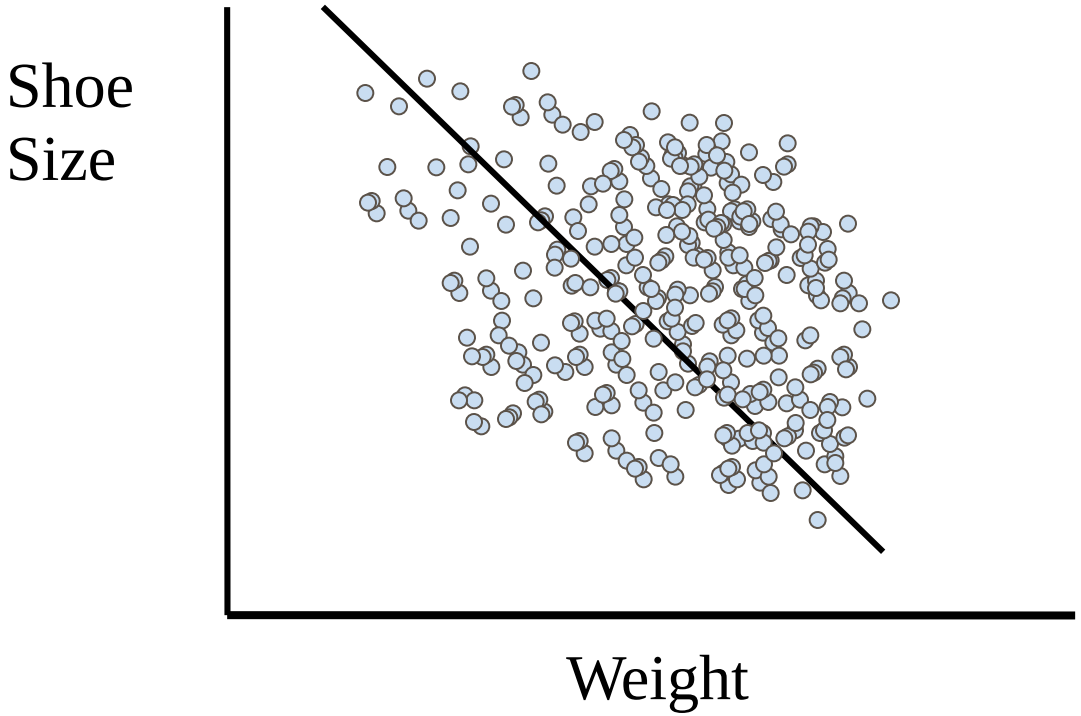
- **Moderate Negative Correlation**

r = -.80



Degree of correlation

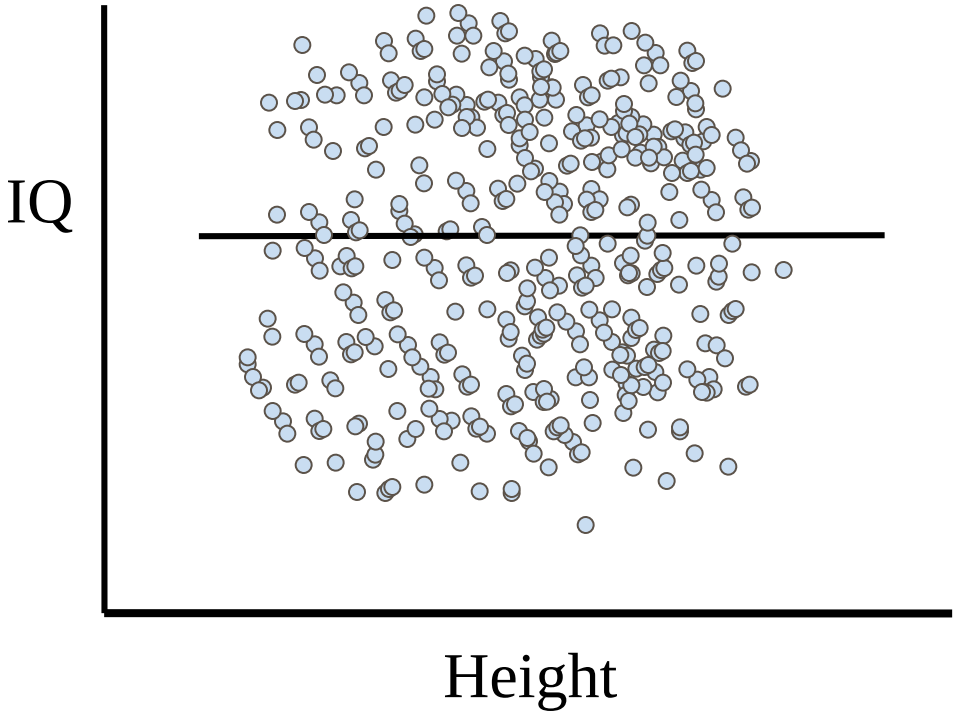
- **Weak negative Correlation**



r = - 0.2

Degree of correlation

- **No Correlation (horizontal line)**



r = 0.0

Advantages and Disadvantage of Scatter Diagram

Advantages:

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage:

- Can not adopt the an exact degree of correlation

Karl Pearson's Coefficient of Correlation

- Pearson's correlation coefficient (r) is the most common correlation coefficient.

