Correlation

F. M. Arifur Rahman

Senior Lecturer, Department of Mathematical & Physical Sciences

Content

- Correlation
  - Simple correlation (Pearson correlation)

Correlation

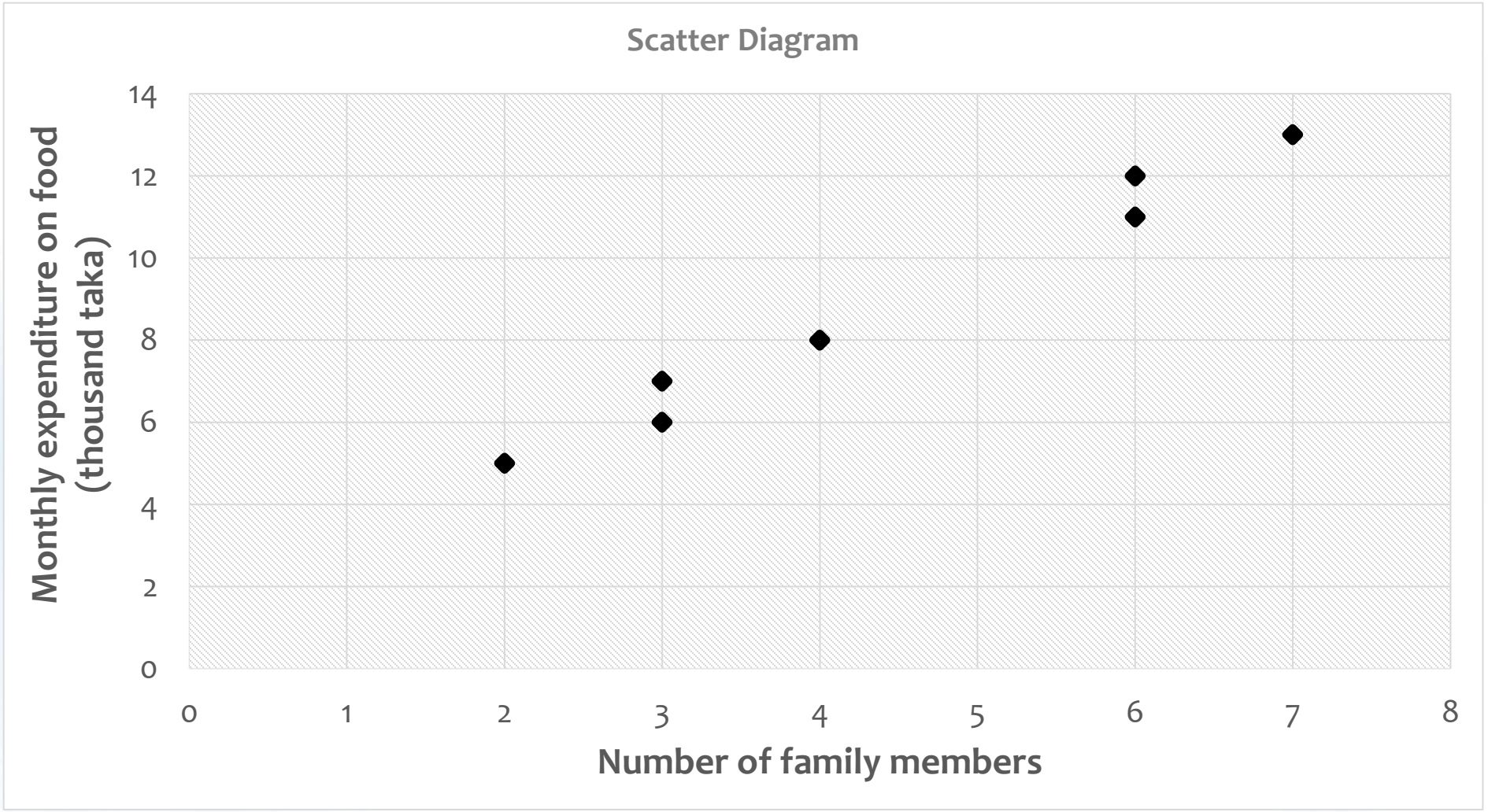
- Relationship between two or more variables
- When variables are found to be related, we often want to know how close the relationship is. This type of analysis is known as correlation analysis
- The primary objective of correlation analysis is to measure-
  - Degree or strength of relationships
  - Direction of relationship
- Correlation does not necessarily mean causation

Correlation

- Example:

