Day 4
Measures of Association and Modelling

Session 1

Measures of Association

Session Outcome

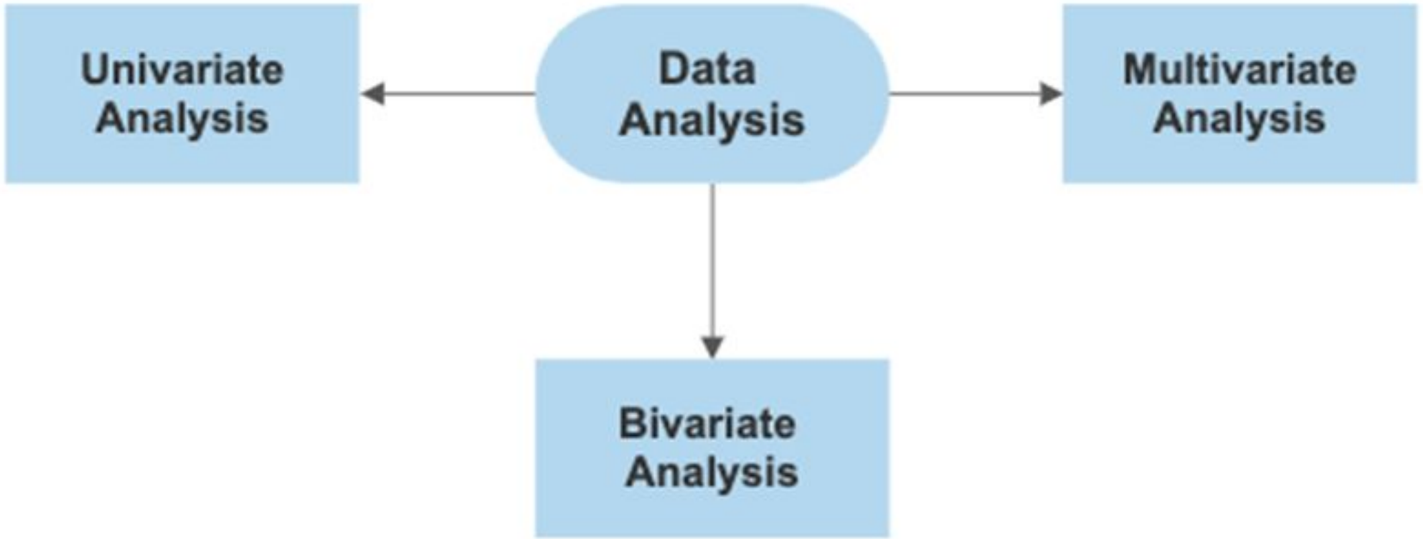
After completing this session, researchers will be able to

- Understand Measures of association (or relationship) between Variables.
- Understand the Concept of Association (or relationship) between Variables.
- Differentiate Different Types of Association
- Examining the tests for categorical variables and continuous variables
- Interpret the findings from the statistical tests.

Session Outline

- Concepts of Measures of Association and its Importance
- Different Measures of Association
- Different types of Correlation
- Contingency Table, Chi-square test and Spearman’s Rank Correlation Coefficients

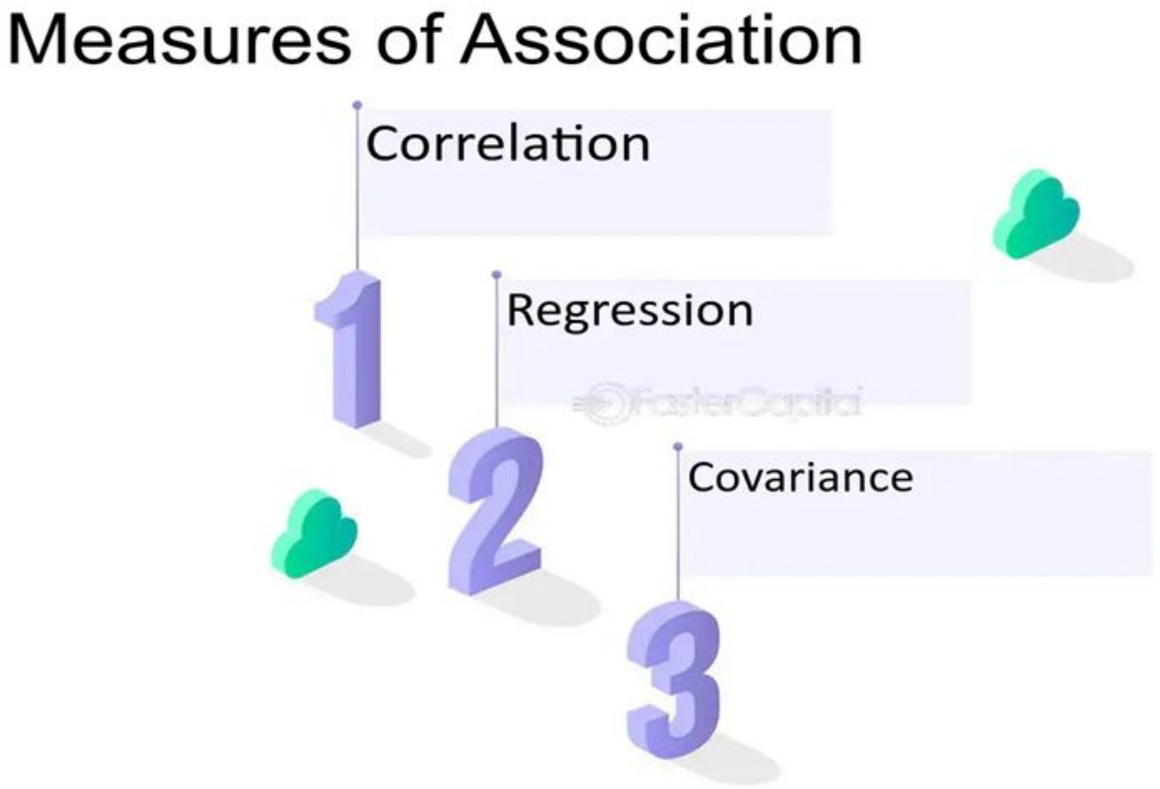
Measures of Association



Measures of Association

Measures of association refer to a wide variety of bivariate statistics that quantify the strength and direction of the relationship between exposure and outcome variables, enabling comparison between different groups.

