Click to prove you're human

The Pearson and Spearman correlation coefficients can range in value from −1 to +1. For the Pearson correlation coefficient to be +1, when one variable increases then the other variable increases by a consistent amount. This relationship forms a perfect line. The Spearman correlation coefficient is also +1 in this case. If the relationship is that one variable increases when the other increases, but the amount is not consistent, the Pearson correlation coefficient is positive but less than +1. The Spearman coefficient still equals +1 in this case. When a relationship is random or non-existent, then both correlation coefficients are nearly zero. If the relationship is a perfect line for a decreasing relationship, then both correlation coefficients are −1. If the relationship is that one variable decreases when the other increases, but the amount is not consistent, then the Pearson correlation coefficient is negative but greater than −1. The Spearman coefficient still equals −1 in this case Correlation values of −1 or 1 imply an exact linear relationship, like that between a circle's radius and circumference. However, the real value of correlation values is in quantifying less than perfect relationships. Finding that two variables are correlated often informs a regression analysis which tries to describe this type of relationship more. Consider a symphony orchestra tuning their instruments before a performance. Each musician adjusts their notes to harmonize with others, ensuring a seamless musical experience. In Data Science, the variables in a dataset can be compared to the orchestra's musicians: understanding the harmony or dissonances between them is crucial. Image source: pixabay.com. Correlation is a statistical measure that acts like the conductor of the orchestra, guiding the understanding of the complex relationships within our data. Here we will focus on two types of correlations: Pearson and Spearman. If our data is a composition, Pearson and Spearman are our orchestra's conductors: they have a singular style of interpreting the symphony, each with peculiar strengths and subtleties. Understanding these two different methodologies will allow you to extract insights and understand the connections between variables. The Pearson correlation coefficient, denoted as r, quantifies the strength and direction of a linear relationship between two continuous variables [1]. It is calculated by dividing the covariance of the two variables by the product of their standard deviations. Pearson coefficient formula Here X and Y are two different variables, and _Xi and _Yi represent individual data points. bar{X} and bar{Y} denote the mean values of the respective variables. The interpretation of r relies on its value, ranging from -1 to 1. A value of -1 implies a perfect negative correlation, indicating that as one variable increases, the other decreases linearly [2]. Conversely, a value of 1 signifies a perfect positive correlation, illustrating a linear increase in both variables. A value of 0 implies no linear correlation. Pearson correlation is particularly good at capturing linear relationships between variables. Its sensitivity to linear patterns makes it a powerful tool when investigating relationships governed by a consistent linear trend. Moreover, the standardized nature of the coefficient allows for easy comparison across different datasets. However, it's crucial to note that Pearson is susceptible to the influence of outliers. If a dataset contains extreme values they can impact the calculation, leading to inaccurate interpretations. Technical concepts can be better understood through practical examples. Let's use Python to show the computation of Pearson correlation and its visualization. Suppose we have two lists representing the hours spent studying (X) and the corresponding exam scores (Y). import numpy as np from scipy.stats import pearsonr import matplotlib.pyplot as plt import seaborn as sns # Generating data points np.random.seed(42) # For reproducibility hours_studied = np.random.randint(8, 25, size=50) exam_scores = 60 + 2 * hours_studied + np.random.normal(0, 5, size=50) # Calculate Pearson correlation coefficient pearson_corr, _ = pearsonr(hours_studied, exam_scores) # Calculate Pearson correlation line coefficients m, b = np.polyfit(hours_studied, exam_scores, 1) # Fit a linear regression line # Scatter plot fig, ax = plt.subplots() ax.scatter(hours_studied, exam_scores, color=sns.color_palette("hls",24)[14], alpha=.9, label='Data points') plt.plot(hours_studied, m * np.array(hours_studied) + b, color='red', alpha=.6, label='Pearson Correlation Line') plt.title("Hours Studied vs. Exam Scores") plt.xlabel("Hours Studied") plt.ylabel("Exam Scores") plt.legend(loc='lower right') ax.spines['top'].set_visible(False) ax.spines['bottom'].set_visible(False) ax.spines['right'].set_visible(False) ax.spines['left'].set_visible(False) ax.xaxis.set_ticks_position('none') ax.yaxis.set_ticks_position('none') Image by the author. Pearson correlation effectiveness diminishes when faced with curvilinear patterns. This limitation arises from Pearson's inherent assumption of linearity, making it ill-suited to capture the nuances of non-linear relationships. Image by the author. Consider a scenario where the relationship between two variables follows a quadratic curve. Pearson correlation might inaccurately suggest a weak or nonexistent relationship due to its inability to capture the non-linear relation. # Generating quadratic data np.random.seed(42) X = np.linspace(-10, 10, 100) Y = X**2 + np.random.normal(0, 10, size=len(X)) # Calculate Pearson correlation coefficient pearson_corr, _ = pearsonr(X, Y) m, b = np.polyfit(X, Y, 1) # Fit a linear regression line # Scatter plot fig, ax = plt.subplots() ax.scatter(X, Y, color=sns.color_palette("hls",24)[14], alpha=.9, label='Data points') plt.plot(X, m * X + b, color='red', alpha=.6, label='Pearson Correlation Line') plt.title("X vs. Y (Quadratic Relationship)") plt.xlabel("X") plt.ylabel("Y") plt.legend(loc='upper