Scatter Diagram

Scatter Plot is a picture of bivariate numerical data in which each observation (ie each pair of values (x,y)) is represented by a point located on a rectangular co-ordinate system. The Horizontal Axis is identified with values of x and the vertical axis with values of y.

Correlation

Pearson's Sample Correlation Coefficient, r, is a measure of the strength of the linear relationship between two variables x and y.

correlation is a term which implies that there is an association between the paired values of 2 variables, where association means that the fluctuations in the values for each variable is sufficiently regular to make it unlikely that the association has arisen by chance.

Correlation:

Correlation is a measure of the relation between two or more variables. The correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between two variables. So correlation analysis is a statistical procedure by which we can determine the degree of association or relationship between two or more variables.

- **Coefficient of correlation:**

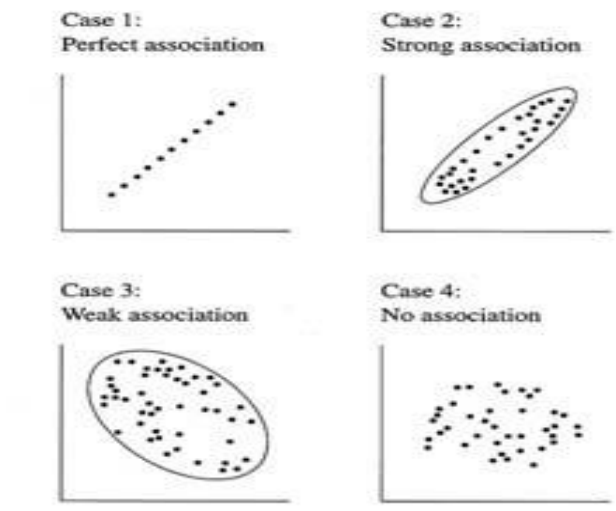
Coefficient of correlation is a measure of such a tendency, i.e. the degree to which the two variables are interrelated is measured by a coefficient which is called the coefficient of correlation.

Properties of r

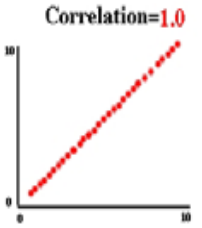
The correct interpretation of r requires an appreciation of some general properties:

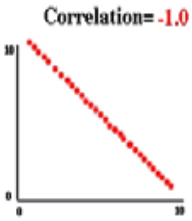
- The value of r does not depend on the unit of measurement for either variable, nor does it depend on which variable is labelled x or y.
- The value of r is between -1 and 1. ($-1 \leq r \leq 1$)
- A positive value of r ($r > 0$) indicates a positive linear relationship between the variables. So as x increases so does y.

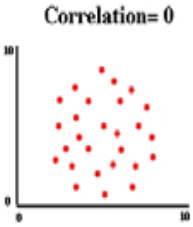
- A negative value of r (r < 0)corresponds to a negative relationship. As x increases y decreases.
- The value r = 1, which indicates the strongest possible positive relationship between x and y results only when all points in the scatter plot lie exactly on a straight line that slopes upward.
- The value r = -1, which indicates the strongest possible negative relationship between x and y results only when all points in the scatter plot lie exactly on a straight line that slopes downward.

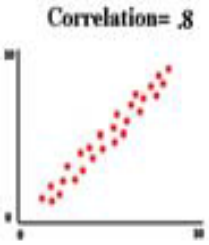


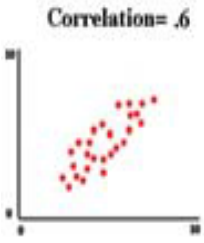
Generalized Scattergrams Showing Strength of Association

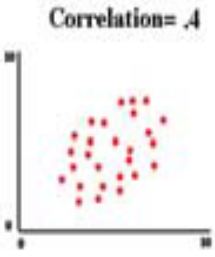


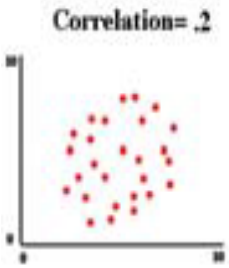


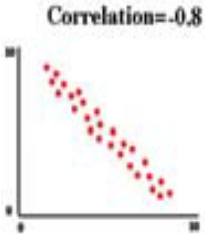


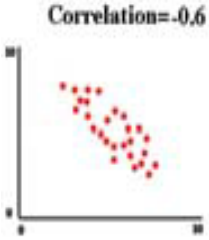


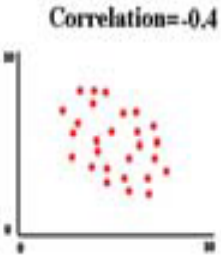


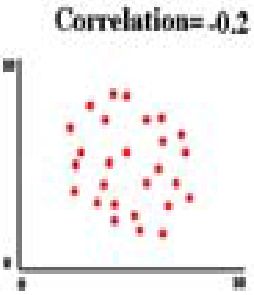












- The value of r is a measure of the extent to which x and y are **linearly related**

  i.e. the extent to which the points in the scatter plot lie close to a straight line.

- A value close to zero does not rule out any strong relationship between

  x and y; there could still be a strong relationship but one that is not linear.

$$\boldsymbol{r = \frac{Cov(x,y)}{S_x S_y}} \text{ where}$$

$$\text{Cov (x,y)} = \sum \frac{xy}{n} - \bar{x}\bar{y}\,,\; Sx = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2}\,,\; and\; Sy = \sqrt{\frac{\sum y_i^2}{n} - \bar{y}^2}$$

OR

$$\boldsymbol{r = \frac{\sum \frac{xy}{n} - \bar{x}\bar{y}}{\sqrt{\left(\frac{\sum x_i^2}{n} - \bar{x}^2\right)\left(\frac{\sum y_i^2}{n} - \bar{y}^2\right)}}}$$

Examples:

- Minimum daily temperature and heating costs
- Interest rate and number of loan applications
- Incomes of husbands and wives when both have full-time jobs
- Ages of boyfriends and girlfriends