Measures of Association

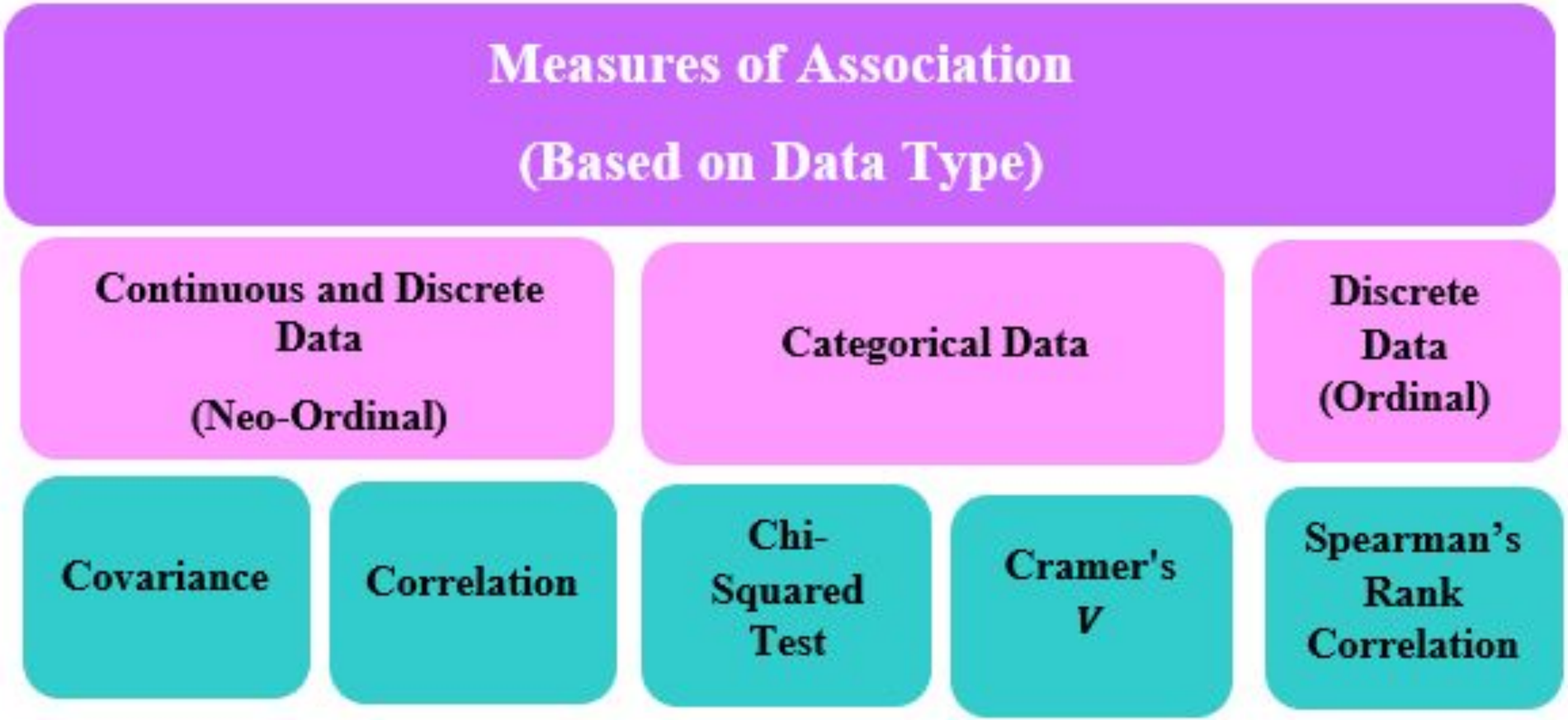


Importance of Measures of Association

Measure of Association helps in the following aspects of primary education:

- **Curriculum Development**
- **Teacher Training and Professional Development**
- **Predictive Modeling**
- **Decision Making**
- **Risk Assessment and Management**
- **Identifying Causal Relationships**
- **Targeted Interventions**
- **Quality Improvement**
- **Research and Exploration**