center') ax.spines['top'].set_visible(False) ax.spines['bottom'].set_visible(False) ax.spines['right'].set_visible(False) ax.spines['left'].set_visible(False) ax.xaxis.set_ticks_position('none') ax.yaxis.set_ticks_position('none') Image by the author. Spearman correlation addresses the limitations of Pearson when applied to non-linear relationships or datasets containing outliers [3]. Spearman's rank correlation coefficient (ρ), denoted as rho, operates on the ranked values of variables, making it less sensitive to extreme values and well-suited for capturing monotonic relationships. Spearman coefficient formula In the above formula, _di represents the difference between the ranks of corresponding pairs of variables, and n is the number of data points. Similar to Pearson, Spearman's coefficient ranges from -1 to 1. A value of -1 indicates a perfect negative monotonic correlation, meaning that as one variable increases, the other consistently decreases. A value of 1 signifies a perfect positive monotonic correlation, illustrating a consistent increase in both variables. A value of 0 denotes no monotonic correlation. Unlike Pearson, Spearman does not assume linearity and is robust in case of outliers. It focuses on the ordinal nature of data, making it a valuable tool when the relationship between variables is more about the order than the specific values. Image by the author. Consider the following practical application using Python, where we have two variables 'A' and 'B' with a non-linear relationship: import numpy as np from scipy.stats import spearmanr from scipy.stats import pearsonr import matplotlib.pyplot as plt # Generating non-linear data np.random.seed(42) A = np.linspace(-10, 10, 100) B = A**5 + np.random.normal(0, 4000, size=len(A)) # Calculate Spearman correlation coefficient spearman_corr, _ = spearmanr(A, B) # Calculate Pearson correlation coefficient pearson_corr, _ = pearsonr(A, B) m, b = np.polyfit(A, B, 1) # Fit a linear regression line # Scatter plot fig, ax = plt.subplots() ax.scatter(A, B, color=sns.color_palette("hls",24)[14], alpha=.9, label='Data points') plt.plot(X, m * X + b, color='red', alpha=.6, label='Pearson Correlation Line') plt.title("A vs. B (Non-linear Relationship)") plt.xlabel("A") plt.ylabel("B") ax.spines['top'].set_visible(False) ax.spines['bottom'].set_visible(False) ax.spines['right'].set_visible(False) ax.spines['left'].set_visible(False) ax.xaxis.set_ticks_position('none') ax.yaxis.set_ticks_position('none') plt.legend() Image by the author. In this example, the scatter plot visualizes a non-linear relationship between variables 'A' and 'B.' Spearman correlation, which does not assume linearity, will be better suited to capture and quantify this non-linear association. You can see that the red line, representing the Pearson Correlation Line, misses the nature of the variables' relationship. We can quantify this measure by comparing the two coefficients: Spearman coefficient is sensibly higher, as it is more suited for this type of relationship. Concluding this introductory guide, let's point out the pros and cons of the two measures. Image by the author. Pearson Correlation coefficient is indeed efficient for linear relationship, and can provide a standardized measure for easy comparison across different datasets. On the other hand, Pearson coefficient is highly sensitive to outliers. Also, assuming linearity may mislead in non-linear scenarios. Most of Pearson coefficient's downsides are addressed by Spearman Correlation coefficient: it is robust in the presence of outliers, and, as it relies on the rank of data points, it is suitable for non-linear relationships. It is important, however, to keep in mind that Spearman coefficient has also some criticalities. In my opinion the most relevant ones are its lower efficiency on large datasets, and the potential loss of information in case of tied ranks. In order to keep this guide concise and practical, my advice is to choose Pearson for linear relationships in normally distributed data without outliers. Opt for Spearman when facing non-linear relationships, ordinal data, or datasets with outliers, as it excels in capturing monotonic associations. Remember that this choice heavily relies on the characteristics of the data. Keep also in mind that this introduction, while touching most important aspects, merely scratched the surface of statistical methodologies. For further study, you can find an interesting set of resources below! In-text references: [1] Laerd Statistics – Pearson Product-Moment Correlation [2] Boston University – Correlation and Regression with R [3] Analytics Vidhya – Pearson vs Spearman Correlation Coefficients [4] Statistics by Jim – Spearman's Correlation Explained Other material: Key Differences between Pearson and Spearman Correlation Coefficient are -AspectPearson Correlation CoefficientSpearman Correlation CoefficientType of RelationshipMeasures linear relationships between variables.Measures monotonic relationships, where variables move consistently in one direction (not necessarily linearly).Data TypeWorks with continuous interval or ratio data.Suitable for ordinal, ranked, interval, or ratio data.AssumptionsAssumes linearity and normal distribution of data.Does not require normality or linearity; works well with non-parametric data.Sensitivity to OutliersSensitive to outliers, which can skew the correlation value.Resistant to outliers since it uses ranks instead of raw data.Calculation MethodBased on covariance and standard deviations of raw values.Based on ranking the data points and calculating the difference in ranks.Range of CoefficientRanges from -1 to 1 (negative, positive, or no linear correlation).Ranges from -1 to 1 (negative, positive, or no monotonic correlation).Ideal Use CasesUse when the data follows a normal distribution and shows linear trends.Use for non-linear or ranked data, or when dealing with outliers.Fields of ApplicationFinance, healthcare, machine learning (e.g., stock price correlation).Education, psychology, customer satisfaction surveys (e.g., rank-based analysis).ExampleAnalyzing the relationship between height and weight of individuals.Assessing the