Correlation Analysis

Correlation

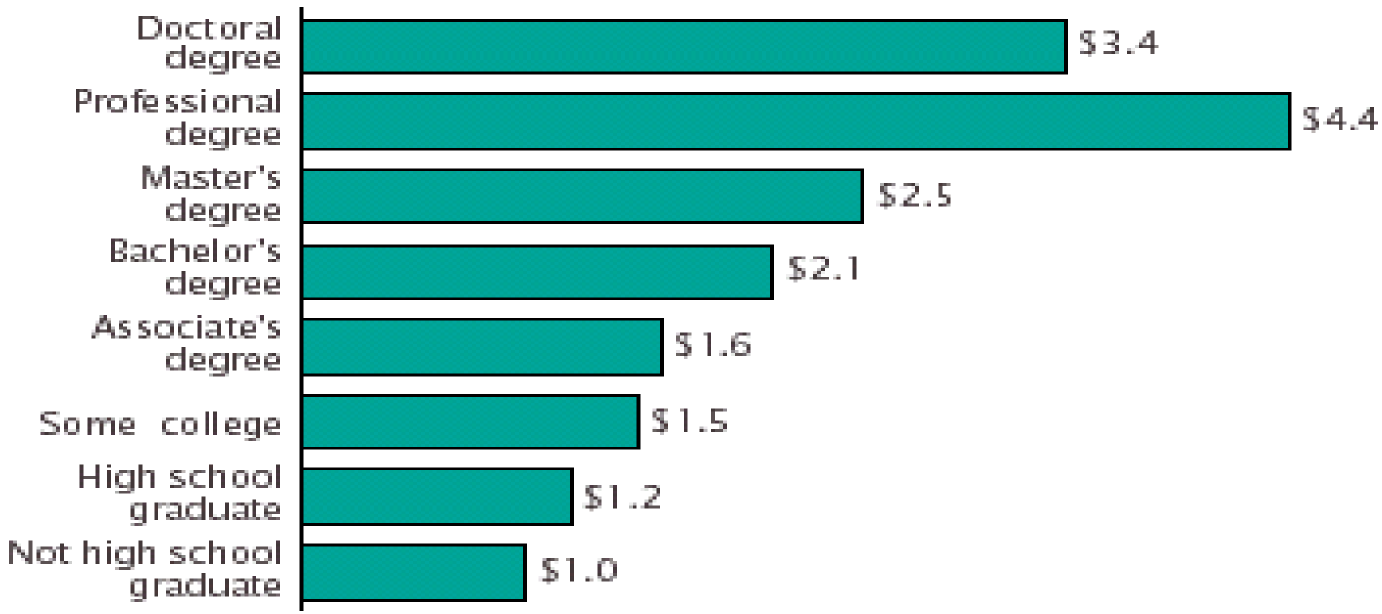
- Correlation is used to measure and describe a *relationship* between two variables.
- Measure of correlation called correlation coefficient which tells about the degree and direction of correlation.
- Correlation analysis measures the closeness of the relationship between variables.
- Ex- Husband & wife's age, sales of a company and expenditure on advertisement

Describing relationships: An example…

Figure 3.

Synthetic Work-Life Earnings Estimates for Full-Time, Year-Round Workers by Educational Attainment Based on 1997-1999 Work Experience

(In millions of 1999 dollars)



Source: U.S. Census Bureau, Current Population Surveys, March 1998, 1999, and 2000.