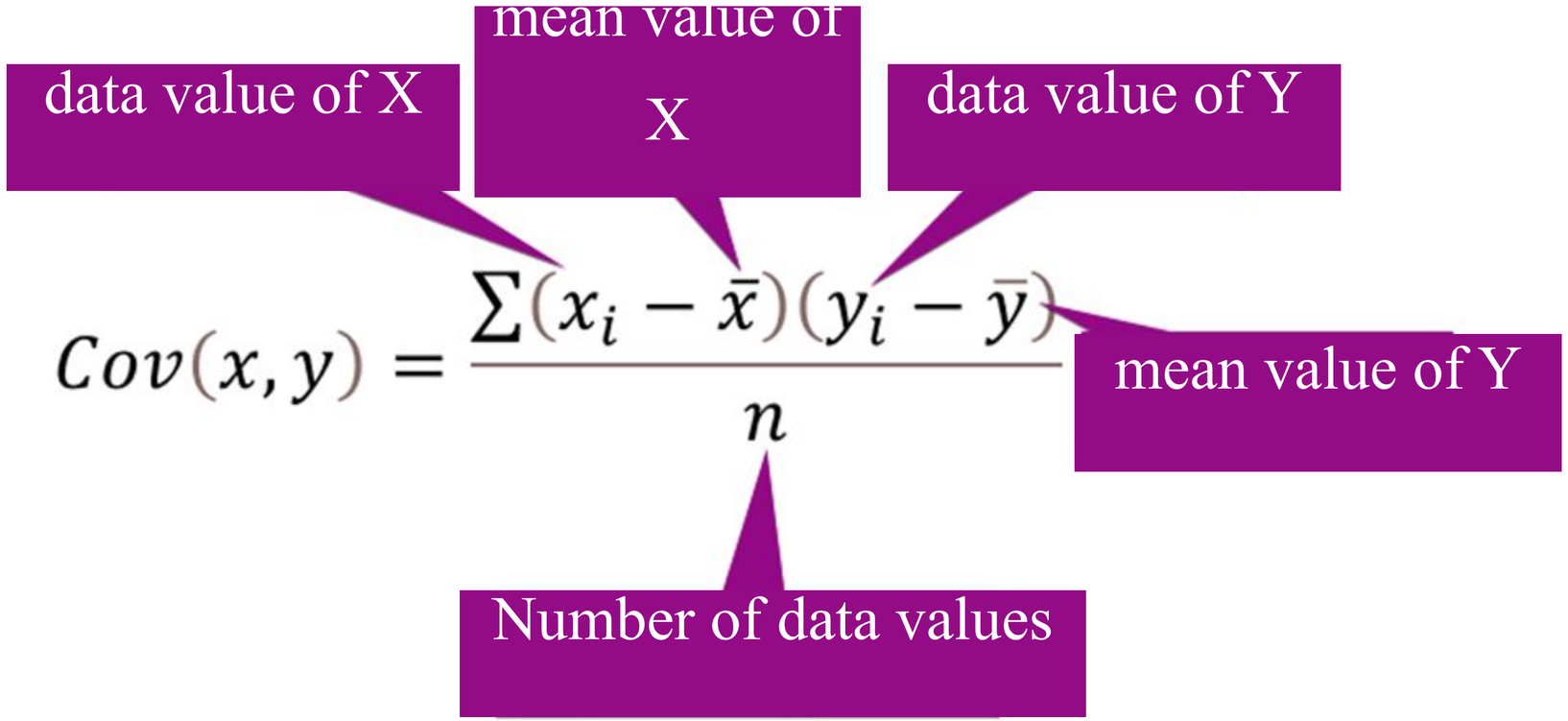
Measures of Association



Covariance

1. Association of Continuous and Discrete Data (Neo- Ordinal)

Covariance: Covariance measures the change in two random variables, indicating their linear relationship. It aids decision-makers in business, economics, and public policy in identifying key factors influencing outcomes and formulating effective strategies.



Covariance

Mathematically, the covariance between two variables, X and Y, can be calculated as:

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Where:

- n is the number of data points
- x_i and y_i are individual data points for variables X and Y respectively
- $\bar{x}$ *and* $\bar{y}$ are the means of variables X and Y respectively.

Interpretation of covariance:

- If the covariance is positive, it means that as one variable increases, the other variable also tends to increase.
- If it is negative, it means that as one variable increases, the other variable tends to decrease.
- If it is close to zero, it means that there is little to no linear relationship between the

Measures of Association

Interpretation of covariance:

- The magnitude of the covariance is not standardized and depends on the scales of the variables, making it difficult to compare covariances across different datasets.

Example: Data is collected from a group of students, focusing on the number of hours spent studying and the test scores they achieved.

Student	Hours Studied	Test Score
1	5	80
2	3	75
3	6	85
4	4	70
5	7	90

Solution : To calculate covariance, you'd first find the mean of each variable:

- Mean Hours Studied: $\bar{x} = \frac{(5+3+6+4+7)}{5} = \frac{25}{5} = 5$ hours
- Mean Test Score: $\bar{y} = \frac{(80+75+85+70+90)}{5} = \frac{400}{5} = 80$

Measures of Association

$x_i - \bar{x}$	(5 - 5) = 0	(3 - 5) = -2	(6 - 5) = 1	(4 – 5) = -1	(7 – 5) = 2
$y_i - \bar{y}$	(80 – 80) = 0	(75 – 80) = -5	(85 – 80) = 5	(70 – 80) = -10	(90 – 80) = 10
$(x_i - \bar{x}) \times (y_i - \bar{y})$	0	10	5	10	20

Finally,

Covariance, $Cov(X, Y) =$

$$\frac{(0+10+5+10+20)}{5} = \frac{45}{5} = 9$$

Interpretation: The covariance of 9 suggests a positive relationship between the number of hours studied and test scores, indicating that as study hours increase, test scores tend to increase as well.

Correlation

Correlation measures the strength and direction of association between two variables, indicating how changes in one variable correspond to changes in another.