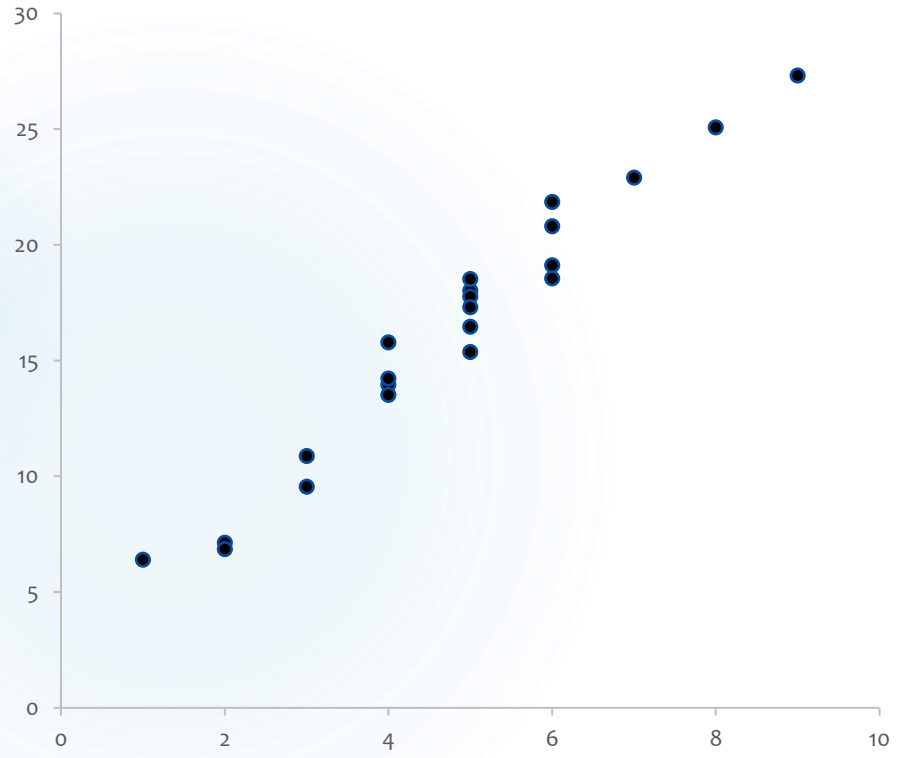
No. of family members, X	Monthly expenditure on food (thousand taka), Y
2	5
3	7
6	11
4	8
7	13
3	6
6	12