relationship between study hours and exam ranks of students.When to Use Each CoefficientUse Pearson if:Your data is continuous and normally distributed.You expect a linear relationship between variables.You are concerned about the precise strength of a linear association.Use Spearman if:Your data is ordinal, ranked, or non-normally distributed.The relationship between variables is monotonic but not necessarily linear.Your data includes outliers that could distort a Pearson analysis.Read More, The procedure to use is, of course, a correlational analysis, but which type should you use? In this guide, we'll walk you through the two main methods you could use for correlation. These methods are called the Pearson correlation and the Spearman correlation. We'll take a look at what each technique involves, when each should be used, and the types of research questions that could be addressed. Also, if you are conducting usage and attitudes (U&A) research or concept testing, we can perform the analysis for you.Before going into detail about the statistical techniques used to perform a correlational analysis, let's quickly define what we mean by correlation. Correlational analysis is a bivariate (two variable) statistical procedure that sets out to identify the mean value of the product of the standard scores of matched pairs of observations. The purpose of this type of analysis is to find out whether changes in one variable produce changes in another. For example, does customer satisfaction increase with the size of discount offered at a grocery store or does employee engagement rise with salary increases? Note that correlation is used to infer whether there is a relationship between the two variables, not whether changes in one variable cause changes in another. In other words, correlation says nothing about causality. In our example above, for instance, employees might be more engaged because they're rewarded with higher salaries. Alternatively, higher levels of engagement might drive managers to increase their wages. Correlation says nothing about which variable impacts the other, but rather tells us whether there is a simple relationship between the variables, the direction of the relationship (positive or negative), and its strength.Of two techniques used to perform correlation analysis, the Pearson correlation method is probably the most recognized and widely used in market and business research. Let's take a look at what the Pearson correlation method is, and how you can use it.The Pearson product moment correlation coefficient can be described as a way to measure the strength of a linear relationship between two variables—which can be used to find out if there is strong association between one variable versus another. Imagine you have two variables—such as employee engagement and employee salaries—plotted on a simple scatter plot graph. The Pearson correlation essentially tries to utilize a scatter plot by drawing a line through the data in order to find out whether the two comparables are covary with one another and to what extent. That is, Pearson correlation coefficient identifies whether:There is a positive correlation between the two variables. That is, whether an increase in employee engagement is associated with an increase in salaries.There is a negative correlation between the two variables. More specifically, whether a rise in salaries is associated with a reduction in employee engagement, or vice versa.There is no relationship between the variables. In other words, changes in salaries and employee engagement are unrelated to one another.Insight into this relationship is a first step in understanding how variables of interest might relate to one another, and could also prompt further causal investigation. The Pearson correlation coefficient test compares the mean value of the product of the standard scores of matched pairs of observations. Once performed, it yields a number that can range from -1 to +1. Positive figures are indicative of a positive correlation between the two variables, while negative values indicate a negative relationship. Furthermore, the value of r represents the strength of the relationship. A Pearson's r that is near the value of 1 is suggestive of a stronger relationship between the two variables. As a rule of thumb, the following values can be used to determine the strength of the relationship:A Pearson correlation coefficient of between 0 and 0.3 (or 0 and -.03) indicates a weak relationship between the two variablesA Pearson correlation coefficient of between 0.4 and 0.6 (or -.04 and -.06) indicates a moderate strength relationship between the two variablesA Pearson correlation coefficient of between 0.7 and 1 (or -.07 and 1) indicates a strong relationship between the two variables.For example, imagine that you've developed some marketing concepts that you've begun testing with some potential customers. For each concept, you're interested in learning whether evaluations of the appeal of the concept are associated with stronger intent to purchase. Comparisons of Concepts A, B and C yield Pearson correlation coefficients of .3 .6 and .9 respectively. Based on these three figures, you can infer the following:For all three market concepts, there is a positive correlation between evaluations of concept appeal and intent to purchase the purchaseHowever, the correlation between concept appeal and intent to purchase is strongest for Concept C, and weakest for Concept For Concept B, there is positive correlation between concept appeal and purchasing intent but the relationship is moderateUsing these inferences, you might decide that Concept C is the most appropriate concept to employ in your next marketing campaign. However, first, you'll need to determine whether the correlation you've observed is statistically significant. Let's look at the formula used to determine Pearson's r in more detail, and how you can combine this formula with a t test to determine significance. The Pearson correlation coefficient coefficient (r) is calculated using the following expression: Where xi represents the values of the x variable in