Department of AI & DS
CSE and CS&IT

COURSE NAME: PROBABILITY, STATISTICS AND QUEUING THEORY

COURSE CODE: 23MT2005

Topic

Correlation analysis

Session - 12

AIM OF THE SESSION

To familiarize To familiarize students with the basic concept of correlation and different methods of studying correlation.

INSTRUCTIONAL OBJECTIVES

This Session is designed to:

1. Demonstrate the correlation and its types with suitable examples
2. Describe different methods of studying correlation
3. List out the properties of correlation
4. Describe the Karlpearson's correlation coefficient.

LEARNING OUTCOMES

At the end of this session, you should be able to:

1. Define correlation and its types.
2. Describe different methods of studying correlation.
3. Summarize the concept of Karlpearson's correlation coefficient with appropriate conclusions.

SESSION INTRODUCTION

CONTENTS

- Definition
- Types of Correlation
- Different methods of Correlation

Correlation

Correlation analysis attempts to measure the strength of such r relationships between two variables by means of a single number called a **correlation coefficient**.

Types of Correlation

Strong Positive Correlation The value of Y clearly increases as the value of X increases.

Strong Negative Correlation The value of Y clearly decreases as the value of X increases.

Weak Positive Correlation The value of Y increases slightly as the value of X increases.

Weak Negative Correlation The value of Y decreases slightly as the value of X increases.

Complex Correlation The value of Y seems to be related to the value of X, but the relationship is not easily determined.

No Correlation there is no connection between the two variables.

Methods of studying correlation

1. Scatter Diagram Method
2. Karlpearson's Correlation coefficient
3. Spearman's Rank correlation

SCATTER DIAGRAM

Scatter diagram also called scatter plot, scatter diagram, dot diagram or scatter is one way to study the relationship between two variables. When the pair of values (X_i, Y_i) for i=1, 2, …, n are plotted on a graph paper, the points show the pattern in which they lie. Such a diagram is known as scatter diagram. If these points lie on a straight line, it is expected that there is a linear relationship between X and Y, otherwise not. It is a pictorial representation of the data. They tell us the direction of the relationship between two variables.

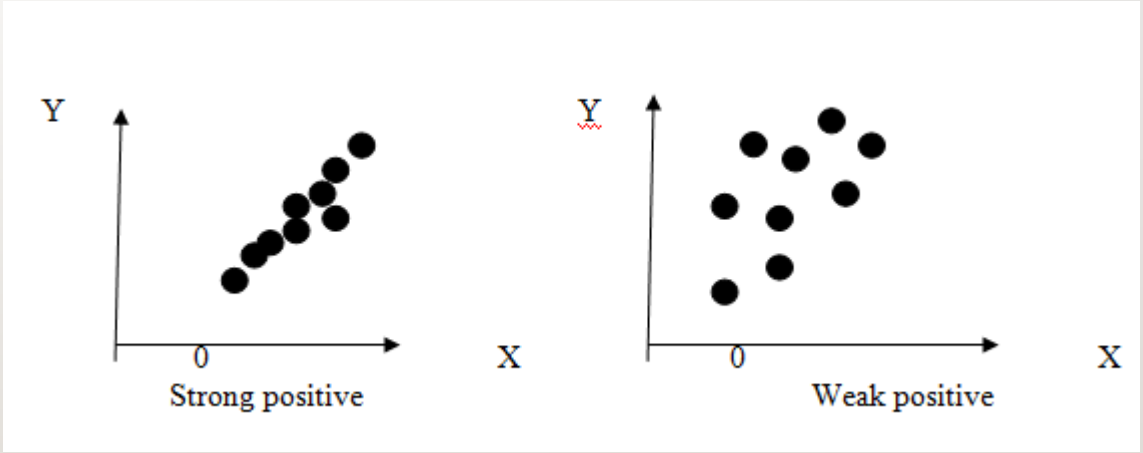
Scatter diagram method

Use a scatter diagram to examine theories about cause-and-effect relationships and to search for root causes of an identified problem. Use a scatter diagram to design a control system to ensure that gains from quality improvement efforts are maintained.

Positive correlation: If the increase of one variable affects the increase of another variable, i.e., two variables are in the same direction then it is said to be positive correlation.

If r=+1 indicates that perfect positive correlation.

If r>0 indicates that positive correlation.