Correlation

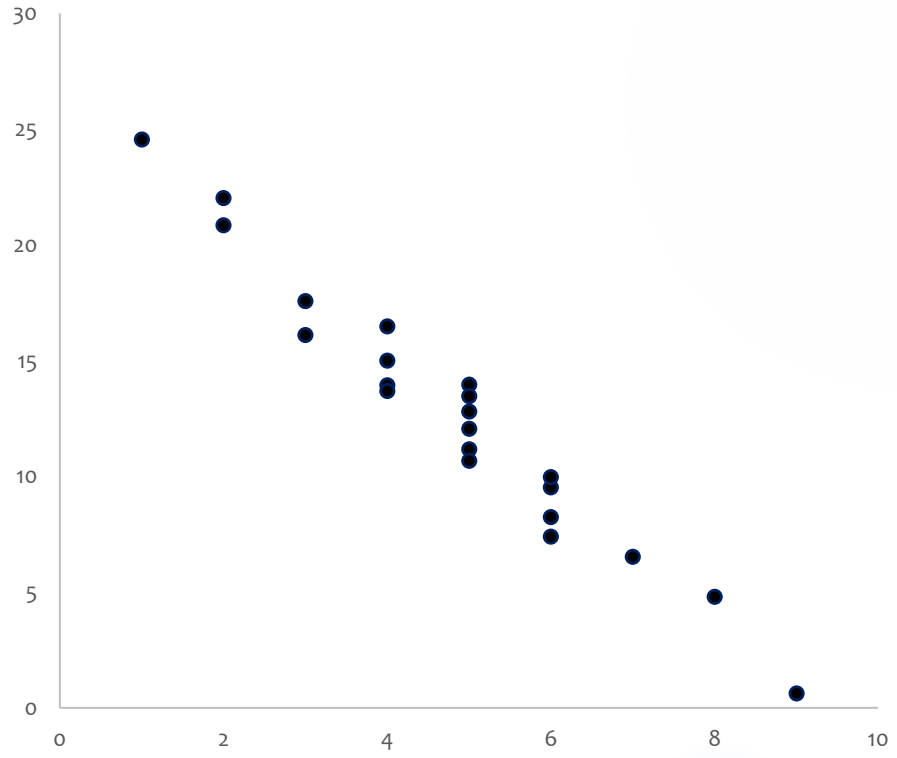


Correlation

Positive correlation

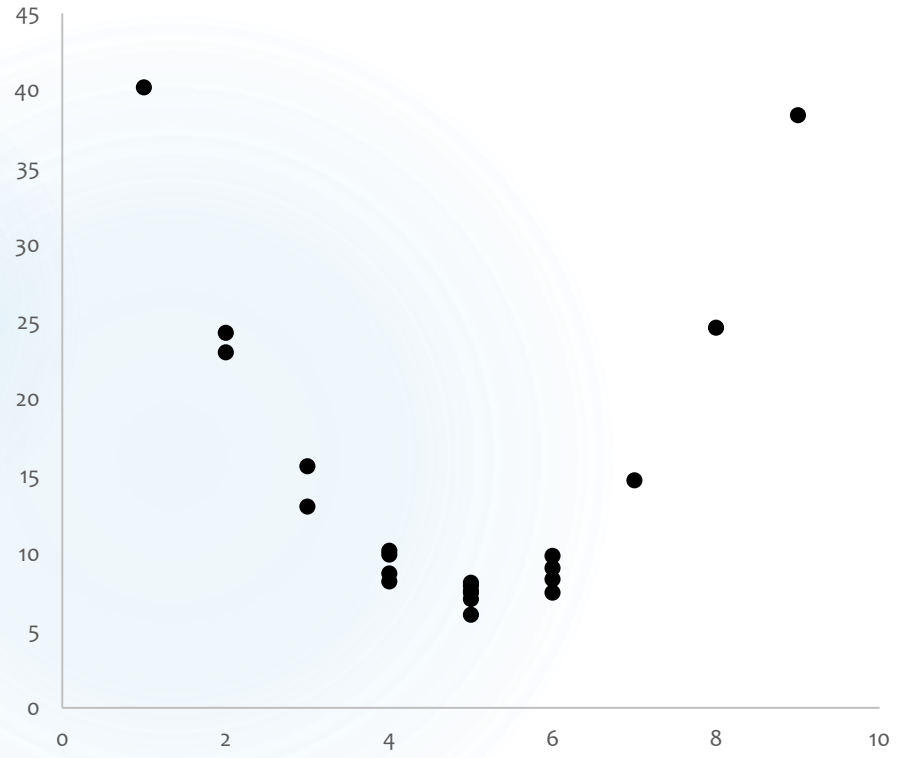


Negative correlation

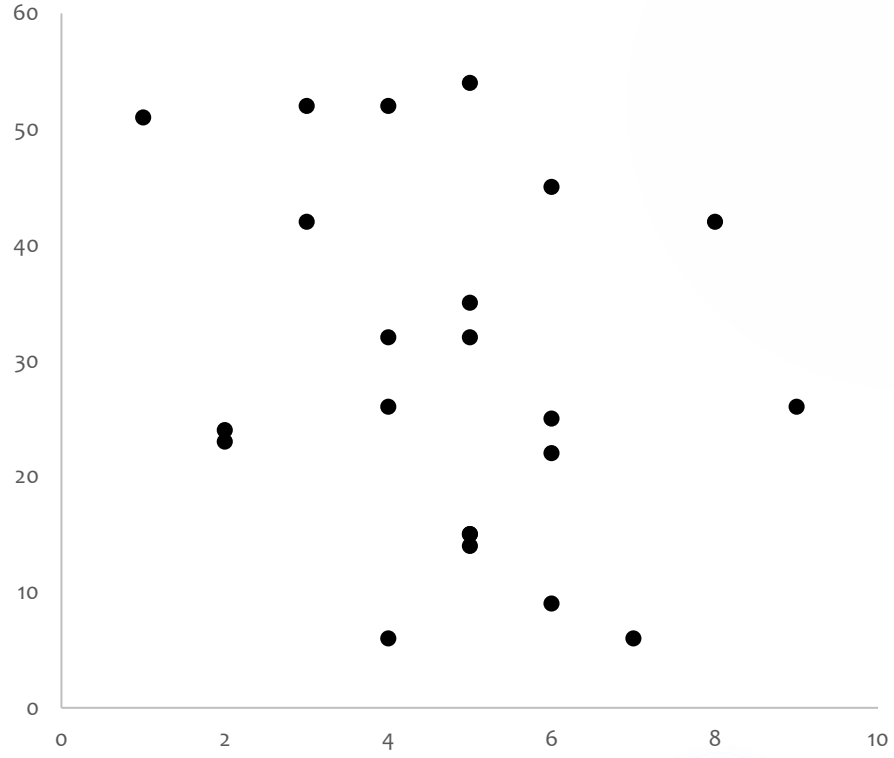


Correlation

Non-linear correlation



No correlation



Simple correlation

- Pearson correlation coefficient-

$$r = \frac{cov(X,Y)}{\sqrt{v(X)v(Y)}}$$

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

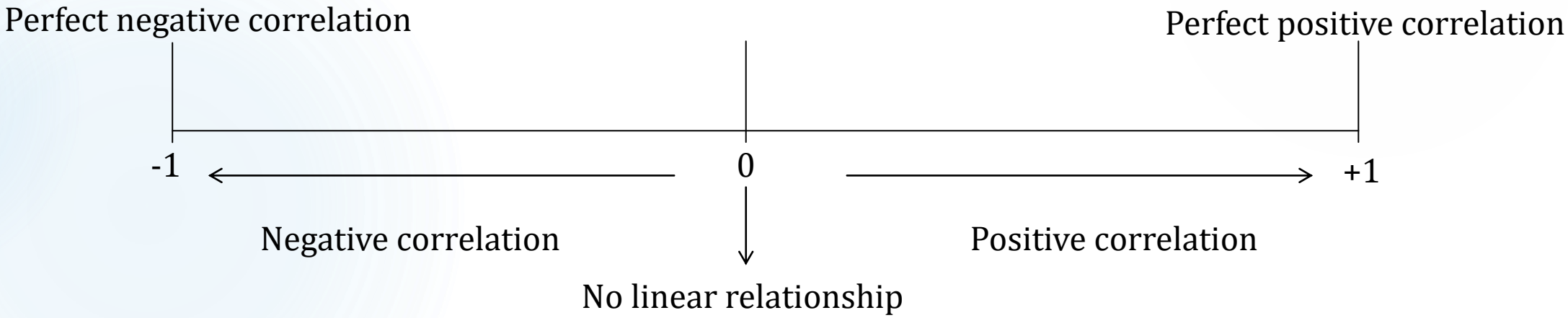
$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

- X & Y are two numerical variables and n is the number of pairs.

Simple correlation

Interpretation

- r>0: Positive linear relationship
- r<0: Negative linear relationship
- r=0: No linear relationship



Simple correlation

Correlation Coefficient	**-1**	**(-.99)-(-.51)**	**-.5**	**(-.49)-(-.01)**	**0**	**.01-.49**	**.5**	**.51-.99**	**1**
Correlation type	Perfect negative	Strong negative	Moderate negative	Weak negative	No correlation	Weak positive	Moderate positive	Strong positive	Perfect positive

Simple correlation

Assumptions:

- Both X & Y are measured on an interval or ratio scales
- The two variables follow bi-variate normal distribution
- The relationship between the variables is linear
- The sample is of adequate size to assume normality

Simple correlation