- The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y.

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$$

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

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

Interpretation of Correlation Coefficient (r)

- If $r = +1$, then the correlation between the two variables is said to be perfect and positive
- If $r = -1$, then the correlation between the two variables is said to be perfect and negative
- If $r = 0$, then there exists no correlation between the variables
- If $0<r<1$, then the correlation between the two variables is said to be partial and positive
- If $-1<r<0$, then the correlation between the two variables is said to be partial and negative

Assumptions of Pearson's Correlation Coefficient

- There is **linear relationship** between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- **Cause and effect relation exists** between different forces operating on the item of the two variable series.

Example: Find the correlation coefficient using Karl Pearson's method for the following data.

x	6	2	10	4	8
y	9	11	5	8	7

Solution:

X	Y	XY	X*X	Y*Y
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
30	40	214	220	340

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

$$= -0.91$$

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also

Limitation of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult
- Value of Correlation Coefficient is affected by the extreme values
- Time consuming methods

Spearman's Rank Coefficient of Correlation

- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation Pearson's correlation coefficient can not be used in such case Spearman's Rank correlation can be used.

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)},$$

$$where, d = Rx - Ry, n : no\ of\ obs$$

$$Rx : Rank\ of\ x, Ry : Rank\ of\ y$$

Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from -1 to +1
- If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction
- If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction
- If R = 0, then there is no correlation

Rank Correlation Coefficient (R)

- **Equal Ranks or tie in Ranks:** In such cases **average ranks** should be assigned to each individual and

$$R = 1 - \frac{6\sum D^2 + AF}{n(n^2 - 1)}$$

$$AF = \frac{1}{12}(m_1(m_1^{\ 2} - 1)) + \frac{1}{12}(m_2(m_2^{\ 2} - 1)) + ...$$

m_i: The number of time an item is repeated

AF: Adjustment Factor

Example:

Find the Spearman's rank correlation coefficient for the following data

x	**39**	**65**	**62**	**90**	**82**	**75**	**25**	**98**	**36**	**78**
y	47	53	58	86	62	68	60	91	51	84

Solution:

X	y	Rx	Ry	D	D*D
39	47	8	10	-2	4
65	53	6	8	-2	4
62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
78	84	4	3	1	1

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

$$= 0.82$$

Exp. 2:

- A physiologist wants to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs have approximately equal scores in an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them as follows

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Solution:

A	B	R_A	R_B	D	D*D
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	-1	1
14	26	10	4	6	36
30	23	1.5	5	-3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	-3	9
30	20	1.5	7	-5.5	30.25
20	16	7	9.5	-2.5	6.25
28	11	4	11	-7	49
11	21	11	6	5	25

$$R = 1 - \frac{6\sum D^2 + AF}{n(n^2 - 1)}$$

$$AF = \frac{1}{12}(2(2^2 - 1)) + \frac{1}{12}(2(2^2 - 1)) + \frac{1}{12}(2(2^2 - 1))$$

$$= -0.02$$

Advantages of Spearman's Rank Correlation

- This method is simpler to understand and easier to apply compared to Karl Pearson's correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Disadvantages of Spearman's Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study
- Easier to collect co relational data

Regression

It is a statistical technique that defines the functional relationship between two variables.

Regression lines (linear regression)

Methods:

- Scatter Diagram
- Least square method

Regression Line y on x:

$$(y-\bar{y}) = b_{yx}(x-\bar{x})$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

Regression Line on y:

$$(x-\bar{x}) = b_{xy}(y-\bar{y})$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

Properties:

1. r * r = byx * bxy
2. byx and bxy both are of same sign
3. byx and bxy simultaneously can not exceed 1
4. Both lines intersect each other at means
5. If r = +- 1: lines coincides, r=0 : lines are perpendicular

Examples:

1. Obtain the regression lines using the following data and hence find the correlation coefficient.

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

2. In a partially destroyed laboratory records on the analysis of correlation data, only the following are legible:

$$\sigma^2_x = 0.09$$

$$40x - 18y - 214 = 0$$

$$8x - 10y + 66 = 0$$

Find means and std. deviation in y.