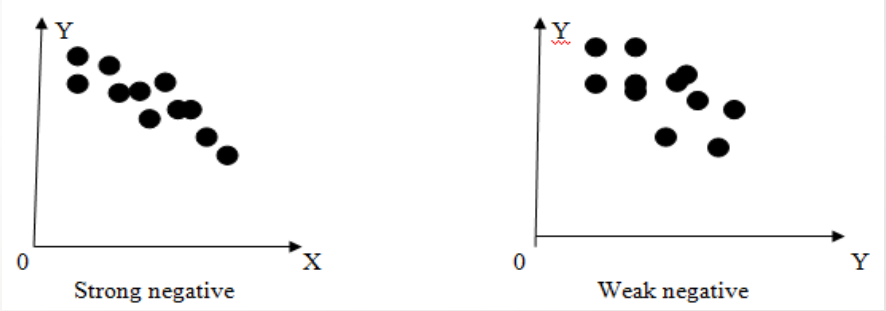


Scatter Diagram method

Negative correlation: If the increase of one variable affects the decrease of another variable, i.e., two variables are in the opposite direction then it is said to be negative correlation.

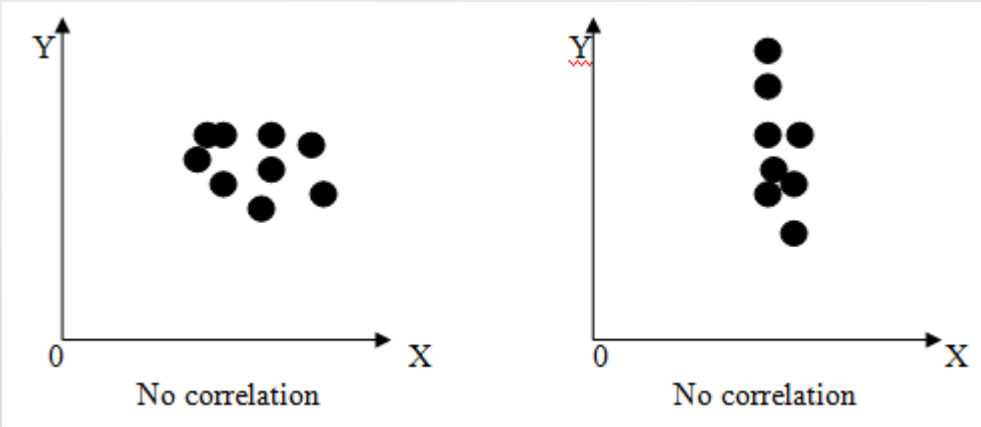
If r=-1 indicates that perfect negative correlation.

If r<0 indicates that negative correlation.



Zero correlation: There is no existence of any relationship between two variables is called zero correlation.

If r=0 indicates that zero correlation or no correlation between the two variables.



Scatter Diagram method

How to use it:

Collect data. Gather 50 to 100 paired samples of data that show a possible relationship.

Draw the diagram: Draw roughly equal horizontal and vertical axes of the diagram, creating a square plotting area. Label the axes in convenient multiples (1, 2, 5, etc.) increasing on the horizontal axes from left to right and on the vertical axis from bottom to top. Label both axes.

Plot the paired data. Plot the data on the chart, using concentric circles to indicate repeated data

Title and label the diagram.

Interpret the data. Scatter diagrams will generally show one of six possible correlations between the variables:

Karlpearson's correlation coefficient

In the regression we have assumed that the independent regressor variable x is a physical or scientific variable but not a random variable. But in the applications of regression technique, it is more realistic to assume that both X and Y are random variables and the measurements $\{(x_i, y_i); i=1,2,..,n\}$ are observations from a population having the density f(x,y).

For instance, if we studied the relationship between impurities in the air and incidence of a certain disease, input and output of wastewater treatment plant or the relationship between the textile strength and the hardness of aluminum between the textile strength and the hardness of aluminum.

