Correlation

Md. Mahfuzur Rahman

Senior Lecturer, Statistics

Content

- Correlation
  - Simple correlation (Pearson correlation)
  - Rank correlation (Spearman correlation)

Correlation

- In many applications, we may want to study the underlying nature of relationships among the variables. Furthermore, we may also want to utilize these relationships for predicting or estimating the values for some variables on the basis of the given values for the other variables. By exploring the underlying relationships, we can explore very important findings and can provide necessary inputs required for useful decisions. Some examples of these relationships are:

(i) relationship between height and weight,

(ii) relationship between weight and cholesterol level,

(iii) relationship between income and expenditure, etc.

Correlation

- In this type of studies, we are interested in answering several important questions, some of which are:
- (i) **Is there a relationship between the variables? What is the nature of this relationship? What is the strength of this relationship?**
- (ii) **If there is a relationship between the variables, how can we formulate it mathematically? How can we utilize it?**

Correlation

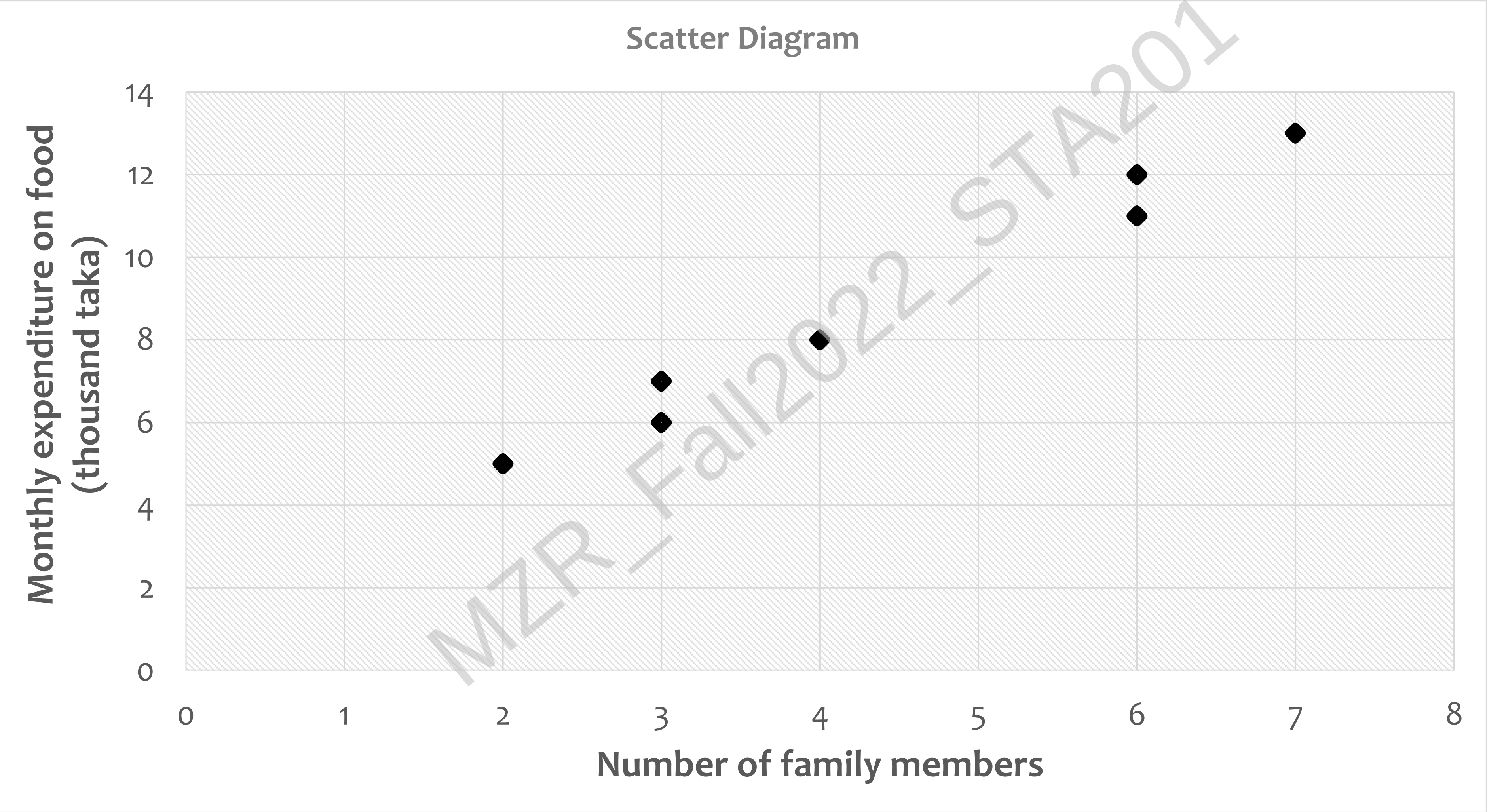
- Relationship between two or more variables
- When variables are found to be related, we often want to know how close the relationship is. This type of analysis is known as correlation analysis
- The primary objective of correlation analysis is to measure-
  - Degree or strength of relationships
  -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

