Statistics Essentials for Data Science

Relation Between Variables

Learning Objectives

By the end of this lesson, you will be able to:

- Discuss the concepts of correlation and causation
- Examine the types of correlation coefficients, such as Karl Pearson's and Spearman's Rank Correlation
- Explain the coefficient of determination

Business Scenario

ABC, an organization that stores a large amount of data, aims to analyze the data and extract meaningful insights by determining the relationship between variables.

To accomplish this goal, the organization must learn how to determine correlation and causation and analyze various types of correlation coefficients, such as Karl Pearson's and Spearman's rank correlation.

Discussion: Relationship Between Variables

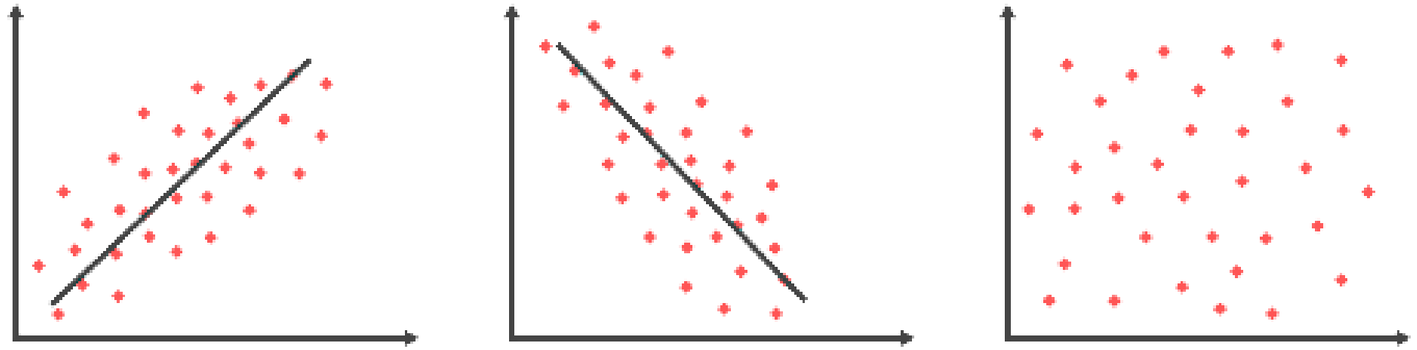
Duration: 15 minutes

- What does correlation mean?
- How do you determine the relationship between variables?

Correlation

Correlation

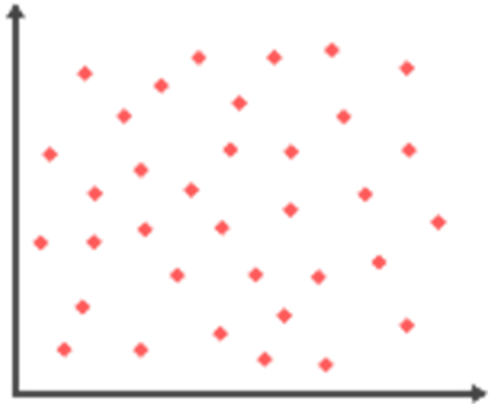
Correlation is a statistical measure that quantifies the extent to which two variables are linearly related.



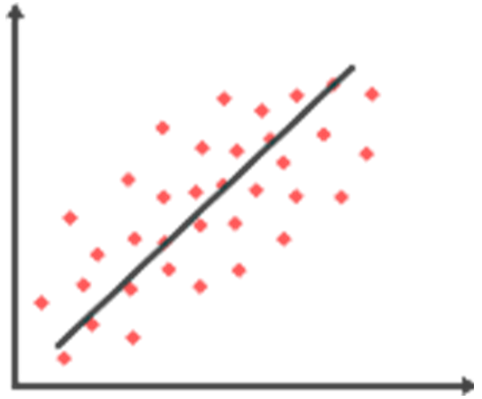
A scatter diagram helps visually illustrate the relationship between variables, providing a clear understanding of their interdependence.

Relationship Between Variables with Scatter Diagram

The nature of the scatter plot provides insights into the relationship between variables.



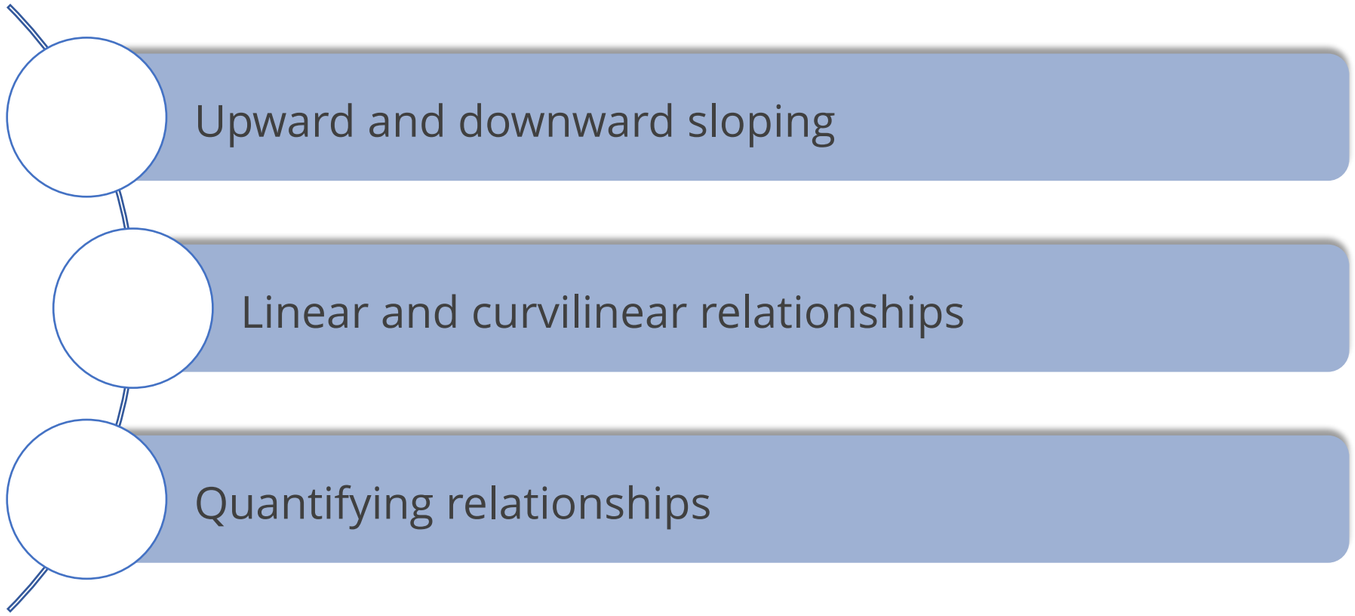
The absence of a band in the scatter plot suggests a lack of relationship between the variables.



The presence of a pattern in the shape of a band within the scatter plot indicates the existence of a relationship between the variables.

Relationship Between Variables with Scatter Diagram

A scatter diagram serves as a powerful visual aid for understanding correlation.
It can provide the following insights:



Upward-Sloping

The bands in Fig (a) and (b) are upward-sloping.

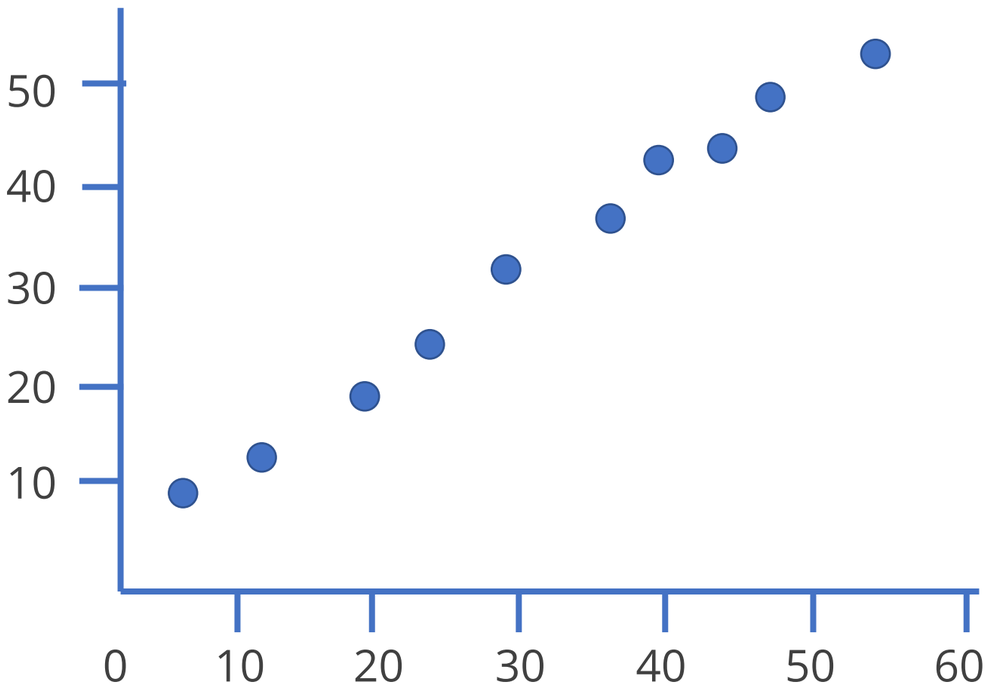


Fig (a)

