Correlation and Regression Analysis

Part 1: Correlation

Meaning of Correlation Analysis

Correlation Analysis: is a statistical technique used to measure the ***degree*** and ***direction of relation-ship*** between the variables.

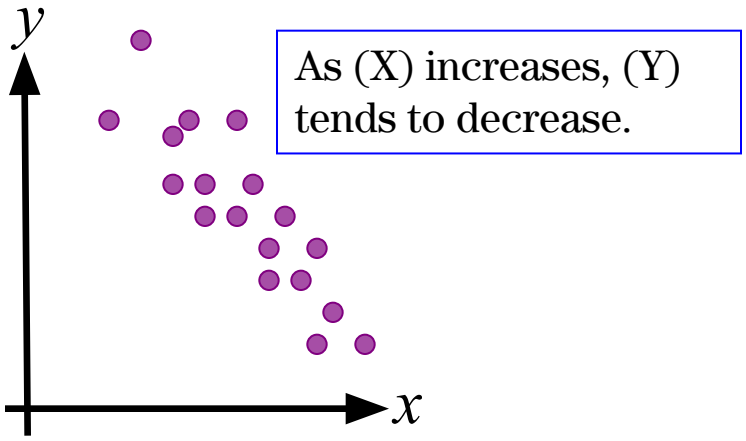
$$-1 \leq r \leq 1$$

Methods of studying Correlation

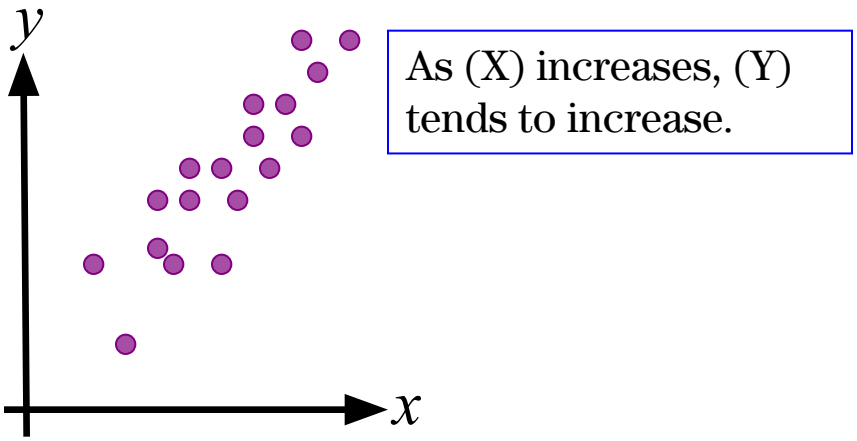
1- Scatter diagram Method.

2- Karl Person's coefficient of Correlation Method.

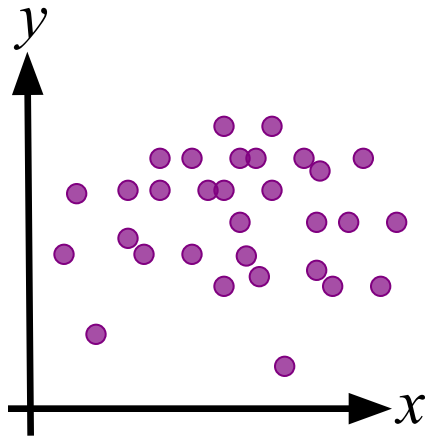
3- Spearman's Rank Correlation Method.