Correlation

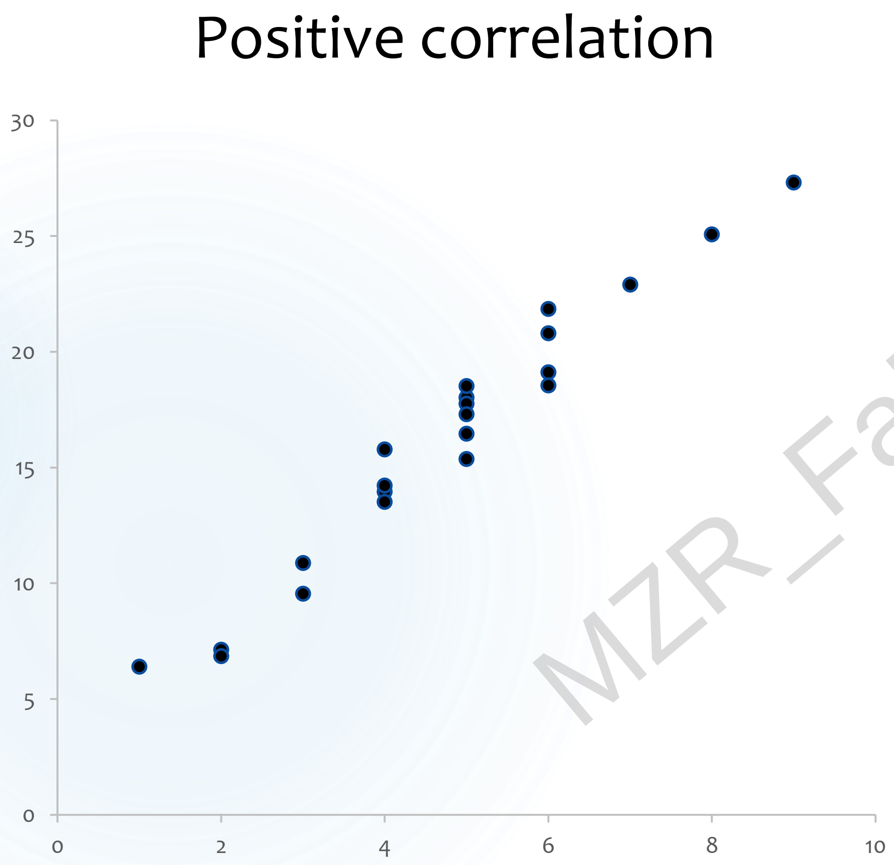
- Example:

