In the world of data science, understanding the relationship between variables is crucial for making informed decisions or building accurate machine learning models. Correlation is a fundamental statistical concept that measures the strength and direction of the relationship between two variables. However, without the right tools and knowledge, calculating correlation coefficients and p-values can be a daunting task for data scientists. This can lead to suboptimal decision-making, inaccurate predictions, and wasted time and resources. In this post, we will discuss what Pearson's r represents, how it works mathematically (formula), its interpretation, statistical significance, and importance for making decisions in real-world applications such as business forecasting or medical diagnosis. We will also explore some examples of using Pearson's r (correlation coefficient) and p-value (used for statistical significance) with real data sets so you can see how this powerful statistic works in action. We will learn to use Python's scipy.stats pearsonr method which is a simple and effective way to calculate the correlation coefficient and p-value between two variables. As a data scientist, it is very important to understand Pearson's r and its implications for making decisions based on data. What is Pearson Correlation Coefficient? Pearson correlation coefficient is a statistical measure that describes the linear relationship between two variables. It is typically represented by the symbol 'r'. Pearson correlation coefficient can take on values from -1 to +1 and it is used to determine how closely two variables are related. It measures the strength of their linear relationship, which means that it indicates whether one variable increases or decreases as the other variable increases or decreases. A Pearson correlation coefficient of 1 indicates a perfect positive (direct) linear relationship, while a Pearson correlation coefficient of -1 indicates a perfect negative (inverse) linear relationship. Furthermore, when Pearson's r is 0 there is no linear relationship between the two variables. It's important to note that correlation does not imply causation. A significant Pearson's r value indicates a linear association, but it doesn't mean that one variable causes the other. Other factors, known as confounding variables, may influence this relationship. Additionally, Pearson's r only measures linear relationships. If the relationship is non-linear, other statistical methods may be more appropriate to describe the association. A study finds a significant positive Pearson correlation coefficient (r) between monthly ice cream sales and the number of drowning incidents. The data show that as ice cream sales increase, the number of drowning incidents also increases. If we mistakenly infer causation from this correlation, we might conclude that eating ice cream leads to an increased risk of drowning. The increase in both ice cream sales and drowning incidents might both be caused by a third variable (confounding variable): the temperature or season (i.e., summer). During summer months, temperatures are higher, which leads to more people buying ice cream. Simultaneously, more people are likely to engage in swimming activities, which increases the risk of drowning incidents. Temperature acts as a confounding variable that is associated with both ice cream sales and drowning incidents. Pearson Correlation Coefficient vs Plots The following plots represent linear relationship vis-a-vis different values of Pearson correlation coefficient. The following is the explanation for the above plots: Direct Linear Relationship (r close to +1): The first plot shows a clear upward trend, indicating that as x increases, y also increases. The points are closely aligned around a straight line, suggesting a strong positive linear relationship. The Pearson Correlation Coefficient for such a dataset would be close to +1, implying that the variables move together in the same direction. No Linear Relationship (r close to 0): The second plot shows a scatter of points with no apparent pattern. There is no discernible slope, and the points do not align around any line. This randomness suggests that there is no linear relationship between x and y. In such a case, the Pearson Correlation Coefficient would be close to 0, indicating no linear correlation between the variables. Inverse Linear Relationship (r close to -1): The third plot shows a clear downward trend, indicating that as x increases, y decreases. The points are closely aligned around a straight line, but this time the line slopes downwards, suggesting a strong negative linear relationship. The Pearson Correlation Coefficient for such a dataset would be close to -1, implying that the variables move in opposite directions. Pearson Correlation