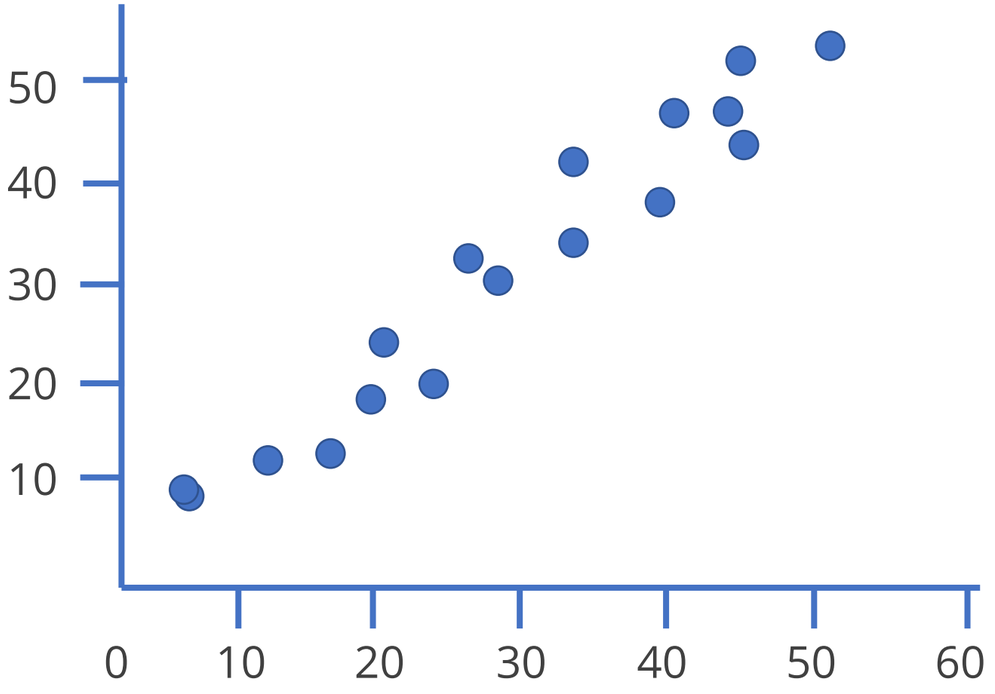


Fig (b)

This indicates that as one variable increases, the other variable also increases.

Downward-Sloping

The bands in Fig (c) and (d) are downward-sloping.

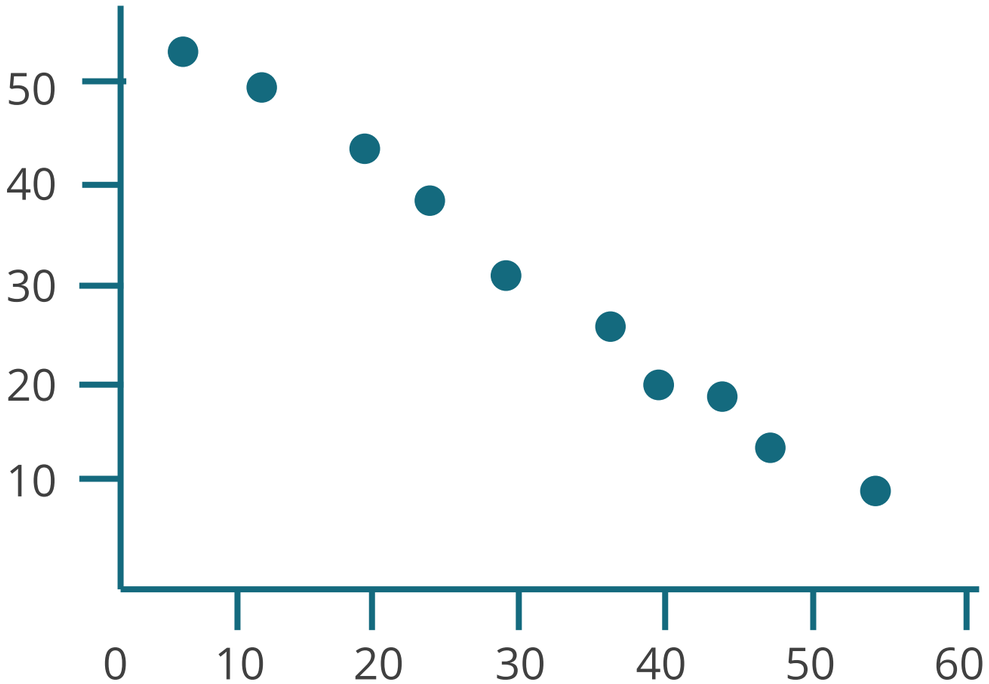


Fig (c)

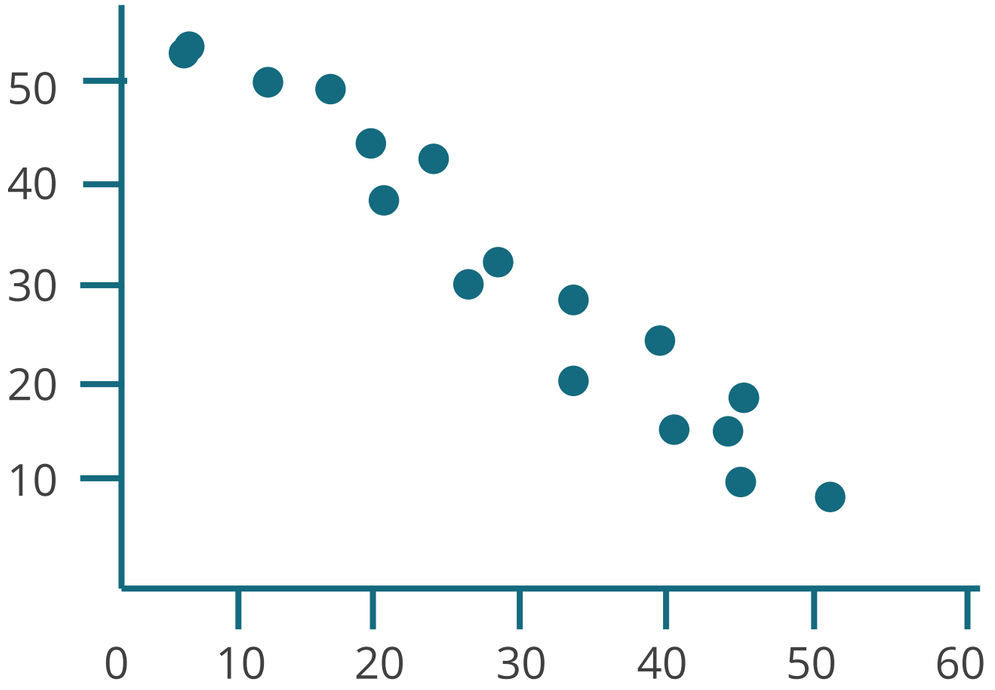


Fig (d)

This indicates that as one variable increases, the other variable decreases.

Width of Bands

The width of the bands in Fig (e) and (f) is narrow.

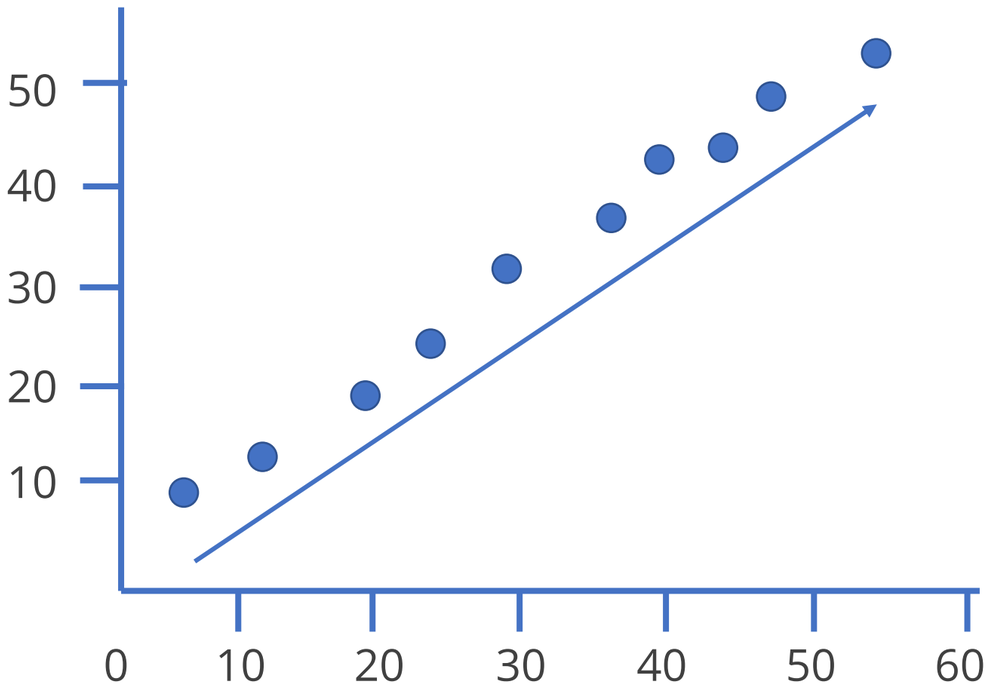


Fig (e)