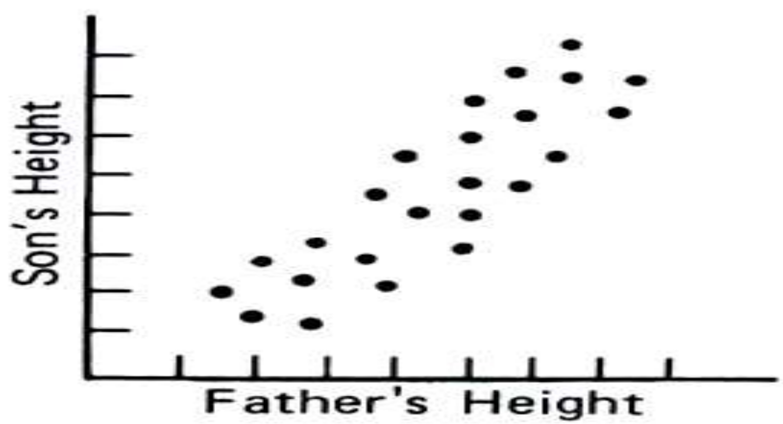
Correlation & Causation

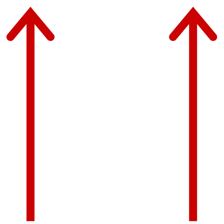
- Correlation $\not\longrightarrow$ Causation
- Causation $\longrightarrow$ Correlation

- Correlation may be coincidental especially in small samples.
- The relationship between variables may be caused by some third variable.
- Both the variables may be influencing each other so that neither can be designated as the cause and other as the effect.

Types of Correlation

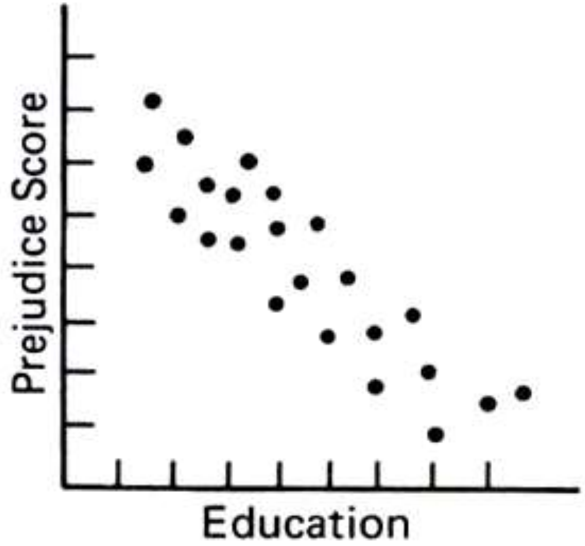
- Positive and negative correlation-
- Depends upon the direction of change of the variables.
- If both the variables are varying in same direction called positive correlation.
- X 2 4 6 8 10 OR X 50 40 30 20 10
- Y 1 3 5 7 9 Y 24 21 19 18 14





Negative Correlation

- The variables are varying in opposite directions.
- X 2 4 6 8 10 OR X 50 40 30 20 10
- Y 9 7 5 3 1 Y 24 26 28 30 32
-



Simple/Partial/Multiple Correlation-

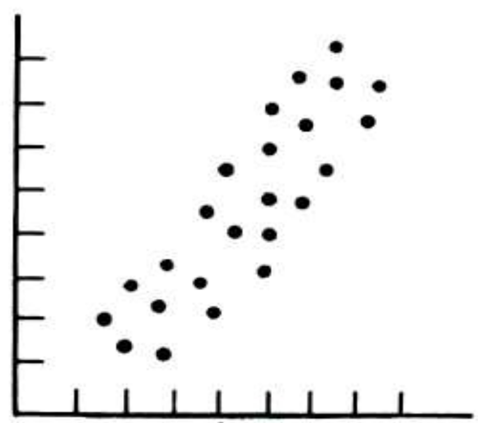
- Distinction between three depends on the number of variables studied.
- When only two variables are studied then simple.
- When three or more variables studied simultaneously then multiple.
- Recognize more than two variables but consider only two variables to be influencing each other and keeping other variables as constant, then partial.

