Correlation analysis

❖ Two Quantitative Variables

The response variable, also called the dependent variable, is the variable we want to predict, and is usually denoted by y.

The explanatory variable, also called the independent variable, is the variable that attempts to explain the response, and is denoted by x.

Example:

Dependent Variable	Independent Variable
Weight of son	Height of son
Expenditure	Income

❖ Scatter Diagram

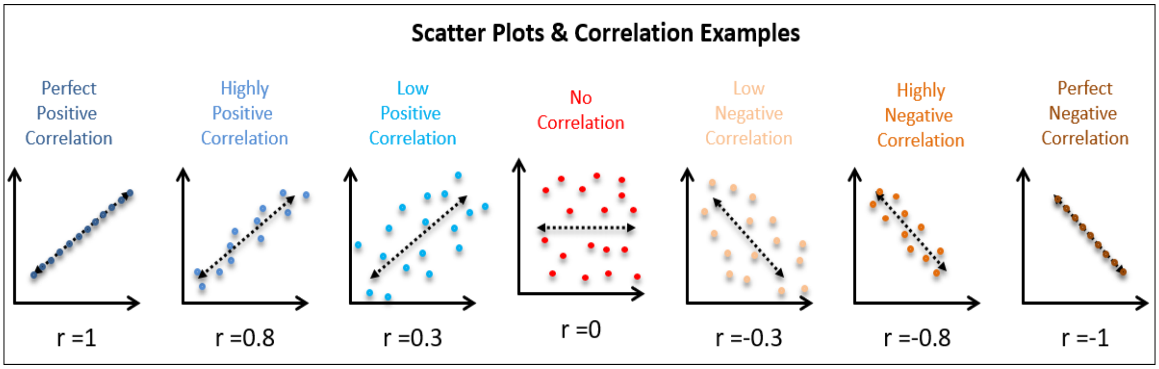
It is the simplest way of the diagrammatic representation of bivariate data. Thus for the' bivariate distribution (X_i, Y_i); $i = 1,2, \ldots, n$, if the values of the variables X and Y be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can get an idea whether the variables are correlated or not, e.g. if the points are very dense, i.e. very close to each other, we should expect a good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is large.

❖ Correlation Analysis

Correlation analysis is used to measure strength of the association or linear relationship between two variables.

- Only concerned with strength of the relationship
- No causal effect is implied.

Example: Relationship between height and weight, income and expenditure etc.



❖ Correlation Coefficient

The correlation coefficient measures the strength of the association between the variables. It is usually denoted by r or ρ (rho).

Let $(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots\cdots\cdots.(x_n, y_n)$ be n pairs of observations of variable x and y. $\bar{\mathrm{x}}$ and $\bar{\mathrm{y}}$ be the mean of x and y respectively. The correlation coefficient between x and y is

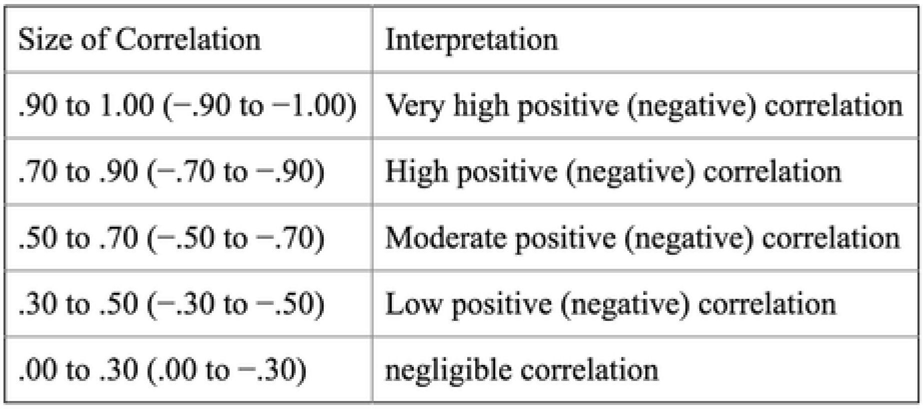
$$r_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\{\sum x_i{}^2 - \frac{(\sum x_i)^2}{n}\}\{\sum y_i{}^2 - \frac{(\sum y_i)^2}{n}\}}}$$

❖ Interpretation of Correlation Coefficient

In statistics, the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. The value of r is always between +1 and –1. To interpret its value, see the following table:



Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

❖ Properties of Correlation Coefficient

Some properties of correlation coefficient are as follows:

- Correlation coefficient measures the linear relationship between two variables.
- Correlation coefficient is a symmetric measure i.e. $r_{xy} = r_{yx}$.
- Correlation coefficient is a pure number.
- Correlation coefficient lies between -1 and $+1$.
- Correlation coefficient is independent on change of origin and scale of measurement.