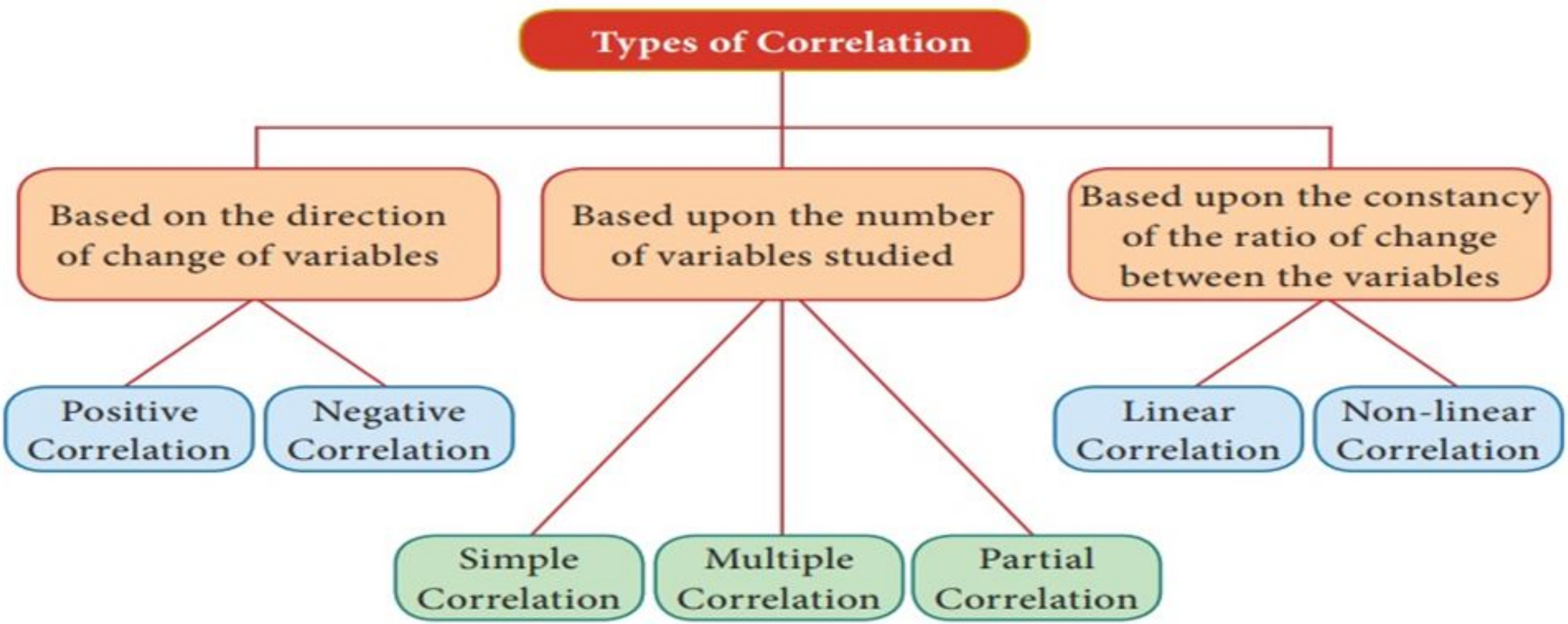
- It ranges from -1 to +1, where,
- 0 implies no linear relationship,
- +1 implies a perfect positive relationship, and
- -1 implies a perfect negative relationship.

Correlation

Importance of Correlation in Primary Education

- **Assessment of Student Performance**
- **Evaluation of Teaching Methods**
- **Identification of Learning Barriers**
- **Resource Allocation**
- **Teacher Development and Support**
- **Policy Formulation and Evaluation**
- **Early Intervention and Remediation**

Types of Correlation



Simple Correlation

Simple Correlation: There are different types of correlation measures, but the most common is the simple correlation, is a measure of the strength and direction of the linear relationship between two continuous variables.

Methods of Simple Correlation:

There are three different methods of measuring the correlation between two variables:

1. Scatter Diagram

2. Karl Pearson's Coefficient of Correlation

3. Spearman's Rank Correlation Coefficient

Simple Correlation

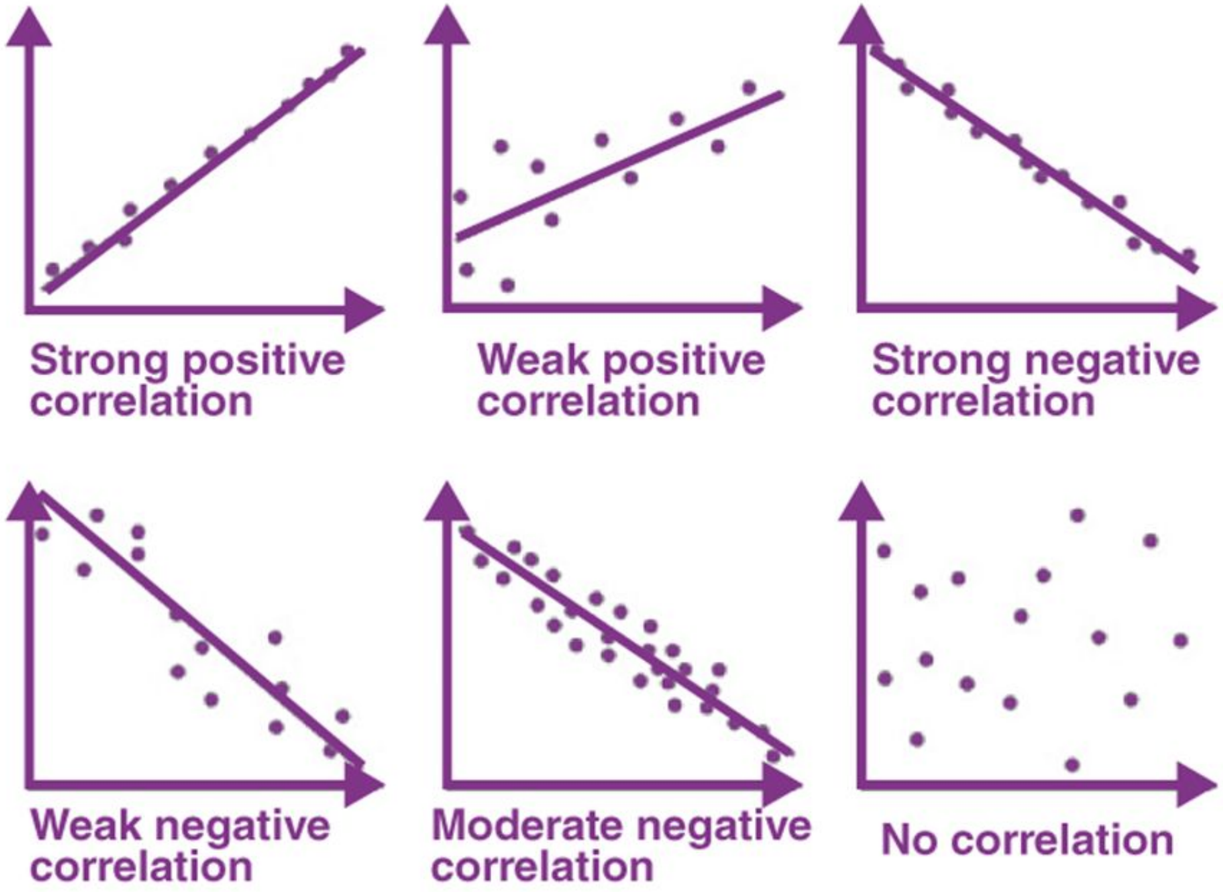
Scatter Diagram: A simple and attractive method of measuring correlation by diagrammatically representing bivariate distribution for determination of the nature of the correlation between the variables is known as Scatter Diagram Method.