Linear/ Non-Linear Relationship

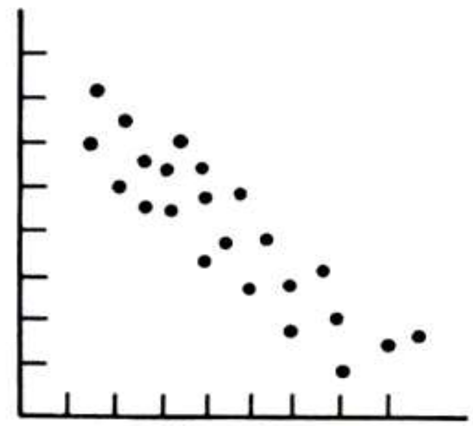
- Depends upon the constancy of the ratio of the change between the variables.
- If the amount of change in one variable tends to bear constant ratio to the amount of change in other variable then it is said to be linear.
- X 10 20 30 40 50
- Y 70 140 210 280 350
- If the amount of change in one variable does not bear a constant ratio to the amount of change in other variable then it is said to be non-linear.

Methods of Correlation

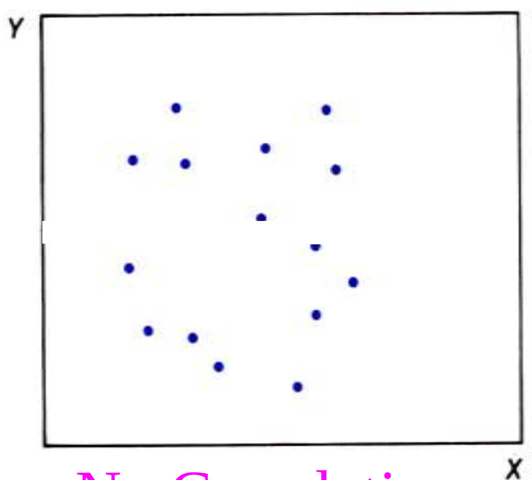
- Scatter Diagram Method
- Simplest device for ascertaining whether two variables are related is to prepare a dot chart.
- Greater the scatter of the plotted points, lesser the relationship between variables.