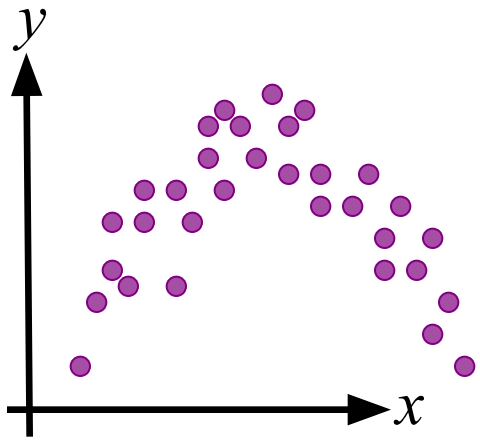
Negative Linear Correlation



Positive Linear Correlation



No Correlation



Nonlinear Correlation

Example:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student:

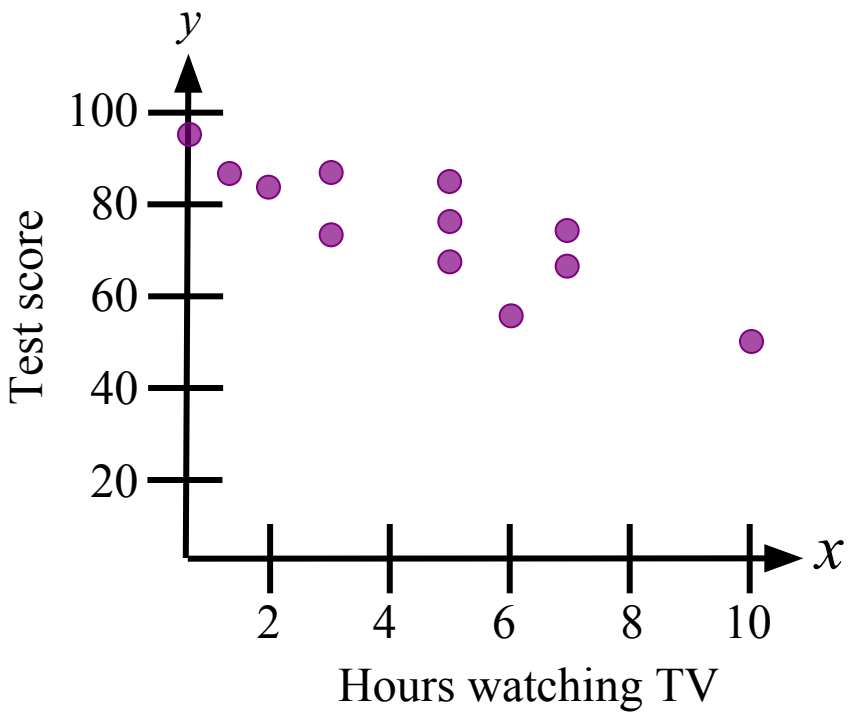
a.) Display the scatter plot.

b.) Calculate the correlation coefficient r.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Continued.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50



There is a strong negative linear correlation.

Correlation Coefficient

Example continued:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \approx -0.831$$

There is a strong negative linear correlation.

Regression Analysis

- Regression Analysis is a statistical tool used to study the **functional relationship** between two or more variables.

$$Y = f(x)$$

Linear model: $Y = a + b\,X$

a is the intercept (the value of variable Y when value of variable $X = 0$).

b is the slope of regression line (the amount of change in the value of Variable Y per unit change in variable X).

Example: Find the equation of the regression line from the following data:

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

$$b = \frac{\sum X_i Y_i - n\,\overline{X}\,\overline{Y}}{\sum X_i^2 - n(\overline{X})^2} = -4.06742$$

$$a = \overline{Y} - b\overline{X} = 93.97$$

$$Y = 93.97 - 4.06742\ X$$

Linear model: Y = a + b X

Parameter	Estimate	Standard Error	T-Statistic	Sig
Intercept (a)	93.97	4.52359	20.7733	0.0000
Slope (b)	-4.06742	0.860012	-4.72949	0.0008

$$Y = 93.97 - 4.06742\ X$$

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Sig
Model	1472.4	1	1472.4	22.37	0.0008
Residual	658.262	10	65.8262		
Total (Corr.)	2130.67	11			

Correlation Coefficient (r) = -0.831296

R-squared (r^2) = 69.1053 percent