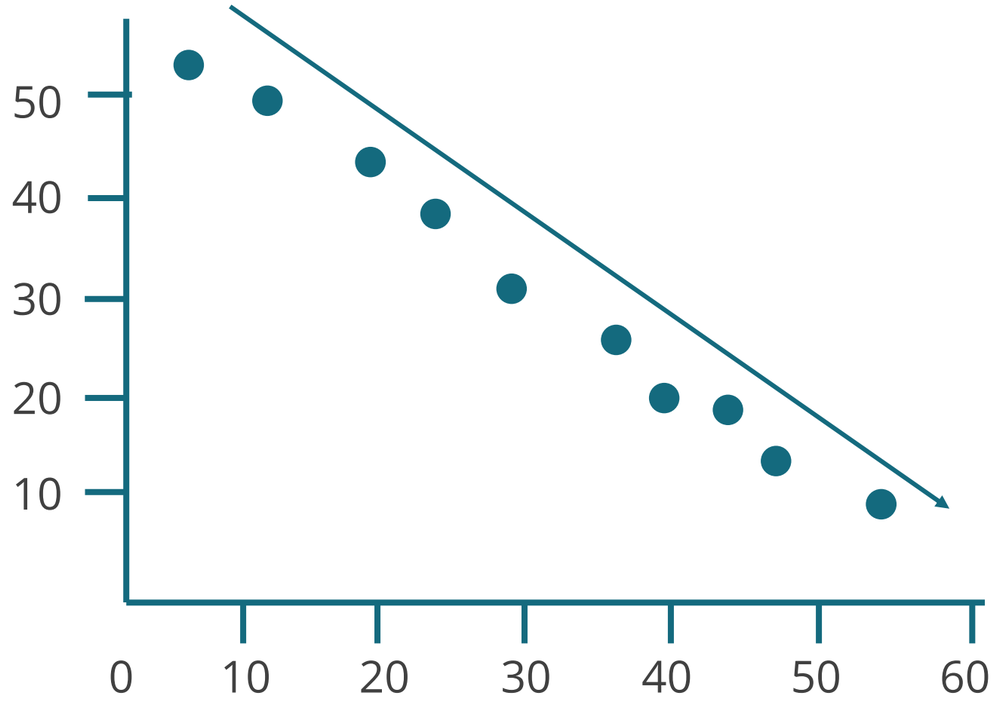


Fig (f)

A narrower band indicates a stronger relationship between the variables.

Width of Bands

The width in Fig (g) and (h) are relatively broader.

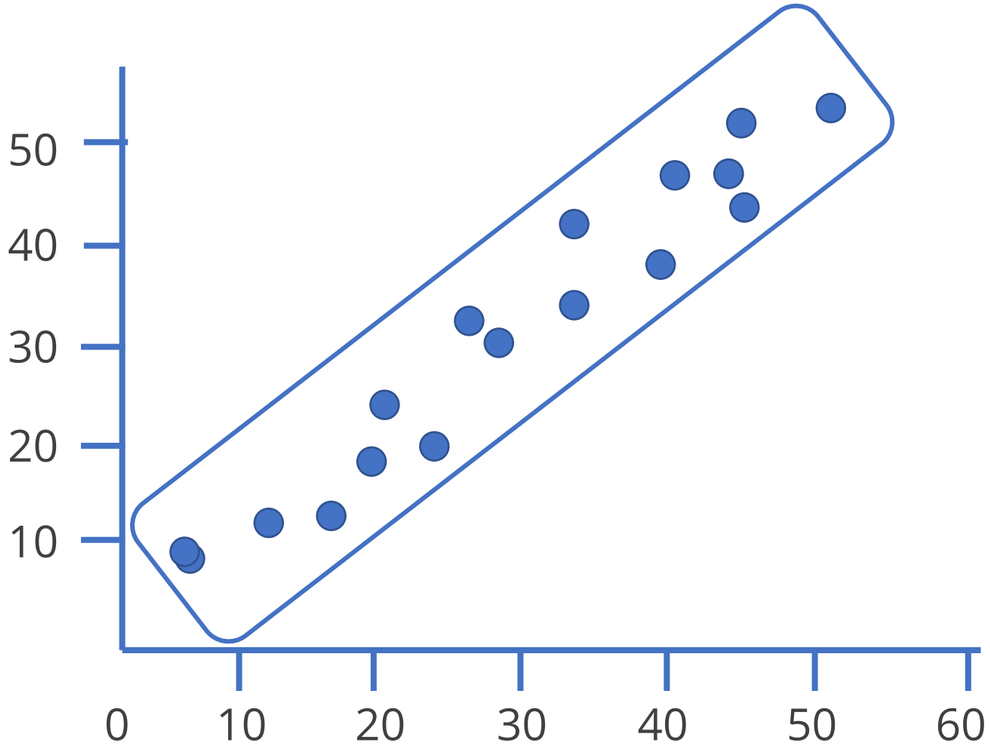


Fig (g)

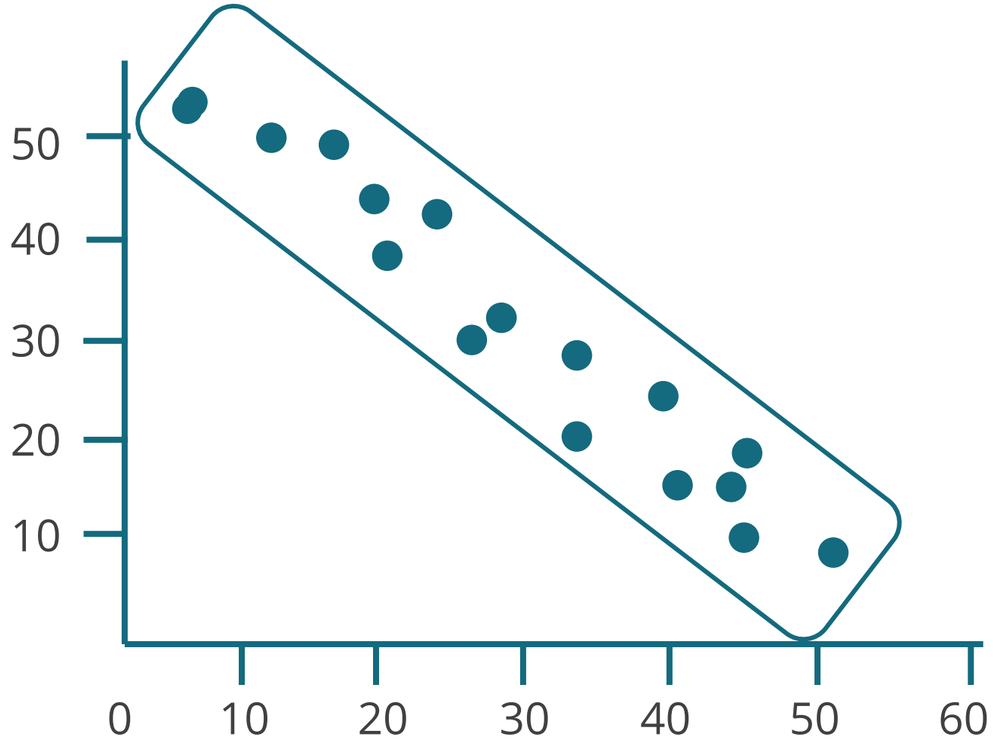
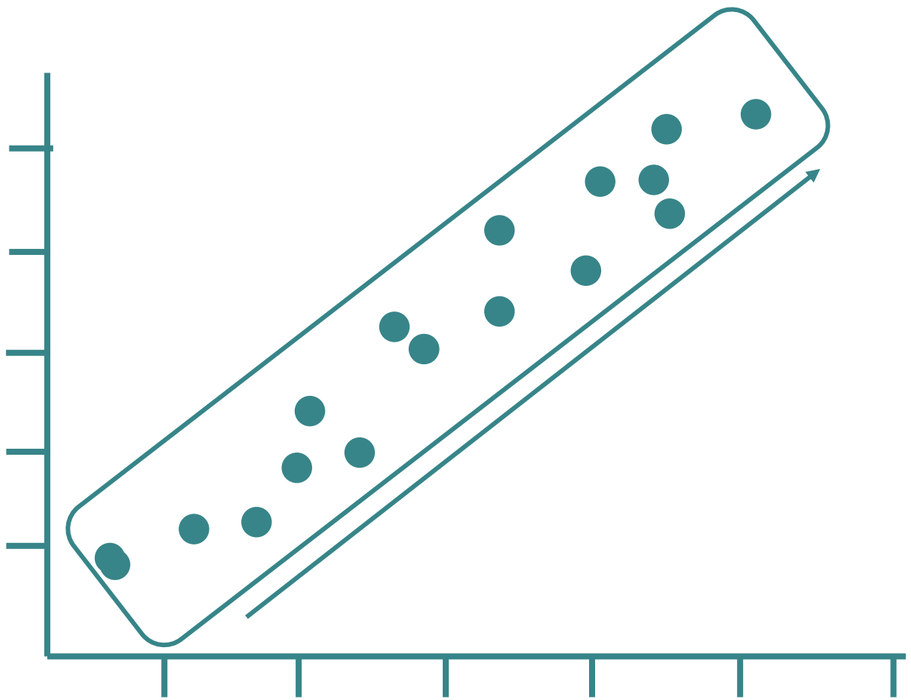