Coefficient – Different Values vs Strength of Relationship When assessing the linear relationship between two variables using correlation analysis, the magnitude of the correlation coefficient (ignoring the sign) provides insight into the strength of the relationship. Here's a more detailed guide to interpreting the absolute value of the correlation coefficient: ±1.00: This represents a perfect correlation, indicating that for every change in one variable, there is a predictable and exact corresponding change in the other variable. In a graph, the data points would lie exactly on a straight line, either upwards or downwards, depending on the sign. ±0.80: When the correlation coefficient approaches this value, it is considered a strong correlation. This suggests a high degree of predictability in the relationship, where changes in one variable are closely followed by changes in the other, though not perfectly. ±0.50: This value signifies a moderate correlation. The relationship between the variables is evident and can be described as substantial, but there are other factors and variability influencing the relationship. ±0.20: This is indicative of a weak correlation, where there is a slight, possibly inconsistent association between the variables. The predictability is low, and while there may be a relationship, it is not strong and could be easily influenced by other variables. 0: A zero or close to zero correlation coefficient means there is no linear correlation between the variables. There's no predictable association that can be discerned from the data; any relationship is likely due to chance or randomness. Pearson Correlation Coefficient – Real-world Examples Pearson correlation coefficient can be used to examine relationships between variables in a variety of real-world applications such as some of the following: In medicine, Pearson's r can be used to measure the strength of the relationship between patient age and cholesterol levels. In finance, Pearson's r can be used to measure the strength of the relationship between stock prices and earnings per share. In business forecasting, Pearson's r can be used to measure the strength of the relationship between sales and marketing efforts. In lifestyle research, Pearson's r can be used to measure the strength of the relationship between exercise habits and obesity rates. Another example is measuring correlation between customer loyalty against customer satisfaction levels and ascertain whether customers who report higher levels of satisfaction also demonstrate higher levels of loyalty or vice versa. Another example could include studying height against weight wherein one might use Pearson's correlation coefficient to measure if taller individuals tend to weigh more than their shorter counterparts on average or if there is no obvious connection present at all between height and weight when considering real-world data sets. Pearson's correlation coefficient has implications for hypothesis testing as well as other decision-making processes. By measuring the strength of a linear relationship between two variables, researchers can make informed decisions based on their findings which can help guide future research studies or inform corporate policies and practices. Pearson's correlation coefficient also provides a basis for making predictions about future outcomes when given certain inputs or conditions–which is incredibly valuable in various business settings where predicting customer behavior or market trends is critical for success. Pearson Correlation Coefficient – Formula The Pearson Correlation Coefficient formula is given as the following: Pearson Correlation Coefficients should not be taken as definitive proof that there is a relationship between two variables; rather they should only serve as indicators for further investigation which can then lead to more conclusive results regarding such relationships. In addition, Pearson Correlation Coefficients are considered reliable only when sample sizes are large enough and data points are normally distributed; if these conditions are not met then other statistical tests may be necessary in order to determine the significance of any indicated correlations. Scatterplots & Pearson Correlation Coefficient Scatterplots are a powerful way of visualizing data and relationships between two variables. They are graphs that display data points in which the values for two variables are plotted against each other. The x-axis usually displays one variable, and the y-axis displays the other variable. Each point on a scatter plot represents one data set composed of the independent and dependent variables being studied; when plotted in relation to each other, these points form clusters or patterns which allow us to analyze the strength and direction of the relationship between these