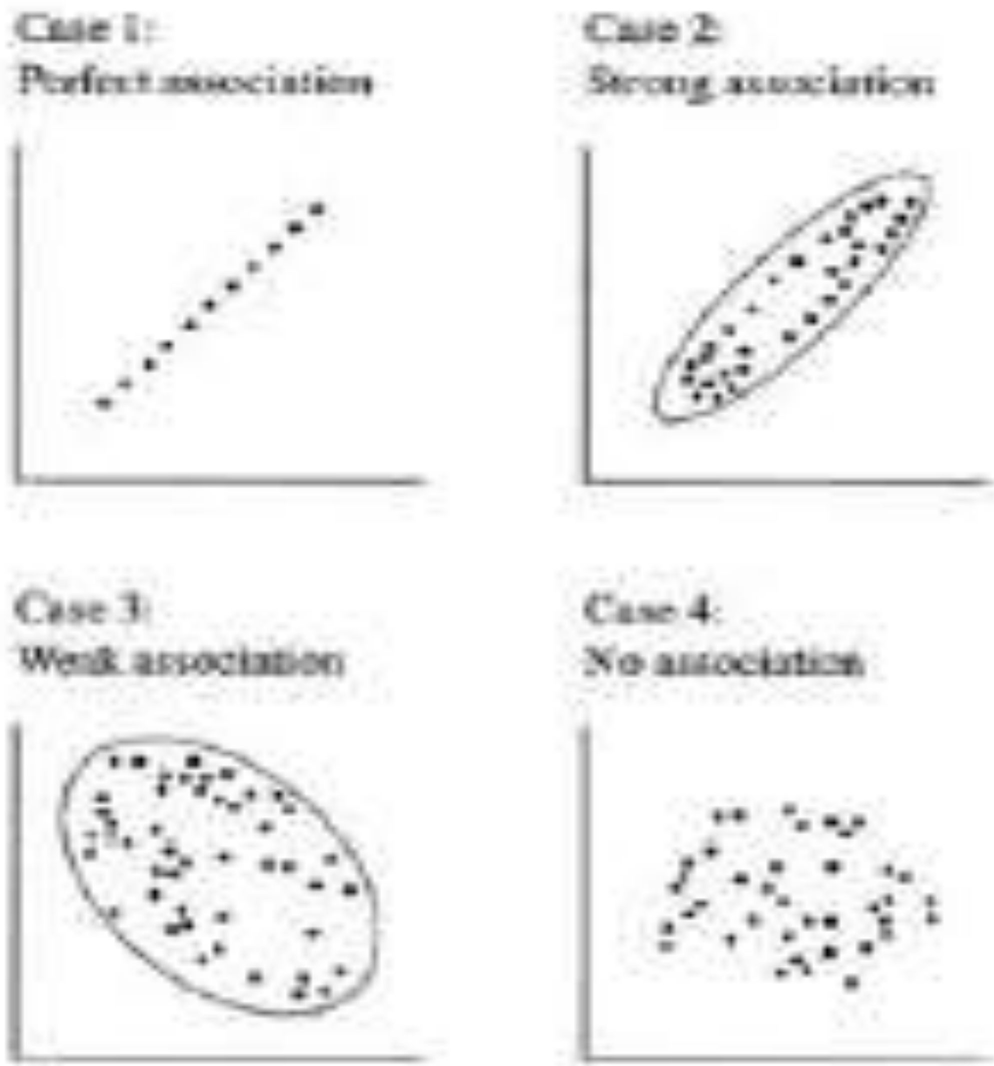
Positive



Negative



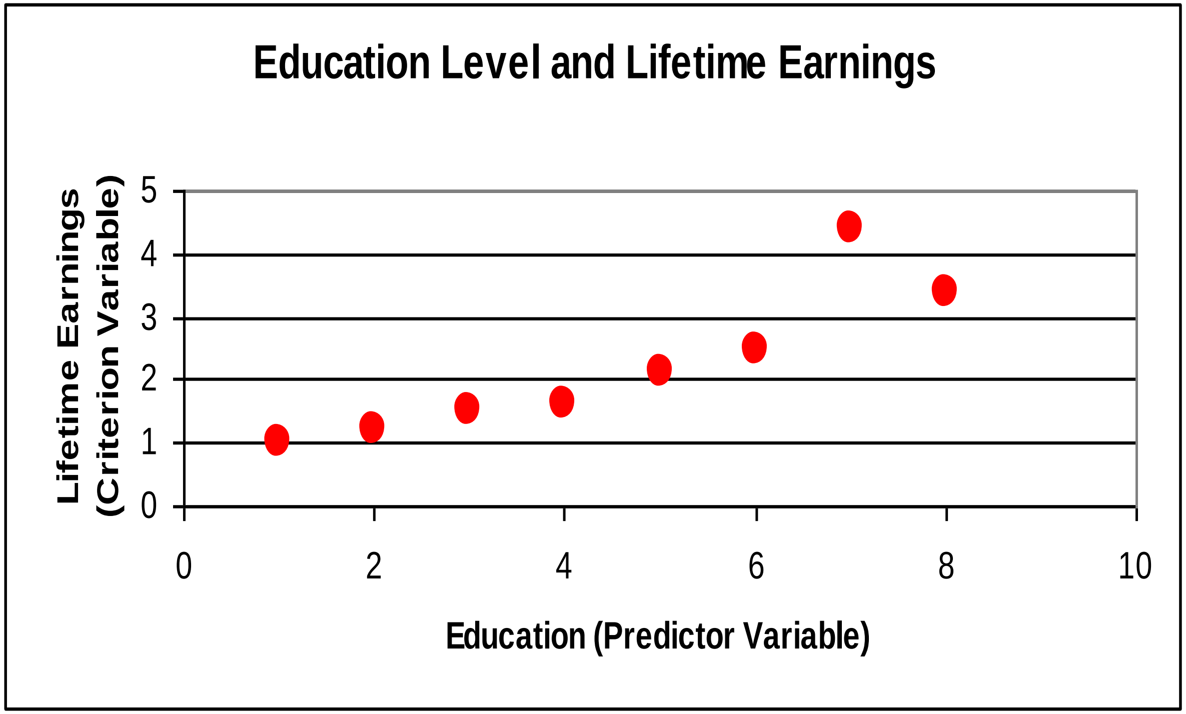
No Correlation


Case 1:
Perfect association
Case 2:
Strong association
Case 3:
Weak association
Case 4:
No association

Scatter Plot

- What is the relationship between level of education and lifetime earnings?