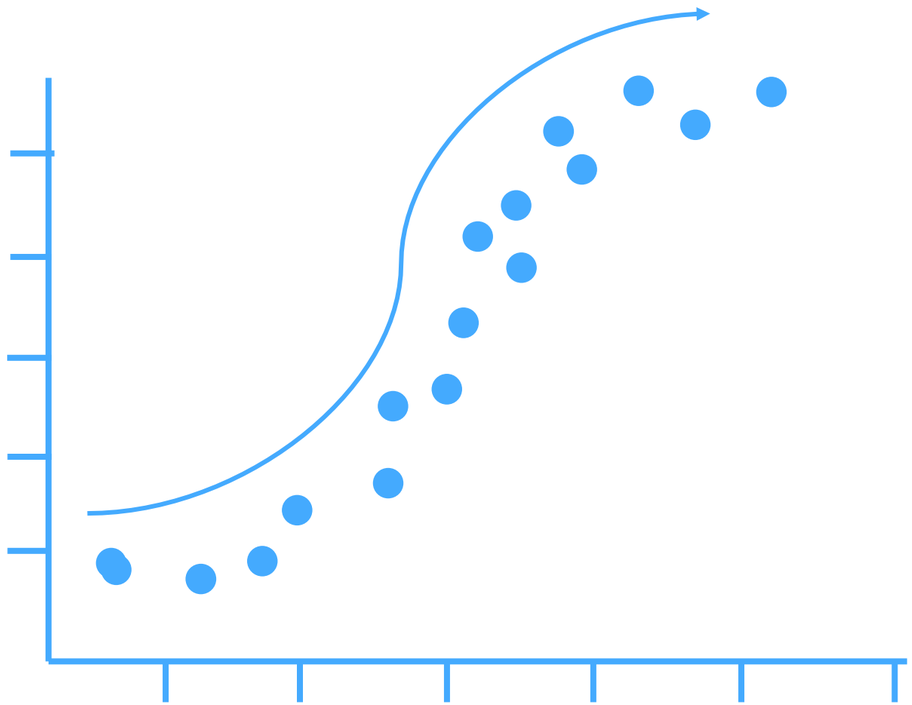
Fig (h)

A broader band indicates greater variability in the variables.

Linear and Curvilinear Relationships

When the band is almost a line or a rectangle, the relationship is linear.





If it is a curve, the relationship is curvilinear.

Quantifying Relationships

A quantitative measure is used to quantify the degree of the relationship.

	Level	Salary
Level	1.000000	0.817949
Salary	0.817949	1.000000

Discussion: Relationship Between Variables

Duration: 15 minutes

- What does correlation mean?
  Answer: Correlation refers to the statistical measure of the strength and direction of the association between two variables. It indicates how closely the variables are related to each other.

- How do you determine the relationship between variables?
  Answer: A scatter diagram serves as a powerful visual aid for understanding correlation. The absence of a band indicates a lack of relationship between the variables. The presence of a band, on the other hand, is indicative of a relationship.

Discussion

Discussion: Correlation and Covariance

Duration: 15 minutes

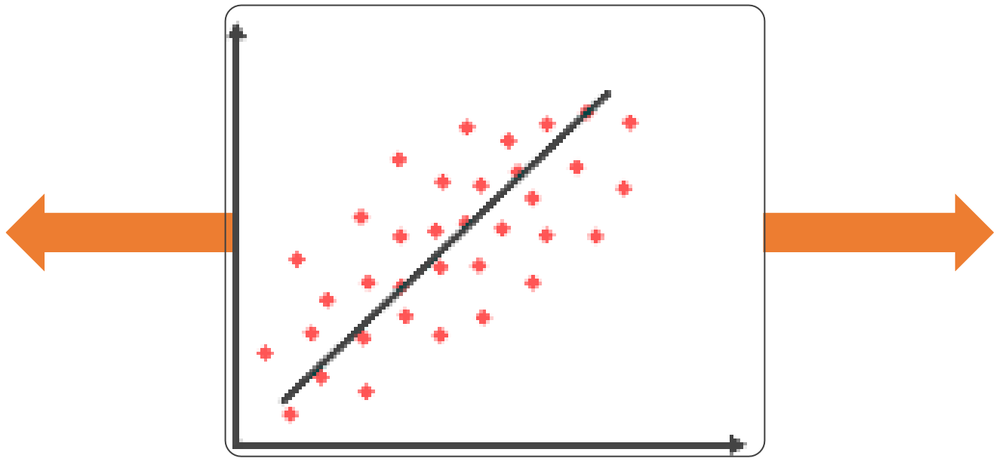
- What are the types of correlation?
- What does covariance mean?

Types of Correlation Coefficients

Types of Correlation Coefficients

There are two types of correlation coefficients:

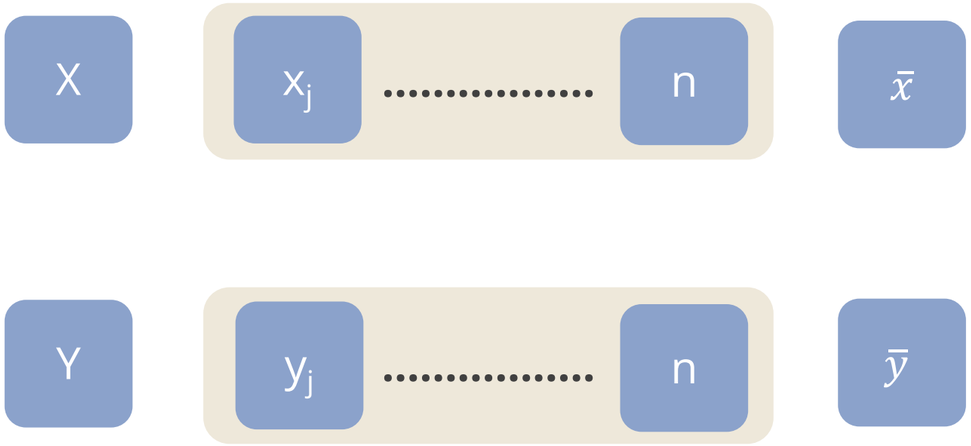
Karl Pearson's Coefficient of Correlation



Spearman's Rank Correlation Coefficient

Karl Pearson’s Coefficient of Correlation

It is a linear correlation coefficient that falls in the value range of -1 to +1.



In a data set comprising n pairs of observations, the x_j values correspond to one characteristic (X), while the y_j values correspond to another characteristic (Y).

Karl Pearson's Coefficient of Correlation

Let $\bar{x}$ and $\bar{y}$ respectively be the means of the X and Y. Then:

Population covariance of X and Y

$$Cov(X,Y) = \frac{\{\sum (x_j - \bar{x}) * (y_j - \bar{y})\}}{n}$$

The summation extends from 1 to n.

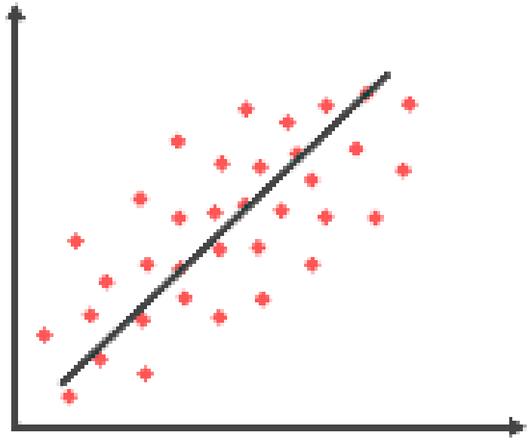
Sample covariance of X and Y

$$Cov(X,Y) = \frac{\{\sum (x_j - \bar{x}) * (y_j - \bar{y})\}}{n - 1}$$

The summation extends from 1 to n – 1.

