In real life situations, especially in social sciences and in business, we often want to know whether two or more variables are related, and if so, how they are related. In this chapter, we discuss relationships between two quantitative/numeric variables.

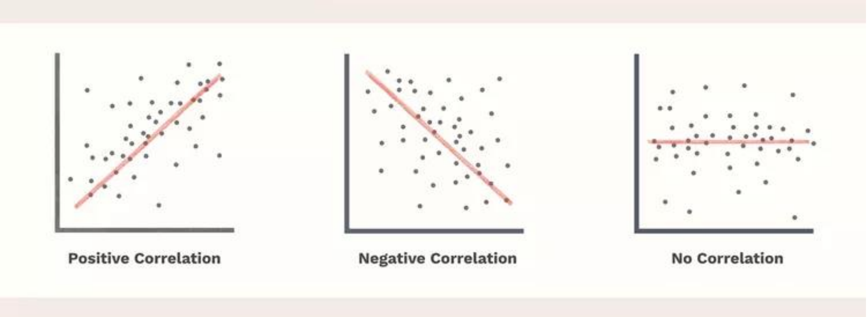
1. Correlation: The study of the statistical relationship between two or more variables which gives us the strength or degree and direction of association or interrelationship is known correlation, and the analysis is termed as correlation analysis.

We can classify correlation based on three categories:

a) ***On the basis of direction:*** Two types-

i. **Positive correlation:** A positive correlation signifies a relationship between two variables where they move together, meaning they change in the same direction. This occurs when one variable decreases as the other decreases, or one variable increases as the other increases.

ii. **Negative correlation:** A negative correlation indicates a relationship between two variables where they move in opposite directions. In other words, when one variable increases, the other variable decreases, and vice versa.



b) ***On the basis of variables:*** Three types-

i. **Simple correlation:** Simple correlation measures the strength and direction of the linear relationship between two variables. It quantifies how changes in one variable are associated with changes in another variable. For example: relationship between demand and supply.

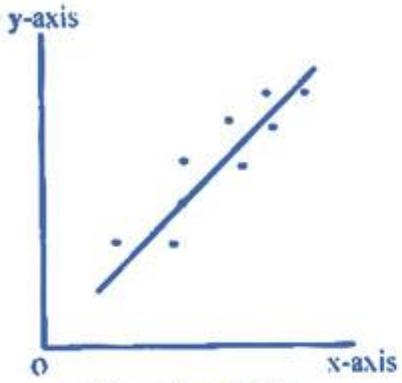
ii. **Multiple correlation:** Multiple correlation measures the strength and direction of the linear relationship among more than two variables. For example, consider the relationship among rainfall, rice production, and rice prices.

iii. **Partial correlation:** Partial correlation measures the strength and direction of the linear relationship among more than three variables where only two influencing variables are studied and rest influencing variables are constant. For example: Consider the relationship among demand, supply, and income, assuming income to be constant.

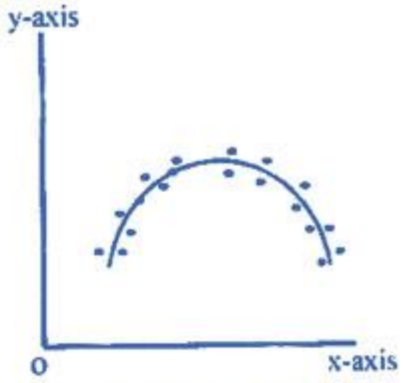
c) ***On the basis of linearity:*** Two types-

i. **Linear correlation:** Linear correlation refers to the statistical relationship between two variables that can be represented by a straight line. When changes in one variable consistently correspond with changes in another variable, and these changes can be plotted as a straight line on a graph, it signifies a linear correlation. In other words, if one variable increase and the other also increases at a constant rate, they are linearly correlated.

ii. **Nonlinear correlation:** Nonlinear correlation, on the other hand, occurs when the relationship between two variables is not accurately represented by a straight line. In this case, changes in one variable do not result in proportional changes in the other variable. Instead, the relationship may follow a curve or another non-straight pattern on a graph, indicating a nonlinear correlation. Nonlinear correlations can take various forms, such as quadratic, exponential, or logarithmic relationships, among others. These relationships cannot be adequately described using a single straight line.