X (Education)	Y (Income)
8	3.4
7	4.4
6	2.5
5	2.1
4	1.6
3	1.5
2	1.2
1	1



Merits/ Demerits of Scatter Diagram

- Useful for gaining a visual impression of the relationship.
- Cant establish the exact degree of correlation between variables, so more quantitative description is needed
- Gives rough indication of nature and strength of relationship between variables.

Karl Pearson's Coefficient of correlation

- Measure of linear correlation
- Widely used method
- Pearsonian Correlation Coefficient is denoted by 'r'.
- The value of r lies between -1 and +1.

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

Pearson's *r*

- Definitional formula:

$$r = \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}}$$

$$r = \frac{COV_{XY}}{(s_x)(s_y)} \qquad COV_{XY} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{n}$$

Computational formula:

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

An Example: Correlation

X Education	Y Income	XY	X^2	Y^2
8	3.4	27.2	64	11.56
7	4.4	30.8	49	19.36
6	2.5	15	36	6.25
5	2.1	10.5	25	4.41
4	1.6	6.4	16	2.56
3	1.5	4.5	9	2.25
2	1.2	2.4	4	1.44
1	1	1	1	1
36	17.7	97.8	204	48.83

$$\sum X = 36$$

$$\sum Y = 17.7$$

$$\sum XY = 97.8$$

$$\sum X^2 = 204$$

$$\sum Y^2 = 48.83$$

$$n = 8$$

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

An Example: Correlation

$\sum X = 36$

$\sum Y = 17.7$

$\sum XY = 97.8$

$\sum X^2 = 204$

$\sum Y^2 = 48.83$

$n = 8$

$$r = \frac{(n)(\sum XY) - (\sum X)(\sum Y)}{\left[\sqrt{n\sum X^2 - (\sum X)^2}\right]\left[\sqrt{n\sum Y^2 - (\sum Y)^2}\right]}$$

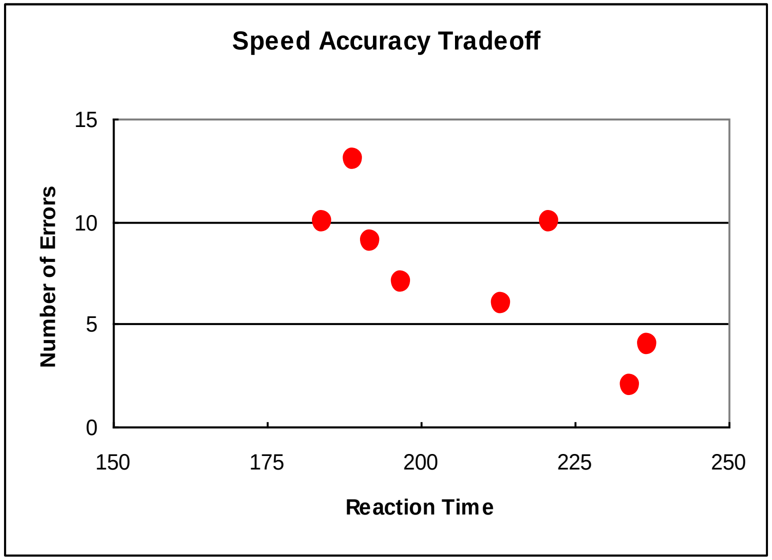
$$= \frac{(8)(97.8) - (36)(17.7)}{\left[\sqrt{8(204) - (36)^2}\right]\left[\sqrt{8(48.83) - (17.7)^2}\right]}$$

$$= .90$$

An Example: Correlation

- Researchers who measure reaction time for human participants often observe a relationship between the reaction time scores and the number of errors that the participants commit. This relationship is known as the speed-accuracy tradeoff. The following data are from a reaction time study where the researcher recorded the average reaction time (milliseconds) and the total number of errors for each individual in a sample of 8 participants. Calculate the correlation coefficient.