a sample, x-bar indicates the mean of the values of the x variable, yi indicates the values of the y variable, and y-bar indicates the mean of the values of the y-variable. S indicates the sum of squares of the x and y variables respectively, and n is the number of observations of x and y variables. After an r value is produced, the next step is to determine whether the value is of statistical significance. The importance of this step cannot be overstated. It is possible to observe two variables that seem to be related to one another, but the relationship is in fact meaningless. For example, you might observe a relationship between concept appeal and intended purchase frequency, leading you to believe that the concept that has the greatest appeal will lead people to spend more. However, if this relationship occurred merely through chance, your marketing campaign might turn out to be an expensive waste of cash. Statistical significance indicates that we are confident of a relationship between the two variables; in other words, that the result did not occur by chance.A t test is used to establish if the Pearson's r statistic differs significantly from zero. Statistical significance (indicated by the probability, or p) indicates whether the observer can be confident of a relationship between the two variables at different levels. For instance, a p value of .05 indicates that there is only a 5% chance that that relationship occurred by chance, while a p value of .10 indicates that there is a 10% chance that the observed correlation is a chance event. The t statistic always has the same sign (+ or -) as the r value and is calculated as follows:t = r* √((n-2) / (1 - r*r))Once the t-value is calculated, it can be compared with the critical value from a standard t-table at the appropriate degrees of freedom (n-1) and the level of confidence (p value) you wish to maintain in order to determine the significance, and therefore the extent to which the correlation you have observed is meaningful.Introducing MomentiveLearn what your audience really wants with an AI-powered solution. Shape your product and marketing strategy with our Usage and Attitudes solution.Practical applications of the Pearson correlation coefficientThe Pearson correlation is a relatively simple equation, but its uses are myriad. You can use apply this technique to answer research questions such as:Is there a statistically significant relationship between age, as measured in years, and height, measured in inches?Is there a relationship between job satisfaction, as measured by the JSS, and income, measured in dollars?Is there an association between levels of household income and customer spend?Are higher levels of education associated with greater happiness?The Spearman's test is a non-parametric version of the parametric Pearson bivariate correlation coefficient. What does this mean? Well, parametric tests and non-parametric tests are distinguished on the basis of assumptions that they make about the nature of the data to be analyzed. A parametric statistical test is a test that makes clear assumptions about the defining properties, or parameters, of the dataset. For a dataset to be appropriate for the parametric version of correlational analysis (i.e. the Pearson correlation), the following assumptions must be met:Each variable must be continuous in nature. In other words, each variable is able to take on a potentially infinite number of values, such as age, income or score evaluations.The shape of the relationship between the variables must be linear. This means that when a scatter plot of the two variables is drawn, the shape of the line of best fit should approximate a straight line rather than a curve.If either of these assumptions are violated, you should use the nonparametric version of the correlation technique, known as Spearman's correlation, Spearman's rank-order test, or Spearman's rho.What is the Spearman correlation coefficient?Like the Pearson test, the Spearman correlation test examines whether two variables are correlated with one another or not. The Spearman's test can be used to analyse ordinal level, as well as continuous level data, because it uses ranks instead of assumptions of normality. This makes the Spearman correlation great for 3, 5, and 7-point likert scale questions or ordinal survey questions.The Spearman's test is therefore useful where the basic assumptions of linearity and continuous variables necessary to perform a Pearson's bivariate correlation analysis have not been met.The Spearman's rank-order tests determines the strength and direction of the monotonic relationship between two variables measured at ordinal, interval or ratio level. As with the Pearson equivalent, the test will yield a figure of between -1 and +1, and the closer the figure is to 1, the stronger the monotonic relationship. As a rule of thumb, you can use the following figures to determine the effect size:A Spearman's correlation coefficient of between 0 and 0.3 (or 0 and -.03) indicates a weak monotonic relationship between the two variablesA Spearman's correlation coefficient of between 0.4 and 0.6 (or -.04 and -.06) indicates a moderate strength monotonic relationship between the two variablesA Spearman's correlation coefficient of between 0.7 and 1 (or -.07 and 1) indicates a strong monotonic relationship between the two variables.In addition, probability values should be used to determine statistical significance.An example is the best way to understand how to calculate a Spearman's correlation. Imagine you've gathered some data on evaluations of a market concept, and the average price a consumer is willing to pay. Take the following steps:In the first instance, you should create a table from your data. Each set of measurements should be ranked by assigning the ranking 1 to the largest number in a column, 2 to the next largest value, 3 to the third largest and so on (tied scores can be assigned the mean rank). Then, find the difference in the ranks (d). This is the difference between the ranks of the two values on each row, calculated by subtracting the ranking of the second