- Provides a visual understanding of association nature.
- Simplest method without numerical value calculation.

The two steps required to draw a Scatter Diagram or Dot Diagram are as follows:

- Plot the values of the given variables (say X and Y) along the X-axis and Y-axis respectively.
- Show these plotted values on the graph by dots. Each of these dots represents a pair of values.

Simple Correlation



Simple Correlation

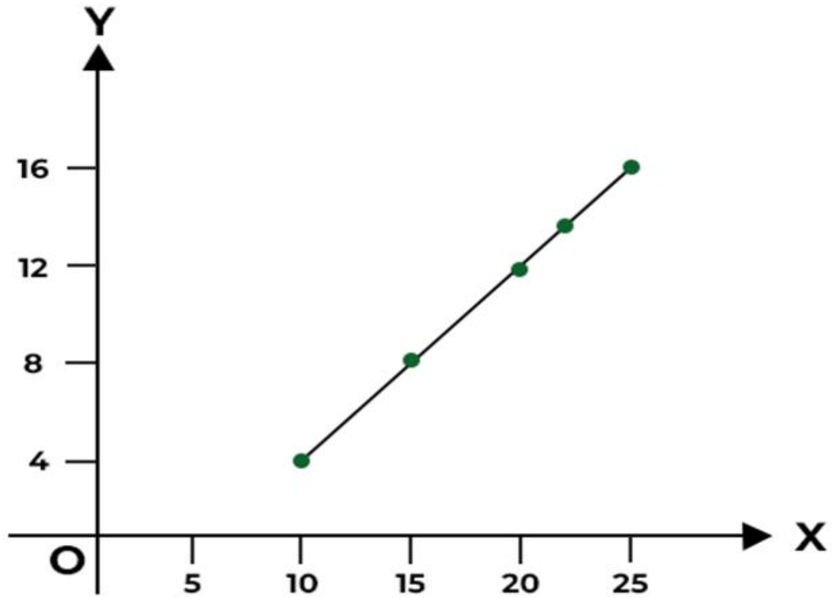
Example: Represent the following values of X and Y variables with the help of a scatter diagram. Also, comment on the type and degree of correlation.

X	10	15	20	22	25
Y	5	9	13	15	17

Solution

Interpretation

The scatter diagram shows that there is an upward trend of the points from the lower left-hand corner to the upper right-hand corner of the graph. In short, there is a Positive Correlation between the values of X and Y variables.



Simple Correlation

2. Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation is also known as Product Moment Correlation or Simple Correlation Coefficient. This method of measuring the coefficient of correlation (r) is the most popular and widely used.

The formula for calculating the Pearson correlation coefficient between two variables X and Y is:

$$r = Corr(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \times \sigma_y}$$

Simple Correlation

To calculate the correlation coefficient manually, follow these steps:

Calculate the mean ($\bar{x}$) of variable X.

Calculate the mean ($\bar{y}$)and standard deviation σ_y of variable Y.

Calculate the correlation coefficient (r) using the formula above.

Alternatively, statistical software such as R, and Python with libraries like NumPy or pandas, or spreadsheet software like Microsoft Excel can be used to calculate the correlation coefficient quickly and accurately.

Simple Correlation

Correlation Coefficient	Interpretation
1	perfect positive correlation
close to 1	strong positive correlation
close to 0.5	weak positive correlation
close to 0	weak or no correlation
close to -0.5	weak negative correlation
close to -0.8	strong negative correlation
close to -1	strong negative correlation