Linear Correlation



Non Linear Correlation

2. Methods of studying correlation: Correlation can be measured by the following methods:

a) Scatter diagram

b) Karl Pearson's correlation coefficient

c) Spearman's rank correlation coefficient

d) Method of least squares

3. Scatter diagram: Below table displays data of age and corresponding glucose level of five respondent.

Age	43	21	25	42	57
Glucose Level	99	65	79	75	87

a) Draw the scatter diagram

b) Is there any relationship between "Age" and "Glucose Level"?

Solution: First, we draw the scatter plot,

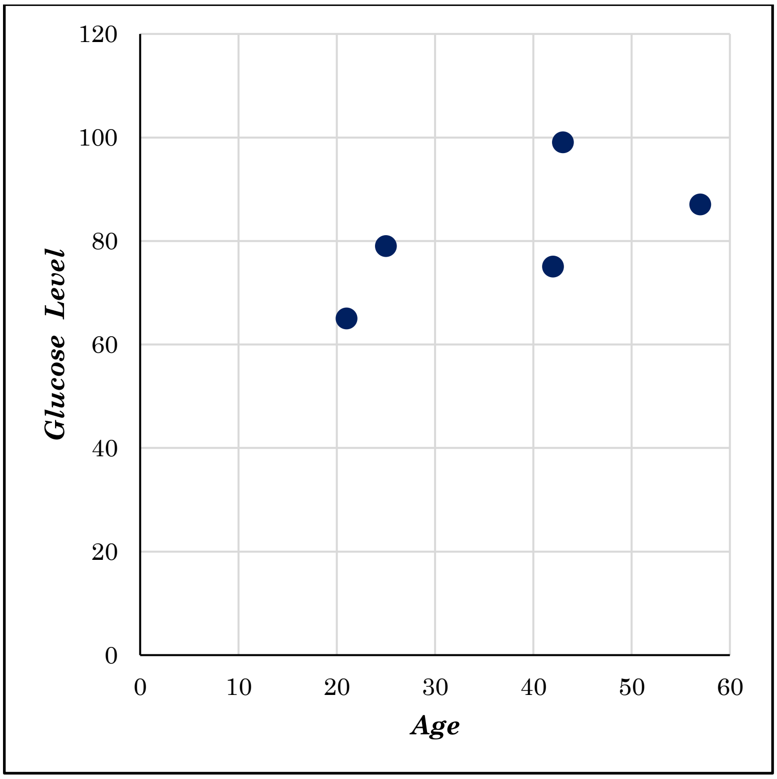


Figure: Scatter diagram for Age-GL data.

Interpretation: With the increase of age, the Glucose Level also increased. Thus, there is a positive correlation between "Age" and "Glucose Level".

4. Limitations of Scatter diagram: Accurate degree and strength of correlation cannot be obtained by scatter diagram.

5. Karl Pearson's correlation coefficient: Karl Pearson's correlation coefficient is computed based on the following assumptions:

- ***a)*** Both variables are measured on an interval or ratio scales.
- ***b)*** The two variables exhibit linear relationship.
- ***c)*** The values on the two variables come from bivariate normal population.
- ***d)*** The sample is of adequate size to assume normality.

Properties:

- a) Correlation coefficient has no unit.
- b) The sign of correlation coefficient gives the direction of the association.
- c) Range of correlation coefficient is between -1 and +1, i.e., $-1 \leq r \leq +1$
- d) If X and Y are independent, the correlation coefficient is zero.

Let, X and Y be two continuous random variables. The relationship between X and Y can be written as,

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

The range of Karl Pearson's correlation coefficient, r between -1 to $+1$. If $r = -1$, we can say that there is perfectly negative correlation between X and Y. If $r = +1$, we can say that there is perfectly positive correlation between X and Y. If $r = 0$, we can say that there is no correlation between X and Y.