value (in this example, price) from the rank of the first (concept evaluation).Finally, square the differences (d2) and then sum them.Now, you have all the data you need to calculate Spearman's rank, using the following formula:In our example, we would first multiply the sum of the d2 values (6) by 6 (i.e. 36). To address the denominator, we would raise the number of observations (concepts) to the power of 3 and then subtract the number of observations (i.e. 43-4 = 60). We can then calculate Spearman's rho as 1-36/60= -.058. This indicates a moderate, negative monotonic correlation between concept evaluation and the price that consumers are willing to pay. A difficult one to interpret!Practical applications of the Spearman's correlation coefficientThe types of research questions that can be addressed through the Spearman correlation method are similar to those addressable through a Pearson analysis. Remember, however, that the main difference is that data can be ordinal in nature, and the relationship should be monotonic. For example, you could use the Spearman correlation coefficient to answer questions like:Is there a statistically significant relationship between participants' level of education and their starting salary?Does income range vary with spend habits?What is the association between size of home and number of inhabitants?Confused about when to use the Pearson correlation and when to use the Spearman's correlation coefficient? Remember that Spearman's correlation determines the strength and direction of the monotonic relationship between your two variables rather than the strength and direction of the linear relationship between your two variables, which is what Pearson's correlation determines. Linear relationships are straight line relationships. Monotonic relationships differ from linear relationships in that the two variables might converge, but not at a constant rate. There are three types of monotonic functions:Monotonically increasing relationshipsThis means that as the x variable increases, the y variable never decreases.Monotonically decreasing relationshipsThis means that as the x variable increases, the y variable never increases.Non monotonic relationshipsThis means that as the x variable increases, the y variable sometimes decreases and sometimes increases.So, that's correlation in a nutshell, and how and when to use it. It can be complicated, but the good news is that if you're planning on a usage and attitudes (U&A) survey or performing some concept testing, we can handle the correlation analysis for you through our Key Driver Analysis feature.Get started with your market researchCollect market research data by sending your survey to a representative sampleGet help with your market research project by working with our expert research teamTest creative or product concepts using an automated approach to analysis and reportingTo read more market research resources, visit our Sitemap. Are you as a data scientist trying to decipher relationship between two or more variables within vast datasets to solve real-world problems? Whether it's understanding the connection between physical exercise and heart health, or the link between study habits and exam scores, uncovering these relationships is crucial. But with different methods at our disposal, how do we choose the most suitable one? This is where the concept of correlation comes into play, and particularly, the choice between Pearson and Spearman correlation coefficients becomes pivotal. The Pearson correlation coefficient is the go-to metric when both variables under consideration follow a normal distribution, assuming there's a linear relationship between them. Conversely, the Spearman correlation coefficient doesn't hinge on the normality of the data and offers a more flexible approach, effectively capturing monotonic relationships even when distributions are non-normal. Choosing the right correlation coefficient is not just a matter of mathematical preference; it can significantly impact the conclusions we draw from our data. In this blog, we'll dive into the intricacies of both the Pearson and Spearman correlation coefficients, demystify their differences, and equip you with the knowledge to select the right one for your data. Correlation is a statistical measure that describes the extent to which two variables change together. It is a crucial concept in data science, as it helps in understanding and quantifying the strength and direction of the relationship between variables. For data scientists, correlation is foundational for feature selection, risk assessment, hypothesis testing, and predictive modeling. A thorough understanding of correlation is essential in discerning which variables in large datasets have the potential to provide insights or predictive power when algorithms are applied. The importance of correlation in the realm of data science cannot be overstated. As data scientists, when we understand the correlation between variables, we can make informed decisions about which variables may influence one another and how they can be harnessed in statistical modeling strategies. This understanding can also help in simplifying machine learning / statistical models by identifying and removing redundant variables, thereby improving computational efficiency and model interpretability. However, it is imperative to remember that correlation does not imply causation. Just because two variables display a strong correlation, it doesn't mean that one variable causes the other to occur. They may be influenced by a third variable, or the observed correlation could be coincidental. Data scientists must be cautious not to leap to conclusions about cause and effect solely based on correlation metrics. Rigorous experimental design and analysis are necessary to establish causative links. The Pearson correlation coefficient, often symbolized as r, is a measure that quantifies the linear relationship between two continuous variables. It is the most widely used correlation statistic to assess the degree of the relationship between linearly related variables. The Pearson coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations. The formula is expressed as: r = \frac{\sum_{i=1}^{n} (X_i – \bar{X})(Y_i – \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i – \bar{X})^2 \sum_{i=1}^{n} (Y_i – \bar{Y})^2}} where: Xi and Yi are the individual sample points indexed with i, X_bar and Y_bar are the means of the X and Y variables, respectively. The Pearson correlation coefficient has several key characteristics: It yields a value between -1 and 1, where 1 implies a perfect positive linear relationship, -1 implies a perfect negative linear relationship, and 0 implies no linear relationship. It is symmetric, meaning the correlation between X and Y is the same as the correlation between Y and X. For the Pearson correlation coefficient to be valid, certain assumptions must be met: The scale of measurement should be interval or ratio. The relationship between the two variables should be linear. The data should be bivariately normally distributed. The data comes from continuous variables. The above can also be taken as criteria that we can use to decide whether to use Pearson correlation coefficient. The Spearman correlation coefficient, denoted as ρ or sometimes as [latex]r_s[/latex], is a non-parametric measure of rank correlation. It is "non-parametric" because it doesn't make any assumptions about the probability distribution of the variables (i.e., they do not need to follow a normal distribution). Instead of calculating the correlation using raw data, it operates on the ranks of the data. "Rank correlation" implies that the correlation is determined by comparing the ranks of the data points, rather than their actual values. Each value is replaced by its rank in the dataset when calculating Spearman's correlation. The Spearman correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function, whether linear or not. Example of Rank Correlation Suppose we are interested in the relationship between the time spent studying for an exam (in hours) and the marks obtained (out of 100). We have data from five students as follows: StudentHours Studied (X)Marks Obtained (Y)A150B470C360D580E255 To calculate the Spearman rank correlation, we would first rank each set of data (hours studied and marks obtained) from lowest to highest. StudentHours Studied (X)Rank RXMarks Obtained (Y)Rank RYA11501B44704C33603D55805E22552 We then calculate the Spearman correlation using these ranks. The idea is that if there is a perfect monotonic relationship, the ranks would match perfectly (i.e., the highest number of hours studied would correspond to the highest marks obtained, and so on). If there is no relationship, the ranks would not correspond at all. Spearman correlation coefficient is primarily used for ordinal data: Ordinal data represent categories with a meaningful order, but the intervals between the categories are not necessarily equal or known. Here's an example of ordinal data where the Spearman correlation coefficient would be appropriate: A company might conduct a survey to assess customer satisfaction with its services. The survey contains two questions where customers rate the following on a scale from 1 to 5: Satisfaction with customer service (1 = Very Unsatisfied, 2 = Unsatisfied, 3 = Neutral, 4 = Satisfied, 5 = Very Satisfied) Likelihood of recommending the service to a friend (1 = Very Unlikely, 2 = Unlikely, 3 = Neutral, 4 = Likely, 5 = Very Likely) Here, the data are ordinal. Each number represents a category that is ranked relative to the others, but the difference in satisfaction between "Very Unsatisfied" and "Unsatisfied" may not be the same as the difference between "Neutral" and "Satisfied." The company is interested in understanding whether there is a relationship between customer satisfaction and their likelihood of recommending the service. In this case, a Spearman correlation coefficient can be used to assess how well the rankings of customer satisfaction correlate with the rankings of their likelihood to recommend. The Spearman coefficient is suitable here because it doesn't assume equal intervals between ranks and is not influenced by the non-linear spacing between the ordinal categories. It simply assesses whether customers who are more satisfied are also more likely to recommend the service (and vice versa), based on the rank order of their responses. If customers who are more satisfied also tend to be more likely to recommend the service, this would result in a high positive Spearman correlation, indicating a strong monotonic relationship. Unlike Pearson's r, which assesses linear relationships and relies on parametric assumptions, the Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. The formula for Spearman's ρ is: \rho = 1 – \frac{6 \sum d_i^2}{n(n^2 – 1)} where: di is the difference between the ranks of corresponding variables, n is the number of observations. The Spearman correlation shares some characteristics with the Pearson correlation: It ranges from -1 to +1, where +1 signifies a perfect increasing monotonic relationship, -1 signifies a perfect decreasing monotonic relationship, and 0 signifies no monotonic relationship. Spearman's ρ is different in that it does not require the assumption of normally distributed data and is less affected by outliers due to its use of ranks rather than raw data. Spearman correlation makes fewer assumptions than Pearson's: It does not require the assumption of linearity or normality. It assumes that the variables are either continuous or ordinal. It is suitable for non-linear relationships, as long as they are monotonic. The above can also be taken as criteria that we can use to decide whether to use Spearman correlation coefficient. The Spearman correlation coefficient also has its limitations: While it can handle non-linear relationships, it only measures monotonic ones; if the relationship is non-monotonic, Spearman's correlation will not be able to capture it effectively. It can sometimes overestimate the strength of a relationship when dealing with ordinal data that have a lot of tied ranks. Similar to Pearson, it does not imply causation. When diving into