Reaction Time	Errors
184	10
213	6
234	2
197	7
189	13
221	10
237	4
192	9



An Example: Correlation

X	X^2	Y	Y^2	XY
184	33856	10	100	1840
213	45369	6	36	1278
234	54756	2	4	468
197	38809	7	49	1379
189	35721	13	169	2457
221	48841	10	100	2210
237	56169	4	16	948
192	36864	9	81	1728
1667	**350385**	**61**	**555**	**12308**

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

$$r = \frac{8(12308) - (1667)(61)}{\left(\sqrt{8(350385) - (1667)^2}\right)\left(\sqrt{8(555) - (61)^2}\right)}$$

$$= -0.77$$

Example-

Sales revenue & profit for cement companies for quarter July-Sept 2017-18.Find r

Company	Revenue (Rs. Crores)	Profit after tax (RS. Crores)
ACC	13	2.5
Grasim Industries	21	3.2
Guj Ambuja Cements	10	2.6
Ultratech Cement	9	1.4
Shree Cements	3	0.8
India Cements	5	1.1

Source: Economic Times , dt. 11th October 2006.

(Ans r =0.916)

Example

- The following table gives indices of industrial production and no. of registered unemployed people(in Lakhs.) Calculate the value of the correlation coefficient.
- Year 1991 92 93 94 95 96 97 98
- Index of prod. 78 89 99 60 59 79 68 61
- No. of
- Unemployed 125 137 156 112 107 136 123 108
- (Ans: r= 0.014)

Interpreting r

- How can we describe the strength of the relationship in a scatter plot?
  - A number between -1 and +1 that indicates the relationship between two variables.
    - The sign (- or +) indicates the direction of the relationship.
    - The number indicates the strength of the relationship.