General guidelines:

Correlation	**Negative**	**Positive**
Weak	-0.29 to -0.10	0.10 to 0.29
Medium	-0.49 to -0.30	0.30 to 0.49
Moderate	-0.50 to -0.79	0.50 to 0.79
Strong	-1.00 to -0.80	0.80 to 1.00

Math: The monthly income and saving data for a sample of 10 garments workers are given below:

Income ($)	60	66	66	66	68	68	70	72	74	80
Savings ($)	5	7	8	9	11	12	14	16	21	27

a) Draw the scatter plot

b) Compute the correlation coefficient value with proper interpretation

Solution:

a)

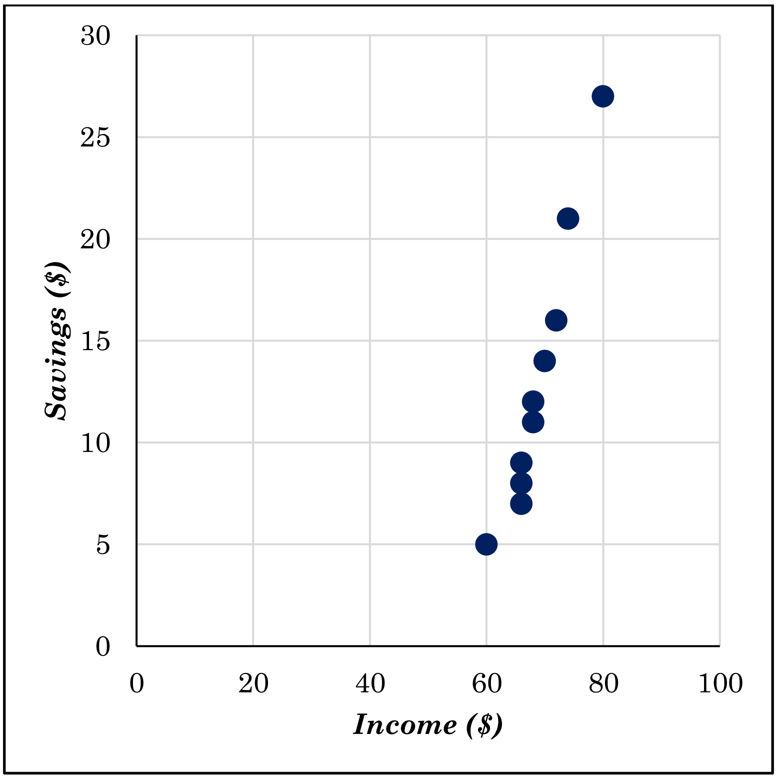


Figure: Scatter diagram of Income-Savings data.

b)

x (income)	y (Savings)	x_i^2	y_i^2	$x_i \times y_i$
60	5	3600	25	300
66	7	4356	49	462
66	8	4356	64	528
66	9	4356	81	594
68	11	4624	121	748

68	12	4624	144	816
70	14	4900	196	980
72	16	5184	256	1152
74	21	5476	441	1554
80	27	6400	729	2160
$\sum x_i = 690$	$\sum y_i = 130$	$\sum x_i^2 = 47876$	$\sum y_i^2 = 2106$	$\sum x_i y_i = 9294$

Now, the correlation coefficient

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}} = \frac{9294 - \frac{690 \times 130}{10}}{\sqrt{47876 - \frac{(690)^2}{10}}\sqrt{2106 - \frac{(130)^2}{10}}} = 0.97$$

The values of $r = 0.97$, suggests a strong positive correlation between income and savings of garments workers. That is, as income increases, there is a strong tendency for saving increase.

Extra:

1. A new type energy bulb was recently introduced in 10 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and recorded the number of units sold in one week for each of the different prices as follows:

Price	250	255	260	272	280	285	290	296	300	304
Number sold	57	41	33	42	30	37	30	28	34	20

a) Plot the data in a scatter diagram and interpret.

b) Compute correlation coefficient and interpret.

2. Compute coefficient of correlation for the following data and comment on the result:

Income	50	720	202	130	140	182
Expenditure	32	378	105	66	84	72

3. Compute coefficient of correlation between x and y and comment.

$\boldsymbol{x}$	11	14	15	17	18	21	23	25
$\boldsymbol{y}$	150	270	270	300	340	360	400	420

4. Mr. Johnson is concerned about the cost to students of textbooks. He believes there is a relationship between the number of pages in the text and the selling price of the book. To provide insight into the problem he selects a sample of eight textbooks currently on sale in the bookstore. **Compute the correlation coefficient.**

Pages	500	700	800	600
Prices ($)	$80 + x$	$70 + x$	$94 + x$	$67 + x$

$\boldsymbol{x = Last\ digit\ of\ your\ ID}$