Simple Correlation

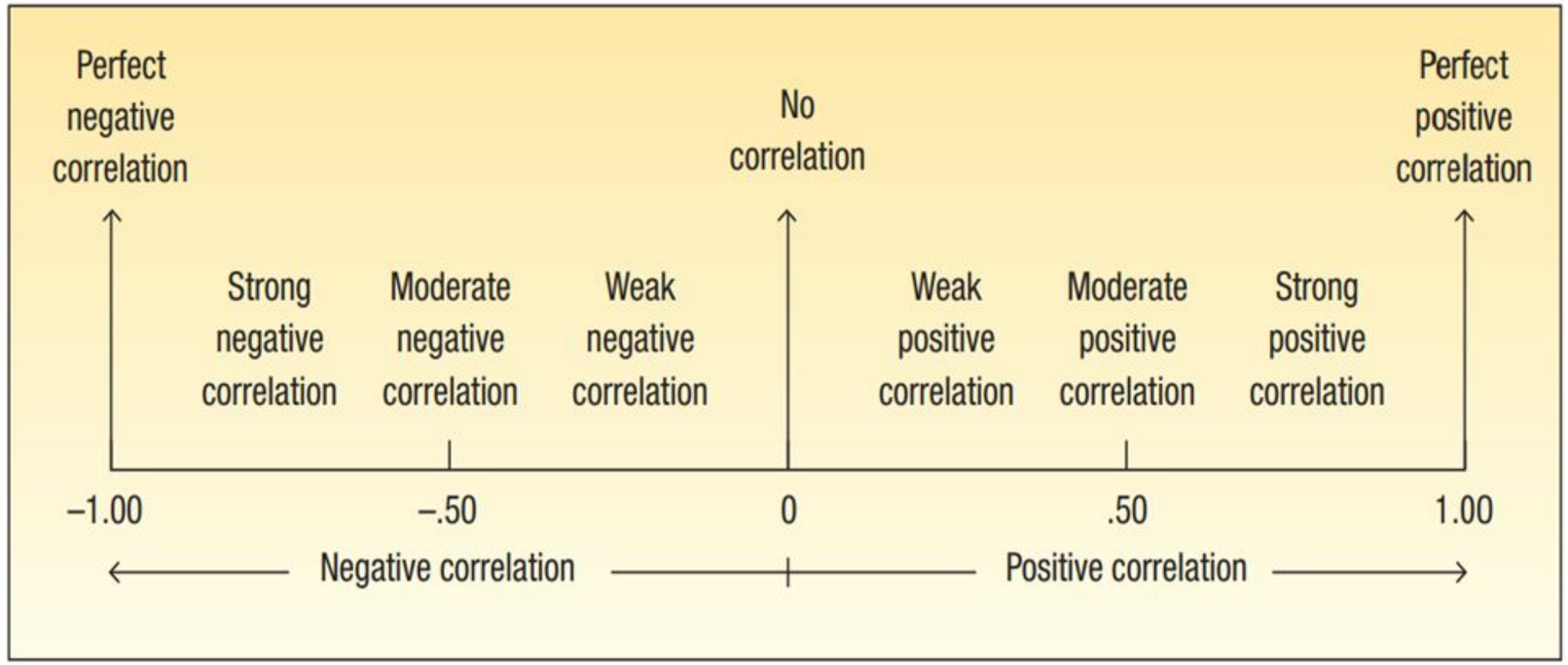


Figure: The strength and direction of the correlation coefficient.

Simple Correlation

Example

To calculate the Pearson correlation coefficient (r) with a sample size of 10, we'll use the provided data for the number of hours of teacher training and average test scores.

The data is given by the following:

School	Hours of Teacher Training (X)	Average Test Score (Y)	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	40	85	6.25	1	-2.5
2	35	78	56.25	36	45
3	50	90	56.25	36	45
4	45	82	6.25	4	-5
5	55	92	156.25	64	100
6	30	75	156.25	81	112.5
7	60	98	306.25	196	245
8	25	70	306.25	196	245
9	65	105	506.25	441	472.5
10	20	65	506.25	361	427.5
Total	425	840	2062.5	1416	1685

Simple Correlation

Let's follow the steps to calculate the Pearson correlation coefficient (r):

Step 1: Calculate the mean $(\bar{x})$ and $(\bar{y})$:

$$\bar{x} = \frac{\sum x}{n} = \frac{40 + 35 + \cdots + 20}{10} = 42.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{85 + 78 + \cdots + 65}{10} = 83.0$$

Step 2: Calculate the sum of the products of deviations:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 160.85$$

Simple Correlation

Step 3: Calculate the standard deviation (σ_x) and (σ_y):

$$\sigma_x = \sqrt{\sum (x_i - \bar{x})^2} = \sqrt{2062.5} = 45.42$$

$$\sigma_y = \sqrt{\sum (y_i - \bar{y})^2} = \sqrt{1416} = 37.62$$

Step 4: Calculate the correlation coefficient (r):

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \times \sigma_y} = \frac{160.25}{45.42 \times 37.62} = 0.986$$

Interpretation: The Pearson correlation coefficient (r) is approximately 0.98. This indicates a strong positive correlation between the number of hours of teacher training and the average test scores of students in the sample of primary schools.

Simple Correlation

Spearman’s Rank Correlation Coefficient

Spearman's rank correlation coefficient is a non-parametric measure of the strength and direction of association between two ranked variables. Spearman’s Rank Correlation Coefficient or Spearman’s Rank Difference Method or Formula is a method of calculating the correlation coefficient of qualitative variables.

It denoted by ρ (rho)

Here's how Spearman's rank correlation coefficient is calculated:

Ranking the Data: For each variable, rank the observations from lowest to highest, assigning ranks in case of ties.

Calculating the Differences: Calculate the differences between the ranks of corresponding observations for both variables.

Simple Correlation

Squaring the Differences: Square each of these differences.

Summing the Squares of Differences: Sum up all the squared differences.

The formula for calculating Spearman's Rank Correlation Coefficient is,

$$\rho = 1 - \frac{6 \sum {d_i}^2}{n(n^2 - 1)}$$

- ρ is the rank correlation coefficient
- d_i represents the differences between the ranks for each pair of observations
- n is the number of observations

Spearman's rank correlation coefficient ranges from -1 to 1:

- $\rho = 1$: Perfect positive monotonic relationship.
- $\rho = -1$: Perfect negative monotonic relationship.
- $\rho = 1$: No monotonic relationship.

Simple Correlation

Example: Suppose, analyzing the correlation between the rankings of primary schools in mathematics and science using Spearman's rank correlation coefficient (ρ), based on a sample of 10 schools.