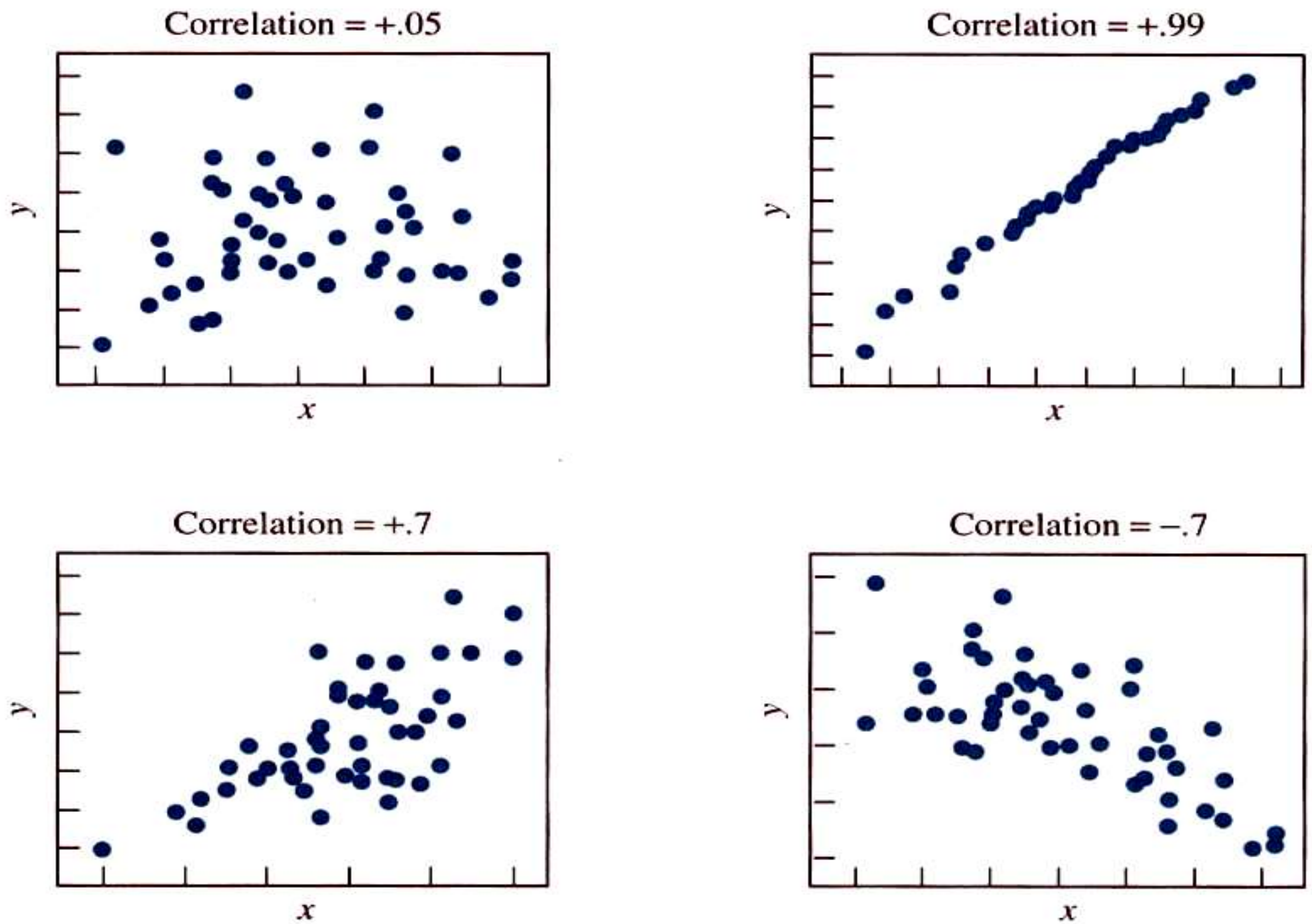
-1 ------------	0 ------------	+1
Perfect Relationship	No Relationship	Perfect Relationship

The closer to –1 or +1, the stronger the relationship.

- When r =+1, perfect positive relationship.
- When r =-1 ,perfect negative relationship.
- When r=0, no relationship
- Close to +1 or -1, closer the relationship between variables.
- Closer to 0, less close the relationship.
- The closeness of relationship is not proportional to r.

Size of Coefficient	General Interpretation
0.8 to 1.0	Very Strong Relationship
0.6 to 0.8	Strong relationship
0.4 to 0.6	Moderate relationship
0.2 to 0.4	Weak relationship
0.0 to 0.2	Very Weak or No relationship

Correlation Coefficient



Spearman's Rank Correlation Coefficient

- This method is useful for correlation analysis when variables are expressed in qualitative terms like beauty, judgment, intelligence, honesty etc.
- Spearman's Rank correlation coefficient is defined as

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

- Where R:Rank Correlation coefficient
- D: difference of rank between items of two series.
- N: no. of observations

When ranks are given-

	Rank as per final grade	Rank as per salary offered
A	1	1
B	2	3
C	3	2
D	4	4
E	5	6
F	6	5
G	7	9
H	8	8
I	9	10
J	10	7

When ranks are not given-

- Quotations of Index numbers of security prices of a certain joint stock company are given. Find r-

Year	Debenture Price	Share Price
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Equal ranks

$$R = 1 - \frac{6[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + ...]}{n(n^2 - 1)}$$

- m: number of times whose rank are common
- Obtain rank correlation coefficient between X & Y:-
- X: 50 55 65 50 55 60 50 65 70
- Y: 110 110 115 125 140 115 130 120 115

Practice Example

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60