the realm of data analysis, it becomes crucial to understand the differences between Pearson and Spearman correlation coefficients, as each serves different purposes and is appropriate under varying circumstances. Similarities between Pearson and Spearman Both Pearson and Spearman correlation coefficients aim to measure the strength and direction of the relationship between two variables. They are bounded by the same limits, producing a value between -1 and +1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and 0 suggests no correlation. Additionally, both are symmetric, meaning the correlation from X to Y is the same as from Y to X. Key Differences and When Each Should Be Applied Scale of Measurement: Pearson: Requires interval or ratio scale data with linear characteristics. Spearman: Can primarily be used with ordinal data. Nature of the Relationship: Pearson: Assumes a linear relationship between variables. Spearman: Does not assume linearity, only that the relationship is monotonic. Presence of Outliers: Pearson: Sensitive to outliers, which can significantly affect the correlation. Spearman: More robust to outliers because it uses rank order rather than actual values. Data Distribution Requirements: Pearson: Assumes that both variables are normally distributed. Spearman: No assumption of normal distribution, making it more versatile for different types of data distributions. The Pearson correlation coefficient is a statistical measure of the linear relationship between two variables. It ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship. To represent the Pearson correlation coefficient with a plot, we typically create a scatter plot of the two variables and add the line of best fit to visualize the linear relationship. The slope of the line and how closely the points cluster around the line give a visual indication of the strength and direction of the linear relationship. The following is the plot: Here's the scatter plot with a line of best fit for a dataset with a positive linear relationship. The Pearson correlation coefficient (r) is annotated on the plot, indicating the strength and direction of the linear relationship between the x and y variables. The line of best fit gives a visual representation of this relationship, and the scatter of points around the line indicates how closely they follow a linear pattern. The closer the Pearson coefficient is to 1, the stronger the positive linear relationship between the variables. The Spearman correlation coefficient is used to represent relationship between the variables in the non-normal monotonic dataset. The following plots represent the non-normal monotonic dataset with or without outliers. In the visualizations: Left Plot (Without Outliers): This plot shows the dataset without outliers, and both Pearson and Spearman correlation coefficients are relatively high, indicating a strong monotonic relationship. Since the data is non-normal and monotonic, both coefficients show similar strength due to the lack of outliers and the overall pattern of the data. Right Plot (With Outliers): Here, the data consists of few outliers. You can see that the Pearson correlation coefficient has dropped significantly due to the presence of these outliers. This is because Pearson's correlation is heavily influenced by the actual values, and outliers can have a large impact on the result. In contrast, the Spearman correlation coefficient has not been affected as much by the outliers because it relies on the rank order of the data, which is less sensitive to extreme values. Selecting the appropriate method (Spearman vs Pearson) to measure correlation requires careful consideration of various aspects of the data. Determining the Scale of Measurement in Your Data Firstly, it's essential to identify the scale of measurement. Pearson's correlation is suitable for data measured on an interval or ratio scale—where the intervals between data points are equal. Examples include temperature in Celsius or revenue in dollars. Spearman's correlation is apt for ordinal data or interval/ratio data that do not meet the normality assumption. An example of ordinal data could be a rating scale from 1 to 5, as discussed previously. Assessing the Relationship Between Variables The nature of the relationship between variables is another critical factor. If the relationship is linear, meaning that the change in one variable is proportionally associated with a change in another, Pearson's correlation should be used. If the relationship is monotonic, where the variables tend to move in the same direction but not necessarily at a constant rate, Spearman's correlation is more appropriate. Dealing with Outliers and Non-normal Distributions Outliers can significantly impact the results of a Pearson correlation analysis. If your data contains outliers or is not normally distributed, Spearman's correlation, which uses ranks rather than actual values, can provide a more accurate measure of the relationship. There is no one-size-fits-all approach when choosing between Pearson and Spearman correlation coefficients. Each dataset should be evaluated on its own merits, and the choice should be justified based on the characteristics of the data and the specific research questions posed. 