variables. When plotting scatter plots, Pearson's correlation coefficient can be used to determine how closely related two variables are to each other by measuring the degree of association between them. Pearson's correlation coefficient is calculated using the formula: $r = \sum(x - \bar{x})(y - \bar{y}) / \sqrt{\sum(x - \bar{x})^2 \sum (y - \bar{y})^2}$ where $\bar{x}$ and $\bar{y}$ represent mean values for the respective x and y values. By examining how closely points cluster together on a scatter plot, one can measure both linearity and strength in order to determine Pearson's correlation coefficient value. The picture below represents correlation coefficient in three different scatter plots. The picture below might represent a very high correlation coefficient closer to 1. The picture below might represent decently high correlation coefficient closer to 0.5. The picture below might represent a very low correlation coefficient closer to 0. Pearson Correlation Coefficient Examples This example will illustrate how to use Pearson correlation coefficient (PCC) to determine the correlation between two continuous variables. In the example below, the marks of mathematics and science for a class of students in a school are considered for evaluating correlation. Based on the value of the PCC, data scientists can identify linear relationships between these two variables, providing invaluable insights about the data. Note the usage of PCC formula defined in earlier section. Statistical Significance of Pearson Correlation Coefficient In order to determine whether any given Pearson correlation coefficient has a statistically significant result or not, we will need to go through the following steps: Determine null & alternate hypothesis: The null hypothesis can be stated that there is no relationship between the two variables (r = 0) while the alternate hypothesis is that there is a relationship (r != 0). Determine statistics for hypothesis testing: We will calculate t-statistics and perform t-test with (n-2) degree of freedom Determine level of significance: The level of significance chosen is 0.05 Calculate & compare t-statistics with critical value: We will test the significance by evaluating t-statistics and comparing it with critical value read from t-distribution table at 0.05 significance level. If the t-statistics value is greater than the critical value at 0.05, the null hypothesis can be rejected. This would mean that there is enough evidence to support the alternate hypothesis that there is some relationship between two variables. The following is the formula for calculating the value of t-statistics for determining statistical significance of Pearson correlation coefficient: In the above formula, r is correlation coefficient value and n is sample size. In the example given in earlier section, the t-value will come out to be based on the following calculation. The value of n = 7 and value of r = 0.724. $$ t = \frac{0.724\sqrt(7-2)}{\sqrt(1 – 0.724^2)} $$ $$t = 2.347 $$ One can also calculate p-value and compare it with 0.05 significance level. If p-value is less than 0.05, the Pearson correlation coefficient can be considered to be statistically significant and the null hypothesis rejected in favor of the alternate hypothesis. The degree of freedom = (N-2). In above example, degree of freedom = (7-2) = 5. Thus, p-value by looking at the t-distribution table for a two-tailed test with t-value as 2.347, df = 5 comes out to be 0.0658. Thus, at a 0.05 level of significance, we don't have enough evidence to reject the null hypothesis (no relationship between the marks). Thus, based on given evidence, one can conclude that there is a statistical evidence that the linear variables such as marks of maths and science are not strongly correlated. Recall that a P-value is a statistic that tells us how likely an observed result is due to chance alone. If a Pearson Correlation Coefficient has an associated P-value below 0.05, then it can be considered statistically significant. This means that the Pearson Correlation Coefficient is unlikely to have occurred by chance and thus supports the hypothesis that there is indeed some kind of relationship between the two variables being studied. Here is the Python code for calculating the correlation coefficient and p-value. The data used in the code below is NPX & PeptideAbundance which can be accessed from Kaggle competition (Parkinson disease prediction) . NPX (Normalized protein expression) is the frequency of the protein's occurrence in the sample. PeptideAbundance is the frequency of the amino acid in the sample import pandas as pd import numpy as np from scipy.stats import pearsonr # Read in data from CSV file df = pd.read_csv('data.csv') # Drop rows with NaN or Inf values df = df.replace([np.inf, -np.inf], np.nan).dropna() # Calculate