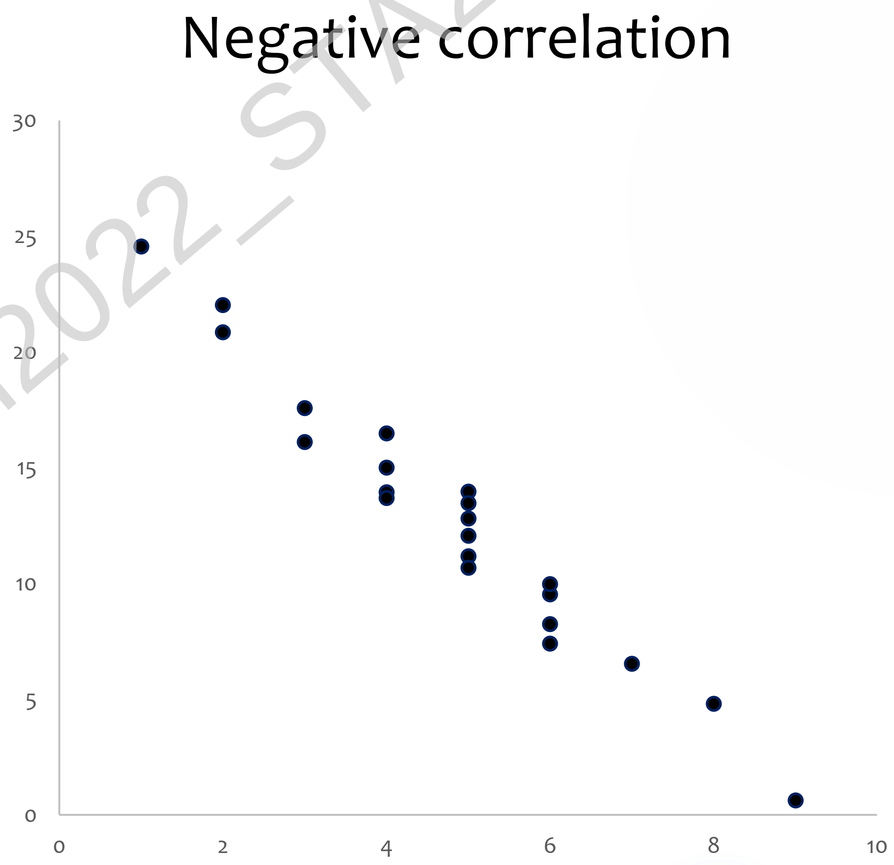
No. of family members, X	Monthly expenditure on food (thousand taka), Y
2	5
3	7
6	11
4	8
7	13
3	6
6	12