Correlation Coefficient

Below is the Karl Pearson's Correlation Coefficient:



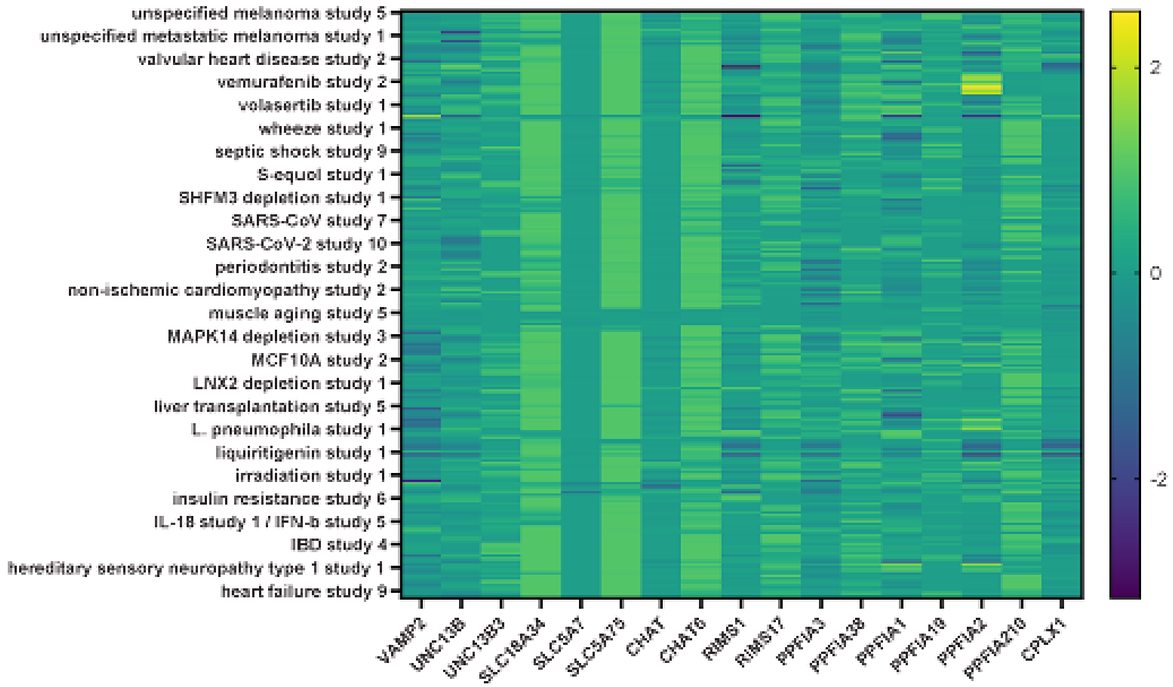
$$r = \frac{\mathrm{cov}(X,Y)}{s_{x^*}s_y}$$

s_x = standard deviation of X

s_y = standard deviation of Y

Correlation Coefficient

The correlation coefficient, often represented by the symbol r, is a statistical measure that calculates the strength and direction of the linear relationship between two variables.



It is important to be cautious and guard against false correlations.

Algebraic Formula of Correlation

The correlation simplifies as:

$$r = \left\{ \frac{(n*(\sum x_j * y_j)) - ((\sum x_j) * (\sum y_j))}{\sqrt{(n*\sum x_j^2 - (\sum x_j)^2)(n*\sum y_j^2 - (\sum y_j)^2)}} \right\}$$

Where,
n = total number of paired data points of x and y
Σ = sum of the values
Σx = sum of all x-values
Σy = sum of all y-values
Σxy = sum of the product of paired x and y values
Σx^2 and Σy^2 = sums of the squares of all x-values and y-values

Discussion: Correlation and Covariance

Duration: 15 minutes

- What are the types of correlation?
  Answer: The two types of correlation coefficients are Karl Pearson's Coefficient of Correlation and Spearman's Rank Correlation Coefficient.

- What does covariance mean?
  Answer: Covariance is a measure of interdependence between two variables.

Karl Pearson’s Correlation Coefficient: Use Cases

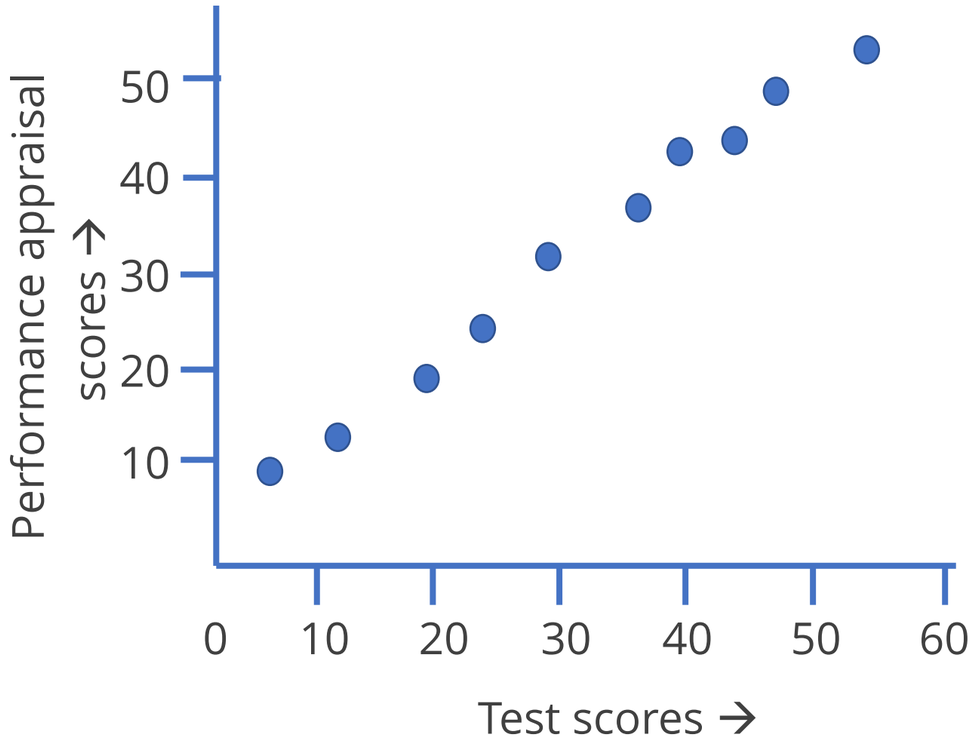
Practical Uses of Karl Pearson’s Coefficient of Correlation

Example 1: Selection of job applicants

The test scores are used to identify candidates for specific posts in an organization.

Practical Uses of Karl Pearson’s Coefficient of Correlation

Assume that individuals' test scores are highly correlated with their scores during later performance appraisals



Well-designed tests prove to be highly effective in personnel selection.

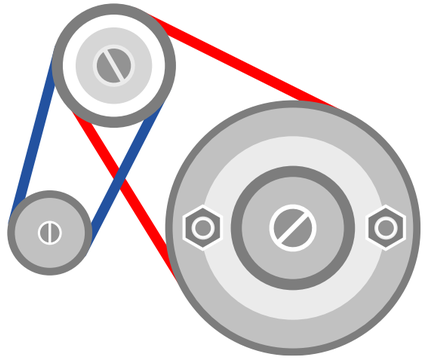
If there is a lack of correlation or a low correlation, it indicates the need to redesign the test.

Practical Uses of Karl Pearson's Coefficient of Correlation

Example 2: The sales in units for certain products are correlated with the demand for spare parts for these products.

Sales in units

Correlated



Demand for spares

Practical Uses of Karl Pearson's Coefficient of Correlation

Knowledge of correlation is useful:

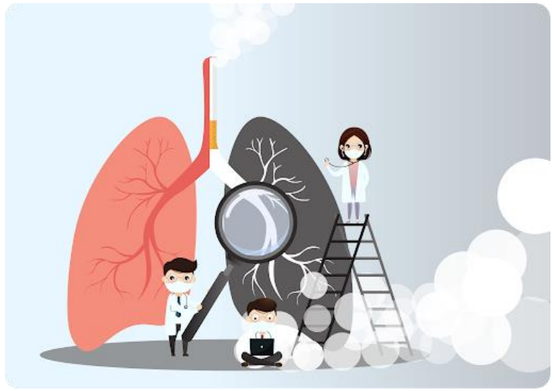
Manufacturing

To make plans for spare parts production in the future

Banking

To find the correlation between income and credit card delinquency rate

Research



To find the correlation between cigarette smoking and longevity

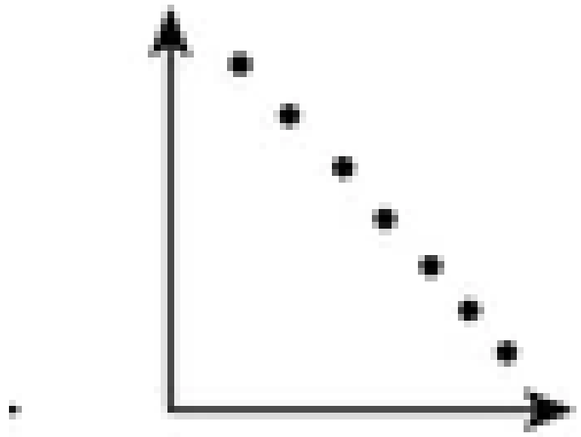
Properties of Pearson’s Correlation Coefficient

The measure is dimensionless.

For instance, the temperature remains constant irrespective of the chosen unit of measurement, demonstrating its inherent value consistency.

Properties of Pearson's Correlation Coefficient

When the correlation coefficient (r) assumes any of the extreme values, it indicates a perfect linear relationship between the two variables.