correlation coefficient corr, pval = pearsonr(df['NPX'], df['PeptideAbundance']) # Print results if pval < 0.05: print(f"The correlation coefficient between NPX and PeptideAbundance is {corr:.3f}, which is significant (p = {pval:.3f}).") else: print(f"The correlation coefficient between NPX and PeptideAbundance is {corr:.3f}, which is not significant (p = {pval:.3f}).") If the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant correlation between the two variables. Otherwise, we fail to reject the null hypothesis, and we conclude that there is no significant correlation. In the above code, the correlation coefficient and the p-value is printed. If the p-value is less than 0.05, it is printed that the correlation is significant; otherwise, it gets printed that the correlation is not significant. In conclusion, Pearson correlation coefficient is a powerful tool that allows for measuring the strength of linear relationships between two variables. It has implications for decision-making processes and research studies as well as real-world applications such as medicine, finance, business forecasting and lifestyle research. Pearson's correlation coefficient also provides statistical significance testing which helps researchers make informed decisions based on their findings. Thus, Pearson correlation coefficient is an invaluable resource to have when conducting any form of quantitative analysis or data exploration. I have been recently working in the area of Data analytics including Data Science and Machine Learning / Deep Learning. I am also passionate about different technologies including programming languages such as Java/JEE, Javascript, Python, R, Julia, etc, and technologies such as Blockchain, mobile computing, cloud-native technologies, application security, cloud computing platforms, big data, etc. I would love to connect with you on Linkedin. Check out my latest book titled as First Principles Thinking: Building winning products using first principles thinking. Medical example data Marketing example data Pearson correlation analysis examines the relationship between two variables. For example, is there a correlation between a person's age and salary? More specifically, we can use the pearson correlation coefficient to measure the linear relationship between two variables. Strength and direction of correlation With a correlation analysis we can determine: How strong the correlation is and in which direction the correlation goes. We can read the strength and direction of the correlation in the Pearson correlation coefficient r, whose value varies between -1 and 1. Strength of the correlation The strength of the correlation can be read in a table. An r between 0 and 0.1 indicates no correlation. An amount of r between 0.7 and 1 is indicates a very strong correlation. Amount of r Strength of correlation 0.0 < 0.1 no correlation 0.1 < 0.3 low correlation 0.3 < 0.5 medium correlation 0.5 < 0.7 high correlation 0.7 < 1 very high correlation From Kuckartz et al.: Statistik, Eine verständliche Einführung, 2013, p. 213 Direction of the correlation A positive relationship or correlation exists when large values of one variable are associated with large values of the other variable, or when small values of one variable are associated with small values of the other variable. A positive correlation results, for example, for height and shoe size. This results in a positive correlation coefficient. A negative correlation is when large values of one variable are associated with small values of the other variable and vice versa. A negative correlation is usually found between product price and sales volume. This results in a negative correlation coefficient. Calculate Pearson correlation The Pearson correlation coefficient is calculated using the following equation. Here r is the Pearson correlation coefficient, xi are the individual values of one variable e.g. age, yi are the individual values of the other variable e.g. salary and $\bar{x}$ and $\bar{y}$ are the mean values of the two variables respectively. In the equation, we can see that the respective mean value is first subtracted from both variables. So in our example, we calculate the mean values of age and salary. We then subtract the mean values from each of age and salary. We then multiply both values. We then sum up the individual results of the multiplication. The expression in the denominator ensures that the correlation coefficient is scaled between -1 and 1. If we now multiply two positive values we get a positive value. If we multiply two negative values we also get a positive value (minus times minus is plus). So all values that lie in these ranges have a positive influence on the correlation coefficient. If we multiply a positive value and a negative value we get a negative value (minus times plus is minus). So all values that are in these ranges have a negative influence on the correlation coefficient. Therefore, if our values are predominantly in the two green areas from previous two figures, we get a positive correlation coefficient and therefore a positive correlation. If our scores are predominantly in the two red areas from the figures, we get a negative correlation coefficient and thus a negative correlation. If the points are distributed over all four areas, the positive terms and the negative terms cancel each other out and we might end up with a very small or no correlation. Testing correlation coefficients for significance In general, the correlation coefficient is calculated using data from a sample. In most cases, however, we want to test a hypothesis about the population. In the case of correlation analysis, we then want to know if there is a correlation in the population. For this, we test whether the correlation coefficient in the sample is statistically significantly different from zero. Hypotheses in the Pearson Correlation The null hypothesis and the alternative hypothesis in Pearson correlation are thus: Null hypothesis: The correlation coefficient is not significantly different from zero (There is no linear relationship). Alternative hypothesis: The correlation coefficient deviates significantly from zero (there is a linear correlation). Attention: It is always tested whether the null hypothesis is rejected or not rejected. In our example with the salary and the age of a person, we could thus have the question: Is there a correlation between age and salary in the German population (the population)? To find out, we draw a sample and test whether the correlation coefficient is significantly different from zero in this sample. The null hypothesis is then: There is no correlation between salary and age in the German population. and the alternative hypothesis: There is a correlation between salary and age in the German population. Significance and the t-test Whether the Pearson correlation coefficient is significantly different from zero based on the sample surveyed can be checked using a t-test. Here, r is the correlation coefficient and n is the sample size. A p-value can then be calculated from the test statistic t. If the p-value is smaller than the specified significance level, which is usually 5%, then the null hypothesis is rejected, otherwise it is not. Assumptions of the Pearson correlation But what about the assumptions for a Pearson correlation? Here we have to distinguish whether we just want to calculate the Pearson correlation coefficient, or whether we want to test a hypothesis. To calculate the Pearson correlation coefficient, only two metric variables must be present. Metric variables are, for example, a person's weight,a person's salary or electricity consumption. The Pearson correlation coefficient, then tells us how large the linear relationship is. If there is a non-linear correlation, we cannot read it from the Pearson correlation coefficient. However, if we want to test whether the Pearson correlation coefficient is significantly different from zero in the sample, i.e. we want to test a hypothesis, the two variables must also be normally distributed! If this is not given, the calculated test statistic t or the p-value cannot be interpreted reliably. If the assumptions are not met, Spearman's rank correlation can be used. Calculate Pearson correlation online with DATAtab If you like, you can of course calculate a correlation analysis online with DATAtab. To do this, simply copy your data into this table in the statistics calculator and click on either the Hypothesis tests or Correlation tab. If you now look at two metric variables, a Pearson correlation will be calculated automatically. If you don't know exactly how to interpret the results, you can also just click on Summary in words! many illustrative examples ideal for exams and theses statistics made easy on 454 pages 6th revised edition (March 2025) Only 8.99 € Free sample "It could not be simpler" "So many helpful examples" The first person to give a mathematical formula for the measurement of the degree of relationship between two variables in 1890 was Karl Pearson. 