Correlation

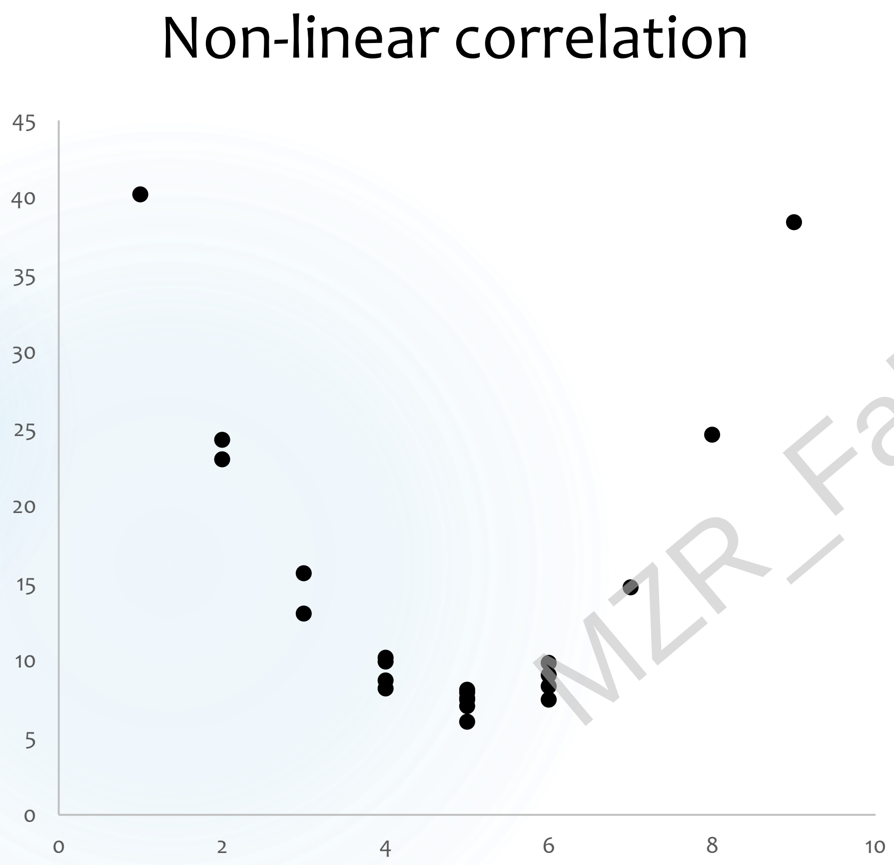


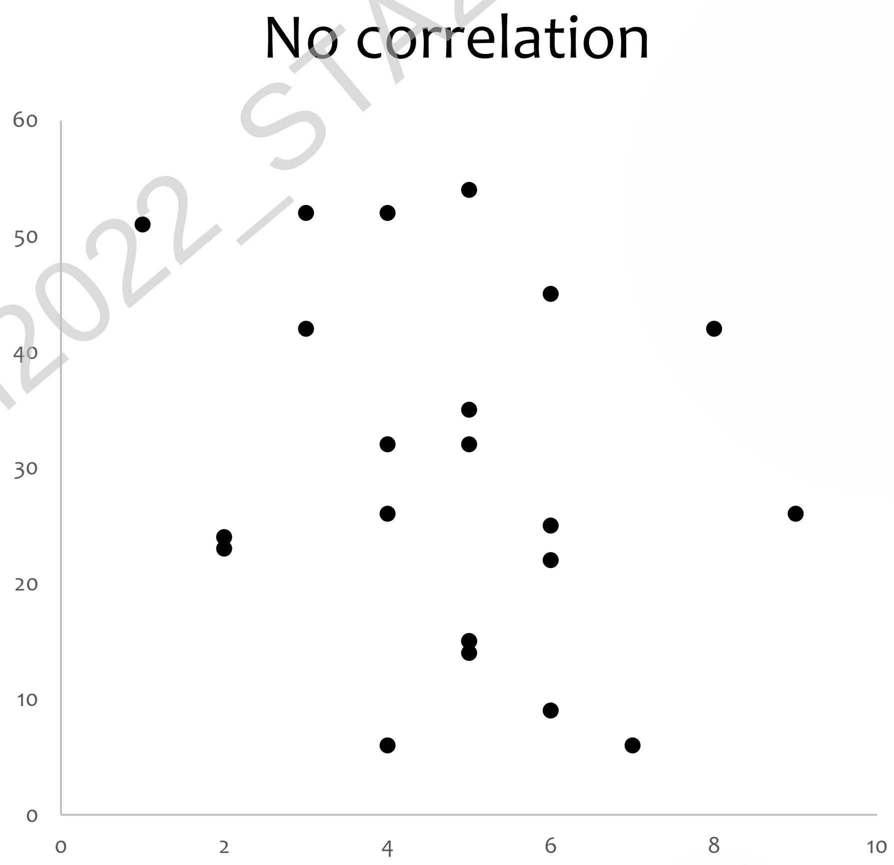
Correlation





Correlation





Simple correlation

- Pearson correlation coefficient-

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

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

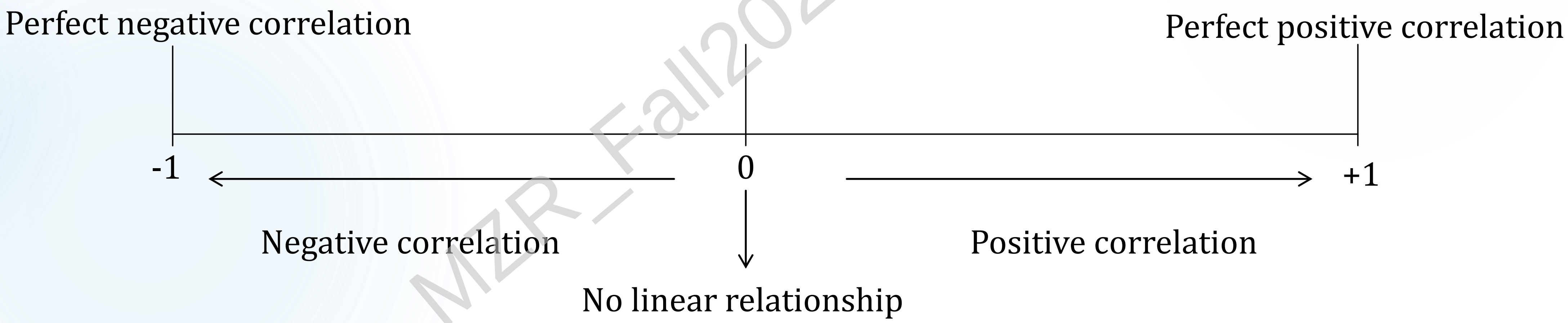
$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left[n\sum x_i^2 - (\sum x_i)^2\right]\left[n\sum y_i^2 - (\sum y_i)^2\right]}} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

- 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.
- $\boldsymbol{r_{xy}} = \boldsymbol{r_{yx}}$, 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

Simple correlation

Example:

An executive manager of a private hospital was interested in studying the relationship between the monthly number of part-time physicians (X) hired in the hospital and the monthly extra profit earned by the hospital in thousands (Y). For this purpose, the manger selected a random sample of ten months and obtained the following data:

i	X_i	Y_i
1	43	175
2	49	180
3	50	186
4	12	95
5	8	75

i	X_i	Y_i
6	32	165
7	51	190
8	30	95
9	35	130
10	23	95

(1) Draw the scatter diagram. What indications does the scatter diagram reveal?

(2) Calculate Pearson's Correlation Coefficient (r).

Rank correlation

Spearman rank correlation (Spearman's rho) $\boldsymbol{r_s}$ is the sample correlation coefficient $\boldsymbol{r}$ applied to the **rank order** data.

Rank correlation

Formula:

$$r_s = \frac{\sum x_i y_i - C}{\sqrt{\left(\sum x_i^2 - C\right)\left(\sum y_i^2 - C\right)}}$$