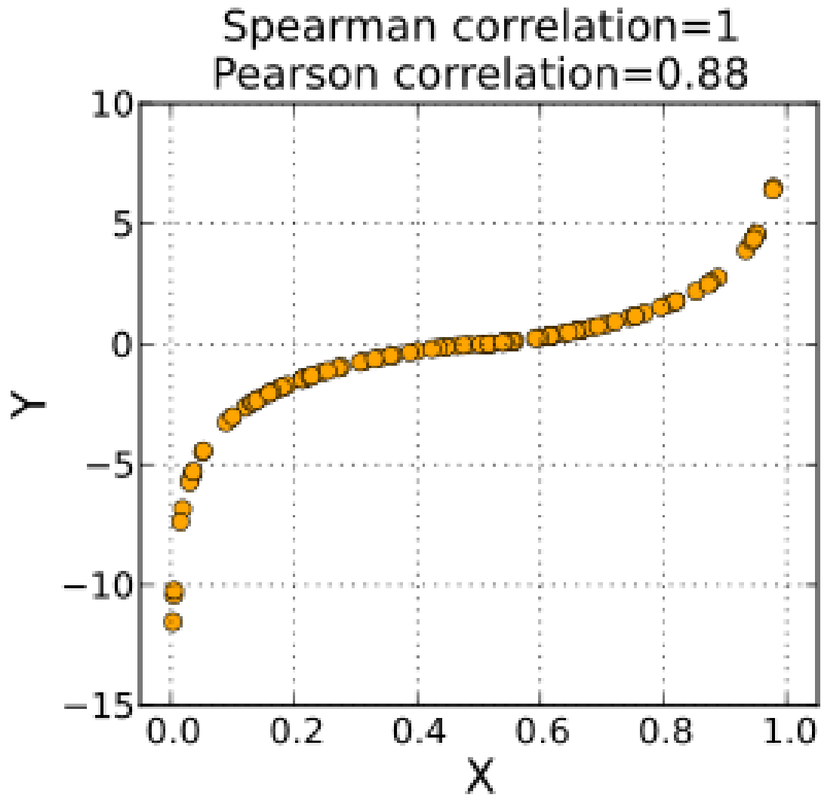
Perfect linear relationship

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

Spearman’s Rank Correlation Coefficient

Spearman’s Rank Correlation Coefficient

It is a nonparametric measure of rank correlation that evaluates the correlation between the rankings of two variables.



It helps determine how well the relationship between the variables can be described using a monotonic function.

Spearman’s Rank Correlation Coefficient

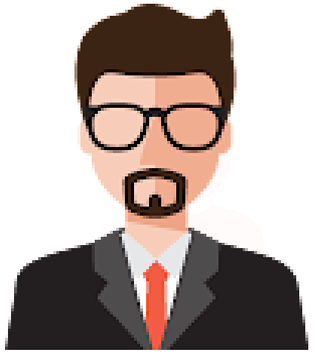
To determine the correlation coefficient, one must examine a data set of two variables representing student scores:

	Examiners	
	X	**Y**
Student 1	0	0
Student 2	4	2
Student 3	7	3
Student 4	10	10

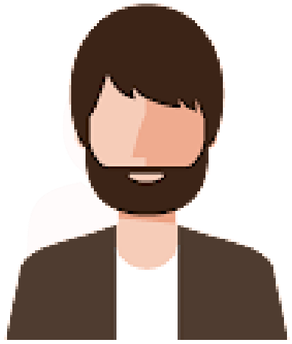
The scores of four students in a test are based on the independent assessments conducted by examiners X and Y.

Spearman’s Rank Correlation Coefficient

Such scores could arise from variations in grading strictness.



Examiner X



Examiner Y

Spearman’s Rank Correlation Coefficient

For example, examiner Y penalizes students more severely for wrong answers.

	Examiners	
	X	Y
Student 1	0	0
Student 2	4	2
Student 3	7	3
Student 4	10	10

For instance, examiner Y applies stricter penalties for incorrect answers. As a result, the correlation between the two examiners is less than 1, with a calculated value of 0.9.

This indicates a high level of agreement between the examiners regarding the relative performance of the candidates.

Spearman’s Rank Correlation Coefficient

To address scenarios like the previous example, Spearman introduced a measure known as the rank correlation coefficient.

Spearman’s Rank Correlation Coefficient

Example 1: Assign ranks to students incorporating the hierarchy in the scores independently for both examiners.

	Graded by X	Rank
Student 1	0	1
Student 2	4	2
Student 3	7	3
Student 4	10	4

	Graded by Y	Rank
Student 1	0	1
Student 2	2	2
Student 3	3	3
Student 4	10	4

Spearman’s Rank Correlation Coefficient

Use the ascending or descending orders consistently in both cases

	X	Rank
Student 1	0	**4**
Student 2	4	**3**
Student 3	7	**2**
Student 4	10	**1**

	Y	Rank
Student 1	0	**4**
Student 2	2	**3**
Student 3	3	**2**
Student 4	10	**1**

The correlation coefficient of the two sets of ranks is referred to as the rank correlation coefficient. This value can be calculated using the same formula.

Formula for Rank Correlation Coefficient

The rank coefficient (r) can be calculated using the following formula:

$$r = 1 - \left\{ \frac{(6*\sum d_j^2)}{[n*(n^2-1)]} \right\}$$

n = number of candidates or pairs of observations

d_j = difference in ranks j^{th} candidate

Calculating Rank Correlation Coefficient

Example 2: Calculating the difference in ranks with coefficient as 1

$$1 = 1 - \left\{ \frac{(6*\sum d_j^2)}{[n*(n^2-1)]} \right\}$$

$$\left\{ \frac{(6*\sum d_j^2)}{[n*(n^2-1)]} \right\} = 0 \longrightarrow \sum d_j^2 = 0$$

Ranks are identical.

Calculating Rank Correlation Coefficient

When the rankings are diametrically opposite, the rank correlation is -1.

Rank 1	Rank 2
1	4
2	3
3	2
4	1

Rank correlation = -1

Discussion

Discussion: Spurious Correlation

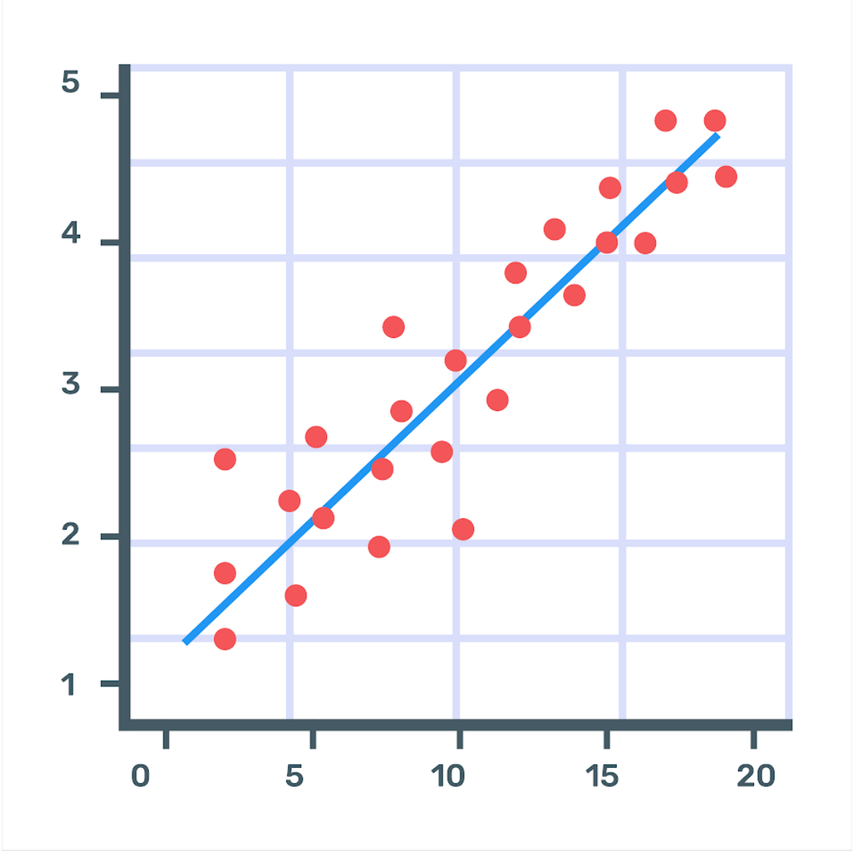
Duration: 15 minutes

- What does spurious correlation mean?
- How is the coefficient of determination utilized?

Causation

Cause and Effect

In a relationship between two variables, one variable serves as the cause, and the other variable as the effect.



- The cause or independent variable is denoted by X.
- The effect or dependent variable is denoted by Y.

Cause and Effect Variables

Example 1: *Cutting speed* is a cause, and its *impact on tool life* is the effect.

Correlation Coefficient

Cause

Effect

Correlation and Causation

Correlation may not always indicate a cause-effect relationship between variables.

Correlation and Causation

Example 2: The total number of students enrolled in schools across different cities may display a correlation when observed over multiple years.

City 1

City 2

However, this does not necessarily imply a cause-effect relationship.

Spurious Correlation

An increase in the number of students in City 1 and City 2 typically stems from a rise in population in both the cities.

Spurious correlation

Increasing students

Increasing population

Such correlations, which may seem related but are not directly causal, are known as spurious correlations.