Correlation coefficient for ungrouped data: The measure of linear association between two variables x and y is estimated by the sample correlation coefficient r, where

$$r = \frac{\text{Covariance}(x, y)}{\text{S. D}(X) * \text{S. D. }(y)}$$

Covariance (x,y) = $\frac{\sum xy}{N} - \left(\frac{\sum x}{N}\right)\left(\frac{\sum y}{N}\right)$, Standard deviation of x (S.D. of x) = $\sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$,

Standard deviation of y (S. D. of y) = $\sqrt{\frac{\sum y^2}{N} - \left(\frac{\sum y}{N}\right)^2}$

Karlpearson's correlation coefficient

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

r is called the Pearson product-moment correlation coefficient.

Properties of r

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

2. r=1 if all paris (x_i, y_i) lie exactly on a straight line having the slope. i.e, there is a perfect linear relationship with positive slope.

3. r>0 if the pattern in the scatter diagram runs from lower left to upper right.

4. r<0 if the pattern in the scatter diagram runs from upper left to lower right.

5. r=-1 if all pairs (x_i, y_i) lie exactly on a straight line having a negative slope, that is, perfect linear relationship with a negative

6. r=0 if there is no relationship.

7. A value of r near -1 or +1 describes strong linear relations

8. A value of r close to zero implies that the linear association is weak. There may still be a strong association along a curve.

EXAMPLES

A researcher wished to determine if a person's age is related to the number of hours he or she exercises per week. The data obtained from a sample is given. State your opinion based on Karl Pearson's coefficient of correlation for the data.

Age x: 18 26 32 38 52 59

Hours y: 10 5 2 3 1.5 1

Solution:

	Age x	Hours y	Xy	x^2	y^2
	18	10	180	324	100
	26	5	130	676	25
	32	2	64	1024	4
	38	3	114	1444	9
	52	1.5	78	2704	2.25
	59	1	59	3481	1
Total	225	22.5	625	9653	141.25

EXAMPLES

Here N=6

Mean of x=$\frac{\sum x}{N}$ =225/6=37.5

Mean of y=$\frac{\sum y}{N}$ =22.5/6=3.75

Standard deviation of x (S.D. of x) = $\sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$ =14.2332

Standard deviation of y (S. D. of y) = $\sqrt{\frac{\sum y^2}{N} - \left(\frac{\sum y}{N}\right)^2}$ =3.0788

Covariance (x,y) = $\frac{\sum xy}{N} - \left(\frac{\sum x}{N}\right)\left(\frac{\sum y}{N}\right)$ = -36.4583

Correlation (r) = $\frac{\text{covariance of (x,,y)}}{(\text{S.D.of x})(\text{S.D.of y})}$ = -0.8320

- There is negative relationship exists between the age and the hours of exercise of the persons. Based on the above data, we conclude that, if the age of person increases then the exercise hours decreases.

SUMMARY

1. Define correlation and its types.
2. Describe different methods of studying correlation.
3. Summarize the concept of Karlpearson's correlation coefficient with appropriate conclusions.

SELF-ASSESSMENT QUESTIONS

The range of simple correlation coefficient is

(a) 0 to ∞…

(b) -∞ to ∞

(c) 0 to 1

(d) -1 to 1

2. If ρ_{xy}=-1, the relation between x and y is of the type:

a) when Y increases, X also increases
b) When Y decreases, X also decreases
c) X is equal to –Y
d) when Y increases, X proportionately decreases

TERMINAL QUESTIONS

1. Describe different methods of studying correlation

2. List out the properties of correlation

3. Summarize the way of representing the relationship between two variables using Scatter diagram method.

4. A study of the amount of rainfall and the quantity of air pollution removed produced the following data

Daily rainfall, x (0.01cm)	**4.3**	**4.5**	**5.9**	**5.6**	**6.1**
Particulate removed, y (ug/m^3)	126	121	116	118	114

i) Make a scatter plot for the given data

ii) Determine the correlation coefficient for the given data.

REFERENCES FOR FURTHER LEARNING OF THE SESSION

Reference Books:

1. Chapter 1 of TP1: William Feller, An Introduction to Probability Theory and Its Applications: Volume 1, Third Edition, 1968 by John Wiley & Sons, Inc.

2. Richard A Johnson, Miller& Freund's Probability and statistics for Engineers, PHI, New Delhi, 11th Edition (2011).

Sites and Web links:

1. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
2. https://spoken-tutorial.org/watch/R/Introduction+to+basics+of+R/English/Methods of studying correlation
3. https://nptel.ac.in/courses/105105150/24

THANK YOU

Team – PSQT EVEN SEMESTER 2024-25