|  |  |  |
| --- | --- | --- |
| Related image | **KONERU LAKSHMAIAH EDUCATION FOUNDATION**  (Deemed to be University estd, u/s, 3 of the UGC Act, 1956) (NAAC Accredited “A++” Grade University)  Green Fields, Guntur District, A.P., India – 522502  **Department of Computer Science and Engineering**  (DST - FIST Sponsored Department) |  |

**B.Tech. II CSE(H) PROGRAM**

**A.Y. 2023-24 ODD, Semester-II**

**Course Code: 22MT2005**

**PROBABILITY, STATISTICS AND QUEUING THEORY**

**Course Outcome-2**

**Session 16:** **SCATTER DIAGRAM/Correlations**

1. **Course Description (Description about the subject)**

A scatter diagram is a type of graph that shows the relationship between two variables. The variables are plotted on the x-axis and y-axis, and the points are connected with lines. The shape of the scatterplot can help to determine the type of relationship between the two variables.

1. **Aim**

To explain identify the linear relationship between two variables using different measures of correlation and regression.

1. **Instructional** **Objectives (Course Objectives)**

To Calculate the linear relationship between two variables using different measures of correlation and regression

1. **Learning** **Outcomes (Course Outcome)**

**CO2**: Students will be able to Explain different measures of central tendency, and dispersion.

1. **Module** **Description** **(CO-2 Description)**

Types of Scatter diagrams, Correlation and Regression

1. **Session** **Introduction**

Scatter diagrams are a useful tool for visualizing the relationship between two variables. They can be used to identify the type of relationship, to identify outliers, and to make predictions.

1. **Session description**

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 (Xi, Yi) 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.

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.

If r>0 indicates that positive correlation.

Y Y

0 X 0 X

Strong positive Weak positive

**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.

Y Y

0 X 0 Y

Strong negative Weak negative

**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.

Y Y

0 X 0 X

correlation No correlation

Correlation coefficients always fall between -1 to +1.

**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:

* **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 any connection between the two variables.

**Correlation:**

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 {(xi, yi); 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 aluminium between the textile strength and the hardness of aluminium.

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

Correlation coefficient measures the linear association between x and y.

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

r is called the Pearson product-moment correlation coefficient.

**Properties of r**

1. -11

2. r=1 if all pairs (xi, yi) 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 (xi, yi) 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.

1. **Activities/ Case studies/related to the session.**

**NA**

1. **Examples & contemporary extracts of articles/ practices to convey the idea of the Session**

**Example:** 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 | x2 | y2 |
| 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** |

Here N=6

Mean of x==225/6=37.5

Mean of y= =22.5/6=3.75

Standard deviation of x (S.D. of x) = =14.2332

Standard deviation of y (S. D. of y) ==3.0788

Covariance (x,y) = = -36.4583

Correlation (r) = = -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.

1. **SAQ's-Self Assessment Questions**
2. Correlation measures the degree of:

a) Causation between two variables

b) Association between two variables

c) Linear relationship between two variables

d) Difference between two variables

1. The correlation coefficient ranges between:

a) -1 and 1

b) 0 and 1

c) -∞ and ∞

d) -1 and 0

1. If the correlation coefficient is close to -1, it indicates:

a) A strong positive relationship

b) A strong negative relationship

c) No relationship between variables

d) A nonlinear relationship

1. If there is a very strong correlation between two variables then the correlation coefficient must be

a. any value larger than 1

b. much smaller than 0, if the correlation is negative

c. much larger than 0, regardless of whether the correlation is negative or positive

d. None of these alternatives is correct

1. The correlation coefficient is used to determine:

a. A specific value of the y-variable given a specific value of the x-variable

b. A specific value of the x-variable given a specific value of the y-variable

c. The strength of the relationship between the x and y variables

d. None of these

**Answers**:

1. b) Association between two variables
2. a) -1 and 1
3. b) A strong negative relationship
4. b) much smaller than 0, if the correlation is negative
5. c) The strength of the relationship between the x and y variables
6. **Summary**

The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. The coefficient is what we symbolize with the r in a correlation report.

1. **Terminal Questions**
   1. 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/m3) | 126 | 121 | 116 | 118 | 114 |

1. Make a scatter plot for the given data
2. Determine the correlation coefficient for the given data.
   1. Calculate the correlation coefficient of the given data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | 50 | 51 | 52 | 53 | 54 |
| y | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 |

* 1. Calculate the correlation coefficient of the given data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | 12 | 15 | 18 | 21 | 27 |
| y | 2 | 4 | 6 | 8 | 12 |

* 1. For the sample data

x 0 1 3 5 8

y 2 4 6 5 9

Draw the scatter plot.

1. Based on the scatter plot, predict the sign of the linear correlation coefficient. Explain your answer.
2. Compute the linear correlation coefficient and compare its sign to your answer to part (b).
   1. The age x in months and vocabulary y were measured for six children, with the results shown in the table

x 13 14 15 16 16 18

y 8 10 15 20 27 30

Compute the linear correlation coefficient for these sample data and interpret its meaning in the context of the problem.

1. **Case Studies (CO Wise)**

**NA**

1. **Answer Key**

**NA**

1. **Glossary**

**NA**

1. **References of books, sites, links Text Books:**

**Textbooks:**

1. Probability and Statistics Rukmangad Achari E. and E. Keshava Reddy
2. Probability and Statistics for Engineers and Scientists” Ronald E. Walpole, Sharon L. Myers and Keying Ye 8th Edition Pearson pub
3. Probability & Statistics for Engineers Dr. J. Ravichandran first Edition Wiley-India

**Reference books:**

1. Hossein Pishro-Nik, Introduction to Probability, Statistics, and Random Processes, 2014, by Kappa Research LLC; ISBN-13: 978-0990637202

**Web Resources**

1. https://ncert.nic.in/textbook.php?kemh1=0- 16
2. https://ncert.nic.in/textbook.php?jemh1=ps-15
3. **Keywords**

Mean, median, mode, central tendency, Interval, continuous data, discrete data.