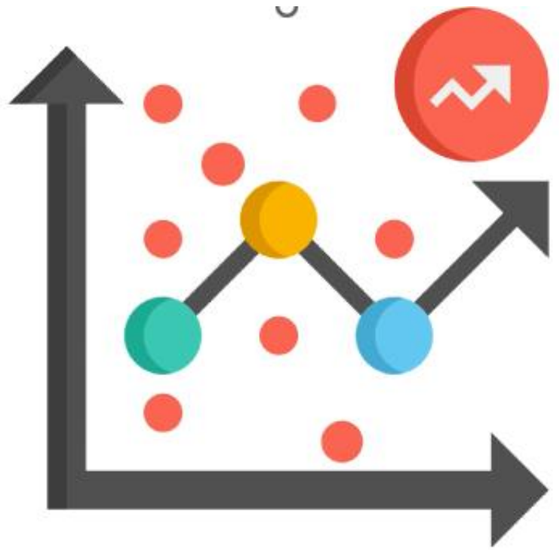
Interpretation of Correlation

Interpreting correlations requires careful consideration.

The action wherein one variable directly impacts another, creating an effect, is known as causation.

Regression

If one needs to study a cause-effect relationship, a predictive tool known as a regression equation is often utilized.



In this context, only linear relationships are considered.

Regression

A data set is composed of n pairs of observations on two variables.

X_i

Independent variable

Y_i

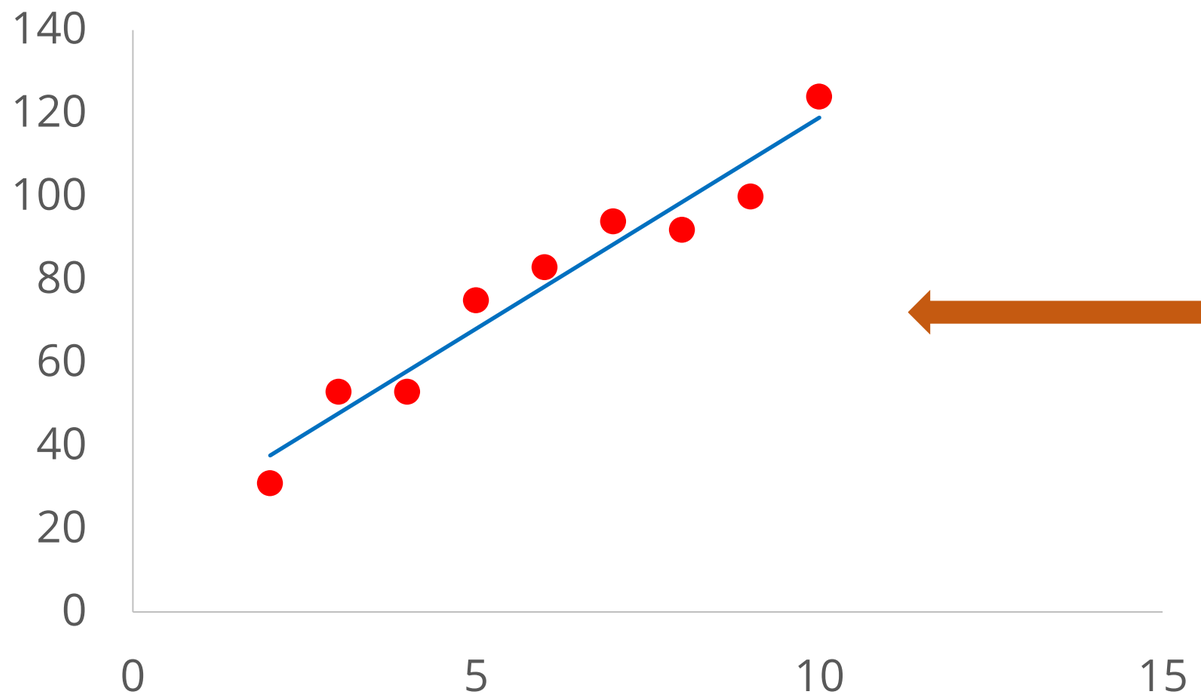
Dependent variable

To predict Y for any given value of X, use a linear equation known as the regression of Y on X

Example of Regression

Consider the given data set and the scatter diagram, as shown:

X	2	3	4	5	6	7	8	9	10
Y	31	53	53	75	83	94	92	100	124



The red points displayed in the graphic representation correspond to the data plots derived from the table.

Example of Regression

Then, the computed values are:

Mean ($\bar{x}$) = $\frac{\Sigma X}{n}$ → $\bar{x}$ = 6 | $\bar{y}$ = 78.3

Standard deviation, (σ) = $\frac{\Sigma(X-\bar{x})^2}{(n-1)}$ → S_x = 2.58 | S_y = 26.97

Correlation = $\epsilon \frac{(X-\bar{x})(Y-\overline{y})}{\sigma x \sigma y}$ → r = 0.97

Prediction Value

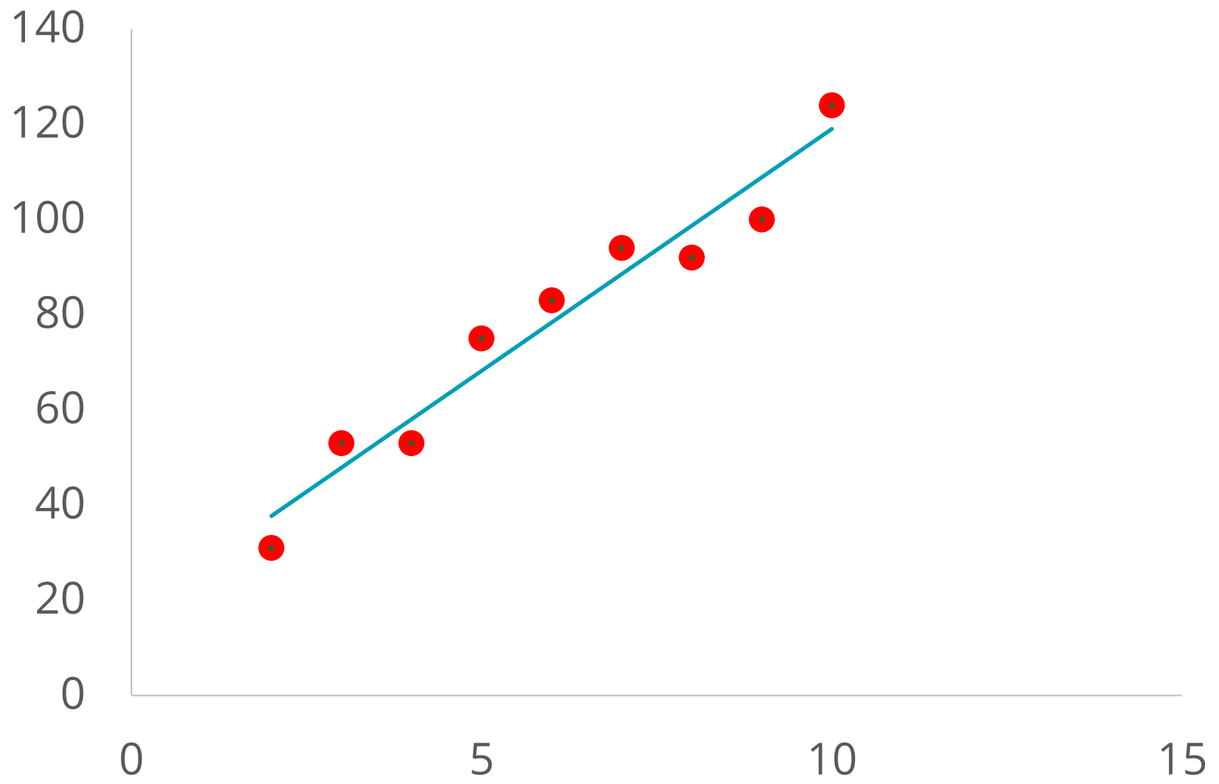
To determine the prediction value, consider the following:

$Est\ (Y_j)$	The predicted value of Y for a given value X_j.
$Est\ (Y_j) = a+(b * X_j)$	Regression of Y on X
$Y_j - Est\ (Y_j) = e_j$	Error in estimation
$Y_j = a +(b*X_j)+ e_j$	ej is the difference between the actual value and the predicted value.

Regression Line

A regression line, also known as a line of best fit, is a straight line that is used to visualize and quantify the correlation between two variables in statistical analysis.

The blue line shown below is a regression line.



Prediction Value

The prediction value is calculated as:

$$\text{Est (Y)} - \bar{y} = r^* (s_y/s_x) * (X - \bar{x})$$

r	Correlation
s_y/s_x	The ratio of the standard deviation of Y to X
$X - \bar{x}$	The difference between X and $\bar{x}$

Thus, the predicted value of Y can be obtained for any value of X.

Calculating Prediction Value

Consider the same data set:

X	2	3	4	5	6	7	8	9	10
Y	31	53	53	75	83	94	92	100	124

Recalling the calculations previously performed:

$\bar{x} = 6$ $\bar{y} = 78.3$ $S_x = 2.58$ $S_y = 26.97$ $r = 0.97$

Calculating Prediction Value

The required regression equation is as shown:

Est (Y) - 78.3 = (0.97 * 26.97/2.58) * (X - 6) = 10.17 * (X - 6)

X = 7 → Est (Y) - 78.3 =10.14

Predicted value Y = 88.44

Coefficient of Determination

Coefficient of Determination

It is a statistical test that determines how well changes in one variable can account for variations in another.

After obtaining this equation, an index for estimating its quality becomes necessary.

The coefficient of determination fulfills this role as a quality estimation index.