Solution:

School	Mathematics Rank	Science Rank	d_i (Difference)	d_i^2
A	1	3	2	4
B	2	5	3	9
C	3	1	2	4
D	4	2	2	4
E	5	6	1	1
F	6	4	2	4
G	7	8	1	1
H	8	7	1	1
I	9	9	0	0
J	10	10	0	0

$$\sum d_i^2 = 4 + 9 + 4 + 4 + 1 + 4 + 1 + 0 + 0 = 28$$

Simple Correlation

Substitute the values into the formula for Spearman's rank correlation coefficient :

$$\rho = 1 - \frac{6 \times \sum {d_i}^2}{n(n^2 - 1)} = 1 - \frac{6 \times 28}{10(10^2 - 1)} \approx 0.830$$

Interpretation: The Spearman's rank correlation coefficient (ρ) indicates a strong positive correlation between primary school rankings in mathematics and science performance.

Chi-squared Test

Chi-square (χ^2) is a statistical test assessing significant association between two categorical variables in a contingency table, assessing independence and requiring knowledge of contingency tables.

Chi-squared Test

Contingency Table: A contingency table, also known as a cross-tabulation or two-way frequency table, is a statistical tool used to display the frequency distribution of categorical variables. Here's a general format for a contingency table:

	Category 1	**Category 2**	**. . .**	**Category m**	**Total**
Group 1	n_{11}	n_{12}	. . .	n_{1n}	n_{1+}
Group 2	n_{21}	n_{22}	. . .	n_{2n}	n_{2+}
. . .	. . .	. . .	. . .	. . .	. . .
Group n	n_{n1}	n_{n2}	. . .	n_{nm}	n_{n+}
Total	n_{+1}	n_{+2}	. . .	n_{+m}	N=n_{++}

- Rows represent the categories of one variable (e.g., groups, conditions).
- Columns represent the categories of another variable (e.g., outcomes, responses).
- n_{ij} represents the frequency or count of observations in the i^{th}row and j^{th} column
- The total of each row ($n_{1+} + n_{2+} + \ldots + n_{n+}$) represents the total count for each category of the first variable.
- The total of each column ($n_{+1} + n_{+2} + \ldots + n_{+m}$) represents the total count for each category of the second variable.
- The grand total (N) represents the total number of observations in the entire dataset.

Simple Correlation

Example:

Let's create a simple example of a contingency table relevant to primary education research. We'll consider the relationship between teaching method (traditional vs. modern) and student performance (Below average, Above average) based on survey data collected from primary school students:

	Traditional Method	Modern Method	Total
Below Average Performance	150	50	200
Above Average Performance	100	200	300
Total	250	250	500

Contingency Tables

Contingency tables are commonly used to visually inspect the relationship between categorical variables and to conduct statistical tests, such as the chi-square test for independence or Fisher's exact test, to determine if there is a significant association between the variables.

a) Pearson Chi-square,

b) Fisher's Exact Test

a) Pearson Chi-square

The Pearson chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. The formula for calculating the Pearson chi-square statistic

Pearson Chi-square

a) Pearson Chi-square

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where

- χ^2 is the Pearson chi-square statistic,
- O_i is the observed frequency for each category,
- E_i is the expected frequency for each category.

The observed frequencies (O_i) are the actual counts observed in each category of the contingency table.

Hypotheses:

Null Hypothesis (H_0):

There is no association between the two variables. They are independent.

Alternative Hypothesis (H_1):

There is an association between the two variables. They are dependent.

Pearson Chi-square

Expected frequencies:

The expected frequencies are the counts that would be expected in each category if there were no association between the variables. They are calculated under the assumption of independence between the two variables. The formula to calculate the expected frequency for each category:

$$E_i = \frac{(\text{row total} \times \text{column total})}{grand\ total}$$

Where:

- Row total, row total is the total count in the row of the contingency table containing category,
- Column total column total is the total count in the column of the contingency table containing category,
- Grand total is the total count of all observations in the contingency table

Pearson Chi-square

Pearson Chi-square Statistics Calculation

- Compares to critical value from chi-square distribution.
- Degrees of freedom equal to (r-1) (c-1).
- If calculated value exceeds critical value, null hypothesis rejected.

Example: Suppose the ministry is interested in studying the association between two categorical variables: type of teaching method used in primary schools (traditional vs. modern) and student performance on standardized tests (below average vs. above average). The ministry collects data from 500 primary schools across the country and creates a contingency table as follows:

Pearson Chi-square

	Traditional Method	Modern Method	Total
Below Average Performance	150	50	200
Above Average Performance	100	200	300
Total	250	250	500

Now, the ministry wants to determine if there is a significant association between the type of teaching method used and student performance on standardized tests.

Step 1 Calculate Expected Frequencies (E_i)

$$E_i = \frac{(\text{row total} \times \text{column total})}{grand\ total}$$

- Expected frequency for cell (1,1): $E_{11} = \frac{(200 \times 250)}{500} = 100$
- Expected frequency for cell (1,2): $E_{12} = \frac{(200 \times 250)}{500} = 100$
- Expected frequency for cell (2,1): $E_{21} = \frac{(300 \times 250)}{500} = 150$
- Expected frequency for cell (2,2): $E_{22} = \frac{(300 \times 250)}{500} = 150$

Pearson Chi-square

Step 2 Calculate Pearson Chi-square Statistic(χ^2)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(150-100)^2}{100} + \frac{(50-100)^2}{100} + \frac{(100-150)^2}{150} + \frac{(200-150)^2}{150}$$

$$\approx 116.67$$

Step 3 Compare χ^2 to Critical Value:

With one degree of freedom (since we have 2 rows and 2 columns), and using a significance level of $\alpha = 0.05$, the critical value from the chi-square distribution is approximately 3.84. Since $\chi^2 = 116.67$ is greater than the critical value of 3.84, we reject the null hypothesis.

Interpretation: The study confirms a significant association between the type of teaching method used in primary schools (traditional vs. modern) and student performance on standardized tests, rejecting the null hypothesis of independence.