Example (continues)

No. of family members, x	Monthly expenditure on food (thousand taka), y	x^2	y^2	xy
2	5			
3	7			
6	11			
4	8			
7	13			
3	6			
6	12			

Simple correlation

Example (continues)

No. of family members, x	Monthly expenditure on food (thousand taka), y	x^2	y^2	xy
2	5	4	25	10
3	7	9	49	21
6	11	36	121	66
4	8	16	64	32
7	13	49	169	91
3	6	9	36	18
6	12	36	144	72
$\sum x = 31$	$\sum y = 62$	$\sum x^2 = 159$	$\sum y^2 = 608$	$\sum xy = 310$

Simple correlation

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$
$$= \frac{7 * 310 - 31 * 62}{\sqrt{[7 * 159 - (31)^2]\,[7 * 608 - (62)^2]}}$$
$$= 0.991$$

Interpretation: So, there is a very strong positive relationship between number of family members and monthly expenditure. That is, both increase or decrease in the same direction.

Simple correlation

Properties:

- ***r*** always measures linear relationships
- ***r***=0 doesn't necessarily mean that X & Y are not related, but that they are not linearly related.
- $r_{xy} = r_{yx}$, i.e. correlation coefficient is a symmetrical measure
- The correlation coefficient is a dimensionless measure, implying that it is not expressed in any units of measurement
- Correlation doesn't mean causation, i.e. correlation doesn't necessarily imply any cause and effect relationship