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 is the most popular and is widely used. It is denoted by 'r', where r is a pure number which means that r has no unit. According to Karl Pearson, "Coefficient of Correlation is calculated by dividing the sum of products of deviations from their respective means by their number of pairs and their standard deviations." Formula of Karl Pearson's Coefficient of CorrelationKarl~Pearson's~Coefficient~of~Correlation(r)=\frac{Sum~of~Products~of~Deviations~from~their~respective~means}{Number~of~Pairs\times{Standard~Deviations~of~both~Series}} Or r=\frac{\sum{xy}}{N\times{\sigma_x}\times{\sigma_y}} Where, N = Number of Pair of Observations x = Deviation of X series from Mean (X-\bar{X}) y = Deviation of Y series from Mean (Y-\bar{Y}) \sigma_x = Standard Deviation of X series (\sqrt{\frac{\sum{x^2}}{N}}) \sigma_y = Standard Deviation of Y series (\sqrt{\frac{\sum{y^2}}{N}}) r = Coefficient of Correlation Methods of Calculating Karl Pearson's Coefficient of CorrelationActual Mean MethodDirect MethodShort-Cut Method/Assumed Mean Method/Indirect MethodStep-Deviation Method1. Actual Mean MethodThe steps involved in the calculation of coefficient of correlation by using Actual Mean Method are: The first step is to calculate the mean of the given two series (say X and Y).Now, take the deviation of X series from \bar{X} and denote the deviations by x.Square the deviations of x and obtain the total; i.e., \sum{x^2} Take the deviation of Y series from \bar{Y} and denote the deviations by y.Square the deviations of y and obtain the total; i.e., \sum{y^2} Multiply the respective deviations of Series X and Y and obtain the total; i.e., \sum{xy} .Now, use the following formula to determine the Coefficient of Correlation:r=\frac{\sum{xy}}{\sqrt{\sum{x^2}\times{\sum{y^2}}}} Example:Use Actual Mean Method and determine the coefficient of correlation for the following data: Solution:\bar{X}=\frac{\sum{X}}{N}=\frac{168}{7}=24 \bar{Y}=\frac{\sum{Y}}{N}=\frac{105}{7}=15 r=\frac{\sum{xy}}{\sqrt{\sum{x^2}\times{\sum{y^2}}}} ∑xy = 336, ∑x2 = 448, ∑y2 = 252 r=\frac{336}{\sqrt{448\times252}}=\frac{336}{\sqrt{1,12,896}}=\frac{336}{336}=1 Coefficient of Correlation = 1 It means that there is a perfect positive correlation between the values of Series X and Series Y. 2. Direct MethodThe steps involved in the calculation of coefficient of correlation by using Direct Method are: The first step is to calculate the sum of Series X (∑X).Now, calculate the sum of Series Y (∑Y).Square the values of X Series and calculate their total; i.e., ∑X2.Square the values of Y Series and calculate their total; i.e., ∑Y2.Multiply the values of Series X and Y and calculate their total; i.e., ∑XY.Now, use the following formula to determine Coefficient of Correlation:r=\frac{N\sum{XY}-\sum{X}.\sum{Y}}{\sqrt{N\sum{X^2}-(\sum{X})^2}{\sqrt{N\sum{Y^2}-(\sum{Y})^2}}} Example:Use Direct Method and determine the coefficient of correlation for the following data: Solution:r=\frac{N\sum{XY}-\sum{X}.\sum{Y}}{\sqrt{N\sum{X^2}-(\sum{X})^2}{\sqrt{N\sum{Y^2}-(\sum{Y})^2}}} =\frac{(7\times2,856)-(168\times105)}{\sqrt{(7\times4,480)-(168)^2\times{\sqrt{(7\times1,827)-(105)^2}}} =\frac{19,992-17,640}{\sqrt{31,360-28,224}\times{\sqrt{12,789-11,025}}} =\frac{2,352}{\sqrt{3,136}\times{\sqrt{1,764}}}=\frac{2,352}{56\times42} =\frac{2,352}{2,352}=1 Coefficient of Correlation = 1 It means that there is a perfect positive correlation between the values of Series X and Series Y. 3. Short-Cut Method/Assumed Mean MethodActual Mean can sometimes come in fractions which can make the calculation of standard deviation complicated and difficult. In those cases, it is suggested to use Short-Cut Method to simplify the calculations. The steps involved in the calculation of coefficient of correlation by using Assumed Mean Method are: First of all, take the deviations of X Series from the assumed mean and denote the values by dx. Calculate their total; i.e., ∑dx.Now, square the deviations of X series and calculate their total; i.e., ∑dx2.Take the deviations of Y Series from the assumed mean and denote the values by dy. Calculate their total; i.e., ∑dy.Square the deviations of Y series and calculate their total; i.e., ∑dy2.Multiply dx and dy and calculate their total; i.e., ∑dxdy.Now, use the following formula to determine Coefficient of Correlation:r=\frac{N\sum{dxdy}-\sum{dx}.\sum{dy}}{\sqrt{N\sum{dx^2}-(\sum{dx})^2}{\sqrt{N\sum{dy^2}-(\sum{dy})^2}}} Where, N = Number of pair of observations ∑dx = Sum of deviations of X values from assumed mean ∑dy = Sum of deviations of Y values from assumed mean ∑dx2 = Sum of squared deviations of X values from assumed mean ∑dy2 = Sum of squared deviations of Y values from assumed mean ∑dxdy = Sum of the products of deviations dx and dy Example:Use Assumed Mean Method and determine the coefficient of correlation for the following data: Solution:r=\frac{N\sum{dxdy}-\sum{dx}.