Fisher’s Exact Test

Fisher's exact test is a statistical method used to determine nonrandom associations between categorical variables, particularly useful in small sample sizes or when chi-square test assumptions are not met. It calculates the probability of obtaining observed data and extreme outcomes.

Consider a contingency table with two categorical variables, A and B, and the observed frequencies are as follows:

	A=1	A=2
B=1	a	b
B=2	c	d

The probability of obtaining this specific arrangement of frequencies, given the row and column marginal totals, is calculated as:

Fisher’s Exact Test

$$P = \frac{\binom{a+b}{a} \times \binom{c+d}{c}}{\binom{n}{a+c}}$$

Where:

- n is the total sample size (sum of all frequencies).
- $\binom{m}{k}$ represents the binomial coefficient, which calculates the number of ways to choose k elements from a set of m elements.

The Fisher's exact test calculates the probability of observing extreme frequencies and compares it to a significance level. If P is less than or equal to the chosen level, the null hypothesis is rejected. It's useful for small sample sizes but computationally intensive for large contingency tables.

Fisher's Exact Test

Example: Let's ministry wants to determine if there is an association between the type of teaching method used (traditional vs. modern) and student performance on a standardized test (pass vs. fail). Suppose the ministry selects a random sample of 50 primary schools and records the following contingency table:

	Traditional Teaching	Modern Teaching
Pass	a=20	b=15
Fail	c=5	d=10

We can calculate P using the formula:

$$P = \frac{\binom{a+b}{a} \times \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{35}{20} \times \binom{15}{5}}{\binom{50}{25}} = 284.20$$

The calculated probability of P exceeding 1 indicates an error due to large numbers or precision issues.

Cramer's V

Cramer's V measures association strength and direction in square contingency tables, an extension of Pearson's chi-squared test for independence. Cramer's V ranges from 0 to 1, where:

- 0 indicates no association between the variables.
- 1 indicates a perfect association between the variables.

The formula to calculate Cramer's V is:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

Where:

- χ^2 is the Pearson chi-squared statistic
- n is the total number of observations in the contingency table
- k is the number of cells in the contingency table (i.e., the number of categories in each variable).

Cramer's V

Interpretation of Cramer's V:

- Small effect (weak association): Close to 0.
- Medium effect (moderate association): Around 0.3.
- Large effect (strong association): Near 0.5 or higher.

Example

Let's consider an example where the relationship between two categorical variables: teaching method (traditional vs. modern) and student performance (below average, average, above average). Suppose we have the following contingency table representing the frequencies of students based on their performance and the teaching method used:

Cramer's V

	Below average	Average	Above average	Total
Traditional	20	30	10	60
Modern	15	25	20	60
Total	35	55	30	120

We want to determine if there's a significant association between teaching method and student performance.

Calculate Chi-Square Test Statistic:

First, we calculate the expected frequencies for each cell assuming independence between teaching method and student per

$$E_i = \frac{(\text{row total} \times \text{column total})}{grand\ total}$$

Cramer's V

$$E_i = \frac{(\text{row total} \times \text{column total})}{grand\ total}$$

- Expected frequency for cell (1,1): $E_{11} = \frac{(60 \times 35)}{120} = 17.5$
- Expected frequency for cell (1,2): $E_{12} = \frac{(60 \times 55)}{120} = 26.5$
- Expected frequency for cell (1,3): $E_{13} = \frac{(60 \times 30)}{120} = 15$
- Expected frequency for cell (2,1): $E_{21} = \frac{(60 \times 35)}{120} = 17.5$
- Expected frequency for cell (2,2): $E_{22} = \frac{(60 \times 55)}{120} = 26.5$
- Expected frequency for cell (2,3): $E_{23} = \frac{(60 \times 30)}{120} = 15$

Step 2 Calculate Pearson Chi-square Statistic(χ^2)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(20-17.5)^2}{17.5} + \frac{(30-26.5)^2}{26.5} + \frac{(10-15)^2}{15} + \frac{(15-17.5)^2}{17.5} + \frac{(25-26.5)^2}{26.5} + \frac{(20-15)^2}{15} = 4.5948$$

Cramer's V

Calculate Cramer's V:

After obtaining the chi-square test statistic and degrees of freedom, we calculate Cramer's V using the formula:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{4.5948}{120(3-1)}} = \sqrt{\frac{4.5948}{240}} = \sqrt{0.0191} \approx 0.1382$$

Interpret Cramer's V:

Cramer's V is approximately 0.2, it would suggest a moderate association between teaching method and student performance in primary education research. This would imply that the choice of teaching method has some impact on student performance, but there are likely other factors at play as well.

Correlation Ratio

The Correlation Ratio measures the strength and direction of the relationship between a nominal and continuous variable by comparing the variance explained by the groups to the total variance in the continuous variable.

Here's how the correlation ratio is calculated:

1. Compute the total sum of squares (SST), which represents the total variance in the dependent variable.
2. Compute the between-group sum of squares (SSB), which represents the variance in the dependent variable that can be attributed to the differences between groups defined by the independent variable(s).
3. Divide the between-group sum of squares (SSB) by the total sum of squares (SST) to obtain the correlation ratio (η^2).

The formula for correlation ratio is defined as

$$\eta^2 = \frac{SSB}{SST} = \frac{\sum_{j=1}^{k} n_j (Y_j - \bar{Y})^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}$$

Correlation Ratio

The correlation ratio ranges from 0 to 1, where:

- 0 indicates no association between the variables.
- 1 indicates a perfect association between the variables.

In the context of ANOVA, the correlation ratio quantifies the proportion of total variance in the dependent variable that is explained by the independent variable(s). It is particularly useful for assessing the strength of association in studies involving categorical variables.

Thank

You