Where, $C = \frac{n\,(n+1)^2}{4}$

And n is the number of pairs.

Rank correlation

Example

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

Rank correlation

Example

No. of family members, x	Monthly expenditure on food (thousand taka), y	Rank of x, a	Rank of y, b	a^2	b^2	ab
2	5	1	1	1	1	1
3	7	2.5	3	6.25	9	7.5
6	11	5.5	5	30.25	25	27.5
4	8	4	4	16	16	16
7	13	7	7	49	49	49
3	6	2.5	2	6.25	4	5
6	12	5.5	6	30.25	36	33
Total	-	-	-	**139**	**140**	**139**

Rank correlation

$$C = \frac{n\,(n+1)^2}{4} = \frac{7\,(7+1)^2}{4} = 112$$

$$r_s = \frac{\sum a_i b_i - C}{\sqrt{\left(\sum a_i^2 - C\right)\left(\sum b_i^2 - C\right)}}$$

$$= \frac{139 - 112}{\sqrt{(139 - 112)(140 - 112\,)}} = 0.982$$

Interpretation: So, there is a very strong positive relationship between number of family members and monthly expenditure.

Rank correlation

Properties:

- Spearman correlation coefficient ranges from -1 to 1 with similar interpretation to that for the simple correlation coefficient r.
- $\boldsymbol{r_s}$ is a measure of monotonicity of a relationship
- Again, correlation doesn't mean causation, i.e. correlation doesn't necessarily imply any cause and effect relationship

Advantages and disadvantages

Advantage of r over r_s:

- ***r*** provides a more accurate result than r_s, when applicable, as ***r*** uses more information than r_s

Advantage of r_s over r:

- r_s is less affected by extreme observations
- r_s can be calculated for curvilinear (non-linear) relationship
- r_s can be calculated for ordinal level of data.