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. Pearson and Spearman correlation coefficients are two widely used statistical measures when measuring the relationship between variables. The Pearson correlation coefficient assesses the linear relationship between variables, while the Spearman correlation coefficient evaluates the monotonic relationship. In this article, we will delve into a comprehensive comparison of these correlation coefficients for correlation analysis. We will explore their calculation methods, interpretability, strengths, and limitations. Understanding the differences between Pearson and Spearman correlation coefficients is crucial for selecting the appropriate measure based on the nature of the data and the research objectives. Also, we are covering the difference between Pearson and Spearman correlation. We will explore Pearson vs Spearman, highlighting their unique applications, and discuss when to use Pearson correlation vs Spearman in data analysis. Let's explore the difference between Pearson vs Spearman Correlation Coefficients! Table of contents Correlation is a bivariate statistical measure that tells us about the association between the two variables. It describes how one variable behaves if there is some change in the other variable. If the two variables are increasing or decreasing in parallel then they have a positive correlation between them and if one of the variables is increasing and another one is decreasing then they have a negative correlation with each other. If the change of one variable has no effect on another variable then they have a zero correlation between them. Correlation coefficients are like universal translators in the world of machine learning and data science. They help us understand the language between variables – how much, and in what direction, they change together. Here's why they're crucial: Finding patterns: Uncovering hidden relationships between features, like what factors influence house prices. Picking the best features: Choosing the most relevant data for machine learning models, making them more efficient. Understanding models: Seeing how models interpret data and identifying potential issues. Read More about this article Type of Correlation Metrics Used by Data Scientists Spearman's correlation, another name for Spearman's rank correlation coefficient, is a statistical tool that dives into how two variables are connecte. Instead of assuming a straight line relationship, it assesses how much one variable tends to go up or down as the other changes along with it. This change, called a monotonic relationship, can be either a steady increase together or a consistent decrease together. Even if the data doesn't form a perfect line, Spearman's correlation can reveal this underlying trend. AspectPearson Correlation CoefficientSpearman Correlation CoefficientPurposeMeasures linear relationshipsMeasures monotonic relationshipsAssumptionsVariables are normally distributed, linear relationshipVariables have monotonic relationship, no assumptions on distributionCalculation MethodBased on covariance and standard deviationsBased on ranked data and rank orderRange of Values-1 to 1-1 to 1InterpretationStrength and direction of linear relationshipStrength and direction of monotonic relationshipSensitivity to OutliersSensitive to outliersLess sensitive to outliersData TypesAppropriate for interval and ratio dataAppropriate for ordinal variables and non-normally distributed dataSample SizeThe Pearson correlation coefficient isn't the most efficient choice for small sample sizes.This method works well with smaller samples and doesn't require normality assumptions.UsageAssessing linear associations, parametric testsAssessing monotonic associations, non-parametric tests The Pearson correlation coefficient also known as linear correlation is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, with values close to -1 indicating a strong negative linear relationship, values close to 1 indicating a strong positive linear relationship, and 0 indicating no linear relationship. The Spearman correlation coefficient is a statistical measure that assesses the strength and direction of a monotonic relationship between two variables. It ranks the data rather than relying on their actual values, making it suitable for non-normally distributed or ordinal data. It ranges from -1 to 1, where values close to -1 or 1 indicate a strong monotonic relationship, and 0 indicates no monotonic relationship. Spearman correlation is valuable for detecting and quantifying associations when linear relationships are not assumed or when dealing with ranked or ordinal scale. Spearman's Rank Correlation: Let's say we want to determine the relationship between the study time (in hours) and the exam scores (out of 100) of a group of students. We have the following data for five students: StudentStudy Time (hours)Exam ScoreA1075B860C1285D655E970 First, we rank the study time and exam scores separately: StudentStudy Time (hours)Rank (Study Time)Exam ScoreRank (Exam Score)A103753B84605C121851D65556E92704 Now, we calculate the differences between the ranks for each pair of data points: P=Rank of Study Time−Rank of Exam Score, Di=Rank of Study Timei−Rank of Exam Scorei Next, we square each (Di) value: The sum of ��2Di2 is 0+1+0+1+4=60+1+0+1+4=6. Determining the association between Girth and Height of Black Cherry Trees (Using the existing dataset "trees" which is already present in r and can be accessed by typing the name of the dataset, list of all the data set can be seen by using the command data()) Below is the code to compute the correlation: > data head(data, 3) Girth Height Volume 1 8.3 70 10.3 2 8.6 65 10.3 3 8.8 63 10.2 > library(ggplot2) > ggplot(data, aes(x = Girth, y = Height)) + geom_point() + + geom_smooth(method = "lm", se =TRUE, color = 'red') Here two assumptions are checked which need to be fulfilled before performing the correlation (Shapiro test, which is test to check the input variable is following the normal distribution or not, is used to check whether the variables i.e. Girth and Height are normally distributed or not) > shapiro.test(data$Girth) Shapiro-Wilk normality test data: data$Girth W = 0.94117, p-value = 0.08893 > shapiro.test(data$Height) Shapiro-Wilk normality test data: data$Height W = 0.96545, p-value = 0.4034 p–value is greater than 0.05, so we can assume the normality > cor(data$Girth,data$Height, method = "pearson") [1] 0.5192801 > cor(data$Girth,data$Height, method = "spearman") [1] 0.4408387 > Pear Pear Pearson's product-moment correlation data: data$Girth and data$Height t = 3.2722, df = 29, p-value = 0.002758 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.2021327 0.7378538 sample estimates: cor 0.5192801 > Spear Spear Spearman's rank correlation rho data: data$Girth and data$Height S = 2773.4, p-value = 0.01306 alternative hypothesis: true rho is not equal to 0 sample estimates: rho 0.4408387 Since the p-value is less than 0.