\sum{dy}}{\sqrt{N\sum{dx^2}-(\sum{dx})^2}{\sqrt{N\sum{dy^2}-(\sum{dy})^2}}} =\frac{(7\times420)-(28\times21)}{\sqrt{(7\times560)-(28)^2\times{\sqrt{(7\times315)-(21)^2}}} =\frac{2,940-588}{\sqrt{3,920-784}\times{\sqrt{2,205-441}}} =\frac{2,352}{\sqrt{3,136}\times{\sqrt{1,764}}}=\frac{2,352}{56\times42} =\frac{2,352}{2,352}=1 Coefficient of Correlation = 1 It means that there is perfect positive correlation between the values of Series X and Series Y. 4. Step Deviation MethodThis method simplifies the calculation of coefficient of correlation as the deviations are taken from assumed means and are divided by a common factor. The steps involved in the calculation of coefficient of correlation by using Step Deviation Method are: First of all, take the deviations of Series X from the assumed mean and divide them by Common Factor (C) to determine step deviation (dx^\prime) . Calculate the total of step deviations; i.e., \sum{dx^\prime} Take the deviations of Series Y from the assumed mean and divide them by Common Factor (C) to determine step deviation (dy^\prime) . Calculate the total of step deviations; i.e., \sum{dy^\prime} Square the step deviation of Series X and determine their total; i.e., \sum{dx^\prime{^2}} Square the step deviation of Series Y and determine their total; i.e., \sum{dy^\prime{^2}} Multiply (dx^\prime) and (dy^\prime) , and determine their total; i.e., \sum{dx^\prime{dy^\prime}} Now, use the following formula to determine Coefficient of Correlation:r=\frac{N\sum{dx^\prime{dy^\prime}}-\sum{dx^\prime}.\sum{dy^\prime}}{\sqrt{N\sum{dx^\prime{^2}}-(\sum{dx^\prime})^2}{\sqrt{N\sum{dy^\prime{^2}}-(\sum{dy^\prime})^2}}} Where, N = Number of pair of observations \sum{dx^\prime} = Sum of deviations of X values from assumed mean \sum{dy^\prime} = Sum of deviations of Y values from assumed mean \sum{dx^\prime{^2}} = Sum of squared deviations of X values from assumed mean \sum{dy^\prime{^2}} = Sum of squared deviations of Y values from assumed mean \sum{dx^\prime{dy^\prime}} = Sum of the products of deviations (dx^\prime) and (dy^\prime) Example:Use Step Deviation Method and determine the coefficient of correlation for the following data: Solution:r=\frac{N\sum{dx^\prime{dy^\prime}}-\sum{dx^\prime}.\sum{dy^\prime}}{\sqrt{N\sum{dx^\prime{^2}}-(\sum{dx^\prime})^2}{\sqrt{N\sum{dy^\prime{^2}}-(\sum{dy^\prime})^2}}} =\frac{(7\times35)-(7\times7)}{\sqrt{(7\times35)-(7)^2}\times{\sqrt{(7\times35)-(7)^2}}} =\frac{245-49}{\sqrt{245-49}\times{\sqrt{245-49}}} =\frac{196}{\sqrt{196}\times{\sqrt{196}}}=\frac{196}{14\times14} =\frac{196}{196}=1 Coefficient of Correlation = 1 It means that there is a perfect positive correlation between the values of Series X and Series Y. Karl Pearson's Coefficient of Correlation | Assumptions, Merits and Demerits Change of Scale and OriginCoefficient of Correlation does not depend upon the change of scale and origin. Change of Origin: If a constant is added or subtracted to the values then it will not have any effect on the value of correlation coefficient.Change of Scale: Similarly, if a constant is multiplied or divided by the values, then it will not have any effect on the value of correlation coefficient.Example:Find the coefficient of correlation from the following figures: Solution:As the coefficient of correlation is not affected by the change in scale and origin of the variables, we will multiply the X Series by 10 and divide the Y series by 100. r=\frac{N\sum{dxdy}-\sum{dx}.\sum{dy}}{\sqrt{N\sum{dx^2}-(\sum{dx})^2}{\sqrt{N\sum{dy^2}-(\sum{dy})^2}}} =\frac{(8\times156)-[(-24)\times(-4)]}{\sqrt{(8\times1,584)-(-24)^2}\times{\sqrt{(8\times44)-(-4)^2}}} =\frac{1,248-96}{\sqrt{12,672-576}\times{\sqrt{352-16}}} =\frac{1,152}{\sqrt{12,096}\times{\sqrt{336}}}=\frac{1,152}{110\times18.3} =\frac{1,152}{2,013}=0.57 Coefficient of Correlation = 0.57 It means that there is a moderate degree of positive correlation between variables X and Y. Correlation: Meaning, Significance, Types and Degree of Correlation Methods of measurements of Correlation Calculation of Correlation with Scattered Diagram Spearman's Rank Correlation Coefficient in Statistics CBSE Class 11 Statistics for Economics Notes Data Collection Methods | Primary and Secondary Data Sources of Data Collection | Primary and Secondary Sources Direct Personal Investigation: Meaning, Suitability, Merits, Demerits and Precautions Indirect Oral Investigation : Suitability, Merits, Demerits and Precautions Difference between Direct Personal Investigation and Indirect Oral Investigation Information from Local Source or Correspondents: Meaning, Suitability, Merits, and Demerits Questionnaires and Schedules Method of Data Collection Difference between Questionnaire and Schedule Qualities of a Good Questionnaire and Types of Questionnaires What are the Published Sources of Collecting Secondary Data? What Precautions should be taken before using Secondary Data? Two Important Sources of Secondary Data: Census of India and Reports & Publications of NSSO What is National Sample Survey Organisation (NSSO)? What is Census Method of Collecting Data? Sample Method of Collection of Data Methods of Sampling Father of Indian Census What makes a Sampling Data Reliable? Difference between Census Method and Sampling Method of Collecting Data What are Statistical