- $r = 0$ indicates that there is no relationship between the variables. $r = +1$ and $r = -1$ indicates the perfect positive and perfect negative relationship between the variables.

Problem #1: The data given below are the amount of export (X in 00 million taka) of fish and amount of total export (Y in 00 million taka) in different years.

X	7	10	13	12	14	16	15
Y	17	22	24	25	28	37	40

Calculate correlation coefficient of X and Y. Also, comment on your result.

Solution:

X	Y	X^2	Y^2	XY
7	17	49	289	119
10	22	100	484	220
13	24	169	576	312
12	25	144	625	300
14	28	196	784	392
16	37	256	1369	592
15	40	225	1600	600
$\sum X$ =87	$\sum Y$ =193	$\sum X^2$ =1139	$\sum Y^2$ =5727	$\sum XY$ =2535

We know that the coefficient of correlation between X and Y is

$$r = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sqrt{\left\{\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right\}\left\{\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right\}}}$$

$$= \frac{2535 - \frac{87 \times 193}{7}}{\sqrt{\left\{1139 - \frac{(87)^2}{7}\right\}\left\{5727 - \frac{(193)^2}{7}\right\}}}$$

$$= \frac{136.286}{\sqrt{57.714 \times 405.714}}$$

$$= \frac{136.286}{153.021} = 0.891$$

Comment: Since $r = 0.891$, therefore we can say that there is strong positive correlation between amount of export (X in 00 million taka) of fish and amount of total export (Y in 00 million taka).

- ❖ **Theorem #1:** Show that correlation coefficient lies between -1 and $+1$.

  Or show that $-1 \le r \le +1$.

Solution:

Let $(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots\cdots\cdots.(x_n, y_n)$ be n pairs of observations of variable x and y. $\bar{x}$ and $\bar{y}$ be the mean of x and y respectively.

The correlation coefficient between x and y is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Let

$$u_i = \frac{(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ and } v_i = \frac{(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

We know that

$$\sum_{i=1}^{n}(u_i \pm v_i)^2 \ge 0$$

$$=> \sum_{i=1}^{n} u_i^2 + \sum_{i=1}^{n} v_i^2 \pm 2\sum_{i=1}^{n} u_i v_i \ge 0$$

$$=> \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \pm 2\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \ge 0$$

$$=> 1 + 1 \pm 2r \ge 0$$

$$=> 1 \pm r \ge 0$$

$1 + r \ge 0$ Or $1 - r \ge 0$

$\therefore r \ge -1$ $\therefore r \le +1$

Therefore, we can say that $-1 \le r \le +1$. (Showed)

- ❖ **Theorem #2:** Show that correlation coefficient is independent on change of origin and scale of measurement.

Solution:

Let $(x_1, y_1), (x_2, y_2), (x_3, y_3), \cdots\cdots\cdots.(x_n, y_n)$ be n pairs of observations of variable x and y. $\bar{x}$ and $\bar{y}$ be the mean of x and y respectively.

The correlation coefficient between x and y is

$$\mathrm{r_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Let

$$u_i = \frac{x_i - a}{h} \qquad \text{and} \qquad v_i = \frac{y_i - b}{k}$$

Where, $a, b, h\ and\ k$ are arbitrary constants. $a\ and\ b$ are origin, and $h\ and\ k$ are scale.

$$=> x_i = a + hu_i \qquad\qquad => y_i = b + kv_i$$

$$\therefore \bar{x} = a + h\bar{u} \qquad\qquad \therefore \bar{y} = b + k\bar{v}$$

We have

$$\mathrm{r_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(a + hu_i - a - h\bar{u})(b + kv_i - b - k\bar{v})}{\sqrt{\sum_{i=1}^{n}(a + hu_i - a - h\bar{u})^2 \sum_{i=1}^{n}(b + kv_i - b - k\bar{v})^2}}$$

$$= \frac{hk \sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{hk\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

$$= \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}} = \mathrm{r_{uv}}$$

Therefore, we can show that correlation coefficient is independent on change of origin and scale of measurement. (Showed)