Source: Investopedia

Coefficient of Determination: Example

Example: Effect of cutting speed (X) on tool life (Y).

Cutting speed	X
Tool life	Y

Values of tool life exhibit variation.

Coefficient of Determination: Example

$Y - \bar{y}$ is the deviation or variation

$Y - \bar{y} = Y - Est(Y) + Est(Y) - \bar{y}$

$Y - Est(Y)$ → An error term highlights the difference between the observed value and the predicted value, which is the value of the deviation.

$Est(Y) - \bar{y}$ → This deviation represents the discrepancy of the estimated value from the average.

As cutting speeds change, predicted values will differ from the average tool life.

Quantifying Quality

When the quality of prediction is good, the error terms are small.

Hence, the above two values will tend to be close to each other.

Quantifying Quality

The following measure, R^2, is utilized to quantify the quality of the regression equation.

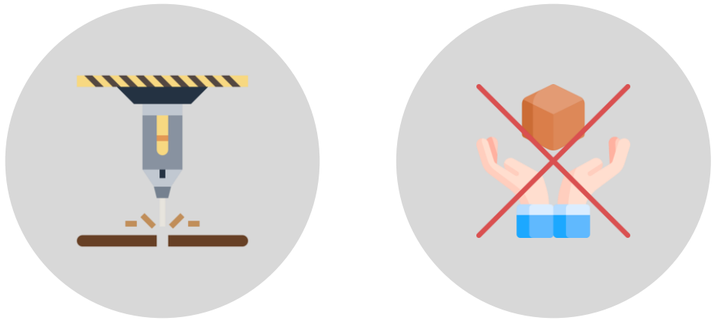
$$R^2 = \{ \sum(Est\ (y_j) - \bar{y}) \sum/\{^2(y_j - \bar{y}^2($$

R^2 is the coefficient of determination.

- It quantifies the degree to which variations in the dependent variable are explained by the regression equation.
- If R^2 is low, the explanatory power of the regression equation might be low.

Variations in Prediction Value

Y may likely be influenced by other factors, in addition to X.

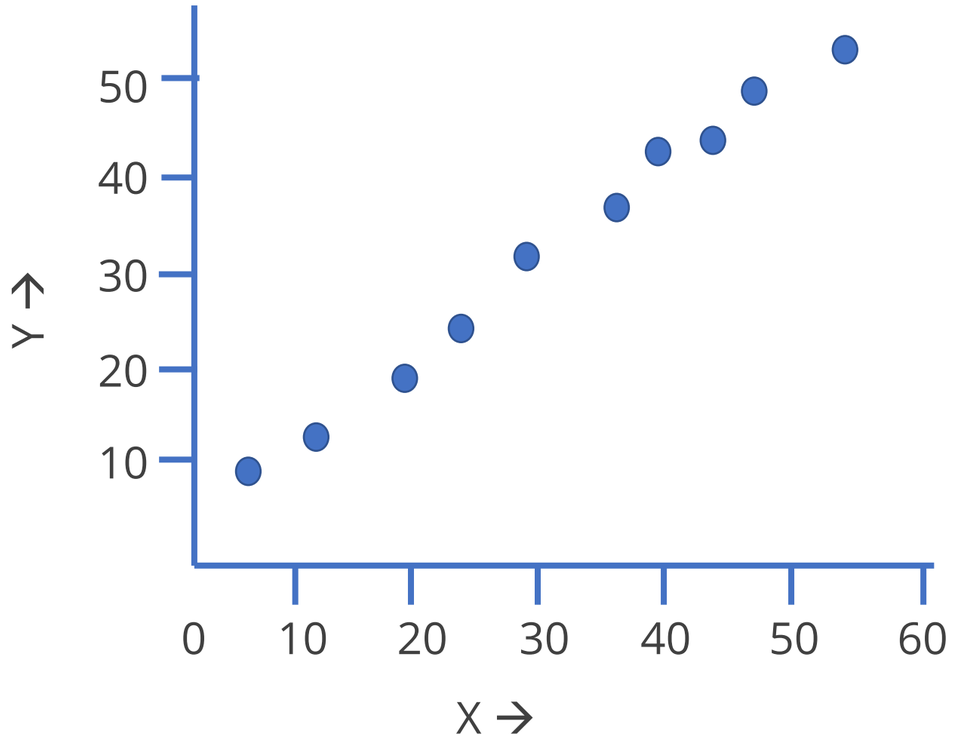


For instance, the cutting speed could be impacted by factors such as improper handling.

Total variation is calculated as follows:
Total variation = Explained variation + Unexplained variation

Coefficient of Determination and Variations

When variables are highly correlated and the regression equation delivers a high-quality prediction, the unexplained variation tends to be low.



Quality of prediction is measured by:

$$\text{Coefficient of determination} = \frac{\text{Explained variation}}{\text{Total variation}}$$

Coefficient of Determination and Variations

The value of the coefficient of determination is non-negative and does not exceed unity.

When r = + or - 1:

100% of the variation is explained, that is, the regression equation accounts for all changes in the dependent variable.

When r = + or – 0.9:

81% of the variations are explained by the regression equation, the remaining 19% is unexplained variation.

TSS and RSS

Difference between RSS and TSS:

Total sum of squares (TSS)

- Measures variation in the observed data
- TSS = $\sum_{i=1}^{n}(yi - \bar{y})^2$

Residual sum of squares (RSS)

- Measures variation in the error between the observed data and modeled values
- RSS = $(y_1 - y_{pred})^2 + (y_2 - y_{pred})^2 + \ldots + (y_n - y_{pred})^2$

Coefficient of Determination: Example

The table below presents the calculated coefficient of determination for the given data:

X	Y	Est (Y)	Y - $\bar{y}$	Est (Y) - $\bar{y}$	TSS	RSS
2	1	37.6665	-47.3333	-40.6668	2240.444	1653.791
3	53	47.8332	-25.3333	-30.5001	641.7776	930.2579
4	53	57.9999	-25.3333	-20.3334	641.7776	413.4484
5	75	68.1666	-3.33333	-10.1667	11.11109	103.3624
6	83	78.3333	4.66667	-3.00e-05	21.77781	9.00E-10
7	94	88.5	15.66667	10.16667	245.4445	103.3612
8	92	98.6667	13.66667	20.33337	186.7779	413.4459
9	100	108.8334	21.66667	30.50007	469.4446	930.2543
10	124	119.0001	45.66667	40.66677	2085.445	1653.786

Coefficient of Determination: Example

From the dataset in the previous slide:

Sum of TSS	=	6544
Sum of RSS	=	6201.707

The coefficient of determination is calculated using the formula = 1 - (RSS/TSS).

Coefficient of Determination: Example

This indicates that 95% of the variations in Y are explained by the regression equation.

Coefficient of determination	=	1- (6201.707/6544)	=	1- 0.9476
			=	0.0523

Discussion: Spurious Correlation

Duration: 15 minutes

- What does spurious correlation mean?
  Answer: A spurious correlation is a term used in statistics to describe a situation where two variables seem to be related to each other, but in reality, there is no causal relationship between them.

- How is the coefficient of determination utilized?
  Answer: The coefficient of determination is used as an index to estimate the quality of the regression equation.

Key Takeaways

- When two variables are related, the relationship can be studied under a concept called correlation.
- The correlation coefficient quantifies the extent to which the variations in one variable influences the variations in the other.
- Causation helps to understand the relationship between two variables. One is the cause and the other is the effect.

Knowledge Check

Knowledge Check 1

_____ is used to analyze the correlation between two variables.

A. Scatter plot

B. Bar graph

C. Pie chart

D. Bubble chart

Knowledge Check 1

_____ is used to analyze the correlation between two variables.

A. Scatter plot

B. Bar graph

C. Pie chart

D. Bubble chart

The correct answer is **A**

A scatter plot is used to analyze the correlation between two variables.

Knowledge Check 2

The formula to find Karl Pearson's Correlation Coefficient is _____.

A. $r = Cov(X, Y)/ (sx* sy)2$

B. $r = Cov(X, Y)/ \sqrt{sx* sy}$

C. $r = Cov(X, Y)/ sx2* sy2$

D. $r = Cov(X, Y)/ sx* sy$

Knowledge Check 2

The formula to find Karl Pearson's Correlation Coefficient is _____.

A. r = Cov(X, Y)/ (sx* sy)2

B. r = Cov(X, Y)/ $\sqrt{sx* sy}$

C. r = Cov(X, Y)/ sx2* sy2

D. r = Cov(X, Y)/ sx* sy

The correct answer is **D**

The formula to find Karl Pearson's Correlation Coefficient is r = Cov(X, Y)/ sx* sy.

Knowledge Check 3

In Spearman's Rank Correlation Coefficient, if rankings are diametrically opposite, the correlation between the rankings is _____.

A. 0

B. -1

C. 0.5

D. +1

Knowledge Check 3

In Spearman’s Rank Correlation Coefficient, if rankings are diametrically opposite, the correlation between the rankings is _____.

A. 0

B. -1

C. 0.5

D. +1

The correct answer is **B**

In Spearman’s Rank Correlation Coefficient, if rankings are diametrically opposite, the correlation between the rankings is -1.

Thank You