Errors? Organization of Data Objectives and Characteristics of Classification of Data Classification of Data in Statistics | Meaning and Basis of Classification of Data Concept of Variable and Raw Data Types of Statistical Series Difference between Frequency Array and Frequency Distribution Types of Frequency Distribution Textual Presentation of Data: Meaning, Suitability, and Drawbacks Tabular Presentation of Data: Meaning, Objectives, Features and Merits Different Types of Tables Classification and Tabulation of Data Diagrammatic Presentation of Data: Meaning , Features, Guidelines, Advantages and Disadvantages Types of Diagrams Bar Graph | Meaning, Types, and Examples Pie Diagrams | Meaning, Example and Steps to Construct Histogram | Meaning, Example, Types and Steps to Draw Frequency Polygon | Meaning, Steps to Draw and Examples Ogive (Cumulative Frequency Curve) and its Types What is Arithmetic Line-Graph or Time-Series Graph? Diagrammatic and Graphic Presentation of Data Measures of Central Tendency in Statistics Arithmetic Mean: Meaning, Example, Types, Merits, and Demerits What is Simple Arithmetic Mean? Calculation of Mean in Individual Series | Formula of Mean Calculation of Mean in Discrete Series | Formula of Mean Calculation of Mean in Continuous Series | Formula of Mean Calculation of Arithmetic Mean in Special Cases Weighted Arithmetic Mean Median(Measures of Central Tendency): Meaning, Formula, Merits, Demerits, and Examples Calculation of Median for Different Types of Statistical Series Calculation of Median in Individual Series | Formula of Median Calculation of Median in Discrete Series | Formula of Median Calculation of Median in Continuous Series | Formula of Median Graphical determination of Median Mode: Meaning, Formula, Merits, Demerits, and Examples Calculation of Mode in Individual Series | Formula of Mode Calculation of Mode in Discrete Series | Formula of Mode Grouping Method of Calculating Mode in Discrete Series | Formula of Mode Calculation of Mode in Continuous Series | Formula of Mode Calculation of Mode in Special Cases Calculation of Mode by Graphical Method Mean, Median and Mode| Comparison, Relationship and Calculation Measures of Dispersion | Meaning, Absolute and Relative Measures of Dispersion Range | Meaning, Coefficient of Range, Merits and Demerits, Calculation of Range Calculation of Range and Coefficient of Range Interquartile Range and Quartile Deviation Partition Value | Quartiles, Deciles and Percentiles Quartile Deviation and Coefficient of Quartile Deviation: Meaning, Formula, Calculation, and Examples Quartile Deviation in Discrete Series | Formula, Calculation and Examples Quartile Deviation in Continuous Series | Formula, Calculation and Examples Mean Deviation: Coefficient of Mean Deviation, Merits, and Demerits Calculation of Mean Deviation for different types of Statistical Series Mean Deviation from Mean | Individual, Discrete, and Continuous Series Mean Deviation from Median | Individual, Discrete, and Continuous Series Standard Deviation: Meaning, Coefficient of Standard Deviation, Merits, and Demerits Standard Deviation in Individual Series Standard Deviation in Discrete Series Standard Deviation in Frequency Distribution Series Combined Standard Deviation: Meaning, Formula, and Example Coefficient of Variation: Meaning, Formula and Examples Lorenz Curveb : Meaning, Construction, and Application You can use the cor() function in R to calculate correlation coefficients between variables. Here are the most common ways to use this function: Method 1: Calculate Pearson Correlation Coefficient Between Two Variables cor(dfx, dfy) Use the Pearson correlation coefficient when calculating the correlation between two continuous variables. (e.g. height and weight) Method 2: Calculate Pearson Correlation Coefficient Between All Numeric Variables in Data Frame cor(df) This method will return a correlation matrix that contains the Pearson correlation coefficient between each pairwise combination of numeric variables in a data frame. Method 3: Calculate Spearman Correlation Coefficient Between Two Variables cor(dfx, dfy, method='spearman') Use the Spearman correlation coefficient when calculating the correlation between two ranked variables. (e.g. rank of a student's math exam score vs. rank of their science exam score in a class) Method 4: Calculate Kendall's Correlation Coefficient Between Two Variables cor(dfx, dfy, method='kendall') Use the Kendall correlation coefficient when you wish to use Spearman Correlation but the sample size is small and there are many tied ranks. The following examples show how to use each method in practice with the following data frame in R that shows the number of hours spent studying, number of practice exams taken, and final exam score for eight different students: #create data frame df