NAME : HARI VIGNESH.R

DATA SCIENCE

Different Methods for Correlations

Link :

<https://easystats.github.io/correlation/articles/types.html>

VIDEO LINK : 2. <https://www.youtube.com/watch?v=PEfQCv9nvSo> (or)

3. <https://www.youtube.com/watch?v=Rx-IA-5zAAw>

Correlations tests are arguably one of the most commonly used statistical procedures, and are used as a basis in many applications such as exploratory data analysis, structural modeling, data engineering, etc. In this context, we present **correlation**, a toolbox for the R language (R Core Team 2019) and part of the **[easystats](https://github.com/easystats/easystats)** collection, focused on correlation analysis. Its goal is to be lightweight, easy to use, and allows for the computation of many different kinds of correlations, such as:

* **Pearson’s correlation**: This is the most common correlation method. It corresponds to the covariance of the two variables normalized (i.e., divided) by the product of their standard deviations.

rxy=cov(x,y)SDx×SDyrxy=cov(x,y)SDx×SDy

* **Spearman’s rank correlation**: A non-parametric measure of correlation, the Spearman correlation between two variables is equal to the Pearson correlation between the rank scores of those two variables; while Pearson’s correlation assesses linear relationships, Spearman’s correlation assesses monotonic relationships (whether linear or not). Confidence Intervals (CI) for Spearman’s correlations are computed using the Fieller, Hartley, and Pearson (1957) correction (see Bishara and Hittner 2017).

rsxy=cov(rankx,ranky)SD(rankx)×SD(ranky)rsxy=cov(rankx,ranky)SD(rankx)×SD(ranky)

* **Kendall’s rank correlation**: In the normal case, the Kendall correlation is preferred to the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient. However, the interpretation of Kendall’s tau is less direct compared to that of the Spearman’s rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events. Confidence Intervals (CI) for Kendall’s correlations are computed using the Fieller, Hartley, and Pearson (1957) correction (see Bishara and Hittner 2017). For each pair of observations (i ,j) of two variables (x, y), it is defined as follows:

τxy=2n(n−1)∑i<jsign(xi−xj)×sign(yi−yj)τxy=2n(n−1)∑i<jsign(xi−xj)×sign(yi−yj)

* **Biweight midcorrelation**: A measure of similarity that is median-based, instead of the traditional mean-based, thus being less sensitive to outliers. It can be used as a robust alternative to other similarity metrics, such as Pearson correlation (Langfelder and Horvath 2012).
* **Distance correlation**: Distance correlation measures both linear and non-linear association between two random variables or random vectors (for more, see Székely, Rizzo, and Bakirov (2007), Székely and Rizzo (2009)). This is in contrast to Pearson’s correlation, which can only detect linear association between two random variables.
* **Percentage bend correlation**: Introduced by Wilcox (1994), it is based on a down-weight of a specified percentage of marginal observations deviating from the median (by default, 20 percent).
* **Shepherd’s Pi correlation**: Equivalent to a Spearman’s rank correlation after outliers removal (by means of bootstrapped Mahalanobis distance).
* **Blomqvist’s coefficient**: The Blomqvist’s coefficient (also referred to as Blomqvist’s Beta or medial correlation; Blomqvist, 1950) is a median-based non-parametric correlation that has some advantages over measures such as Spearman’s or Kendall’s estimates (see Shmid and Schimdt, 2006).
* **Hoeffding’s D**: The Hoeffding’s D statistic is a non-parametric rank based measure of association that detects more general departures from independence (Hoeffding 1948), including non-linear associations. Hoeffding’s D varies between -0.5 and 1 (if there are no tied ranks, otherwise it can have lower values), with larger values indicating a stronger relationship between the variables.
* **Gamma correlation**: The Goodman-Kruskal gamma statistic is similar to Kendall’s Tau coefficient. It is relatively robust to outliers and deals well with data that have many ties.
* **Gaussian rank correlation**: The Gaussian rank correlation estimator is a simple and well-performing alternative for robust rank correlations (Boudt et al., 2012). It is based on the Gaussian quantiles of the ranks.
* **Point-Biserial and biserial correlation**: Correlation coefficient used when one variable is continuous and the other is dichotomous (binary). Point-Biserial is equivalent to a Pearson’s correlation, while Biserial should be used when the binary variable is assumed to have an underlying continuity. For example, anxiety level can be measured on a continuous scale, but can be classified dichotomously as high/low.
* **Winsorized correlation**: Correlation of variables that have been Winsorized, i.e., transformed by limiting extreme values to reduce the effect of possibly spurious outliers.
* **Polychoric correlation**: Correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables.
* **Tetrachoric correlation**: Special case of the polychoric correlation applicable when both observed variables are dichotomous.
* **Partial correlation**: Correlation between two variables after adjusting for the (linear) effect of one or more variables. The correlation test is run after having partialized the dataset, independently from it. In other words, it considers partialization as an independent step generating a different dataset, rather than belonging to the same model. This is why some discrepancies are to be expected for the *t*- and the *p*-values (but not the correlation coefficient) compared to other implementations such as ppcor. Let ex.zex.z be the residuals from the linear prediction of xx by zz (note that this can be expanded to a multivariate zz):

rxy.z=rex.z,ey.zrxy.z=rex.z,ey.z

* **Multilevel correlation**: Multilevel correlations are a special case of partial correlations where the variable to be adjusted for is a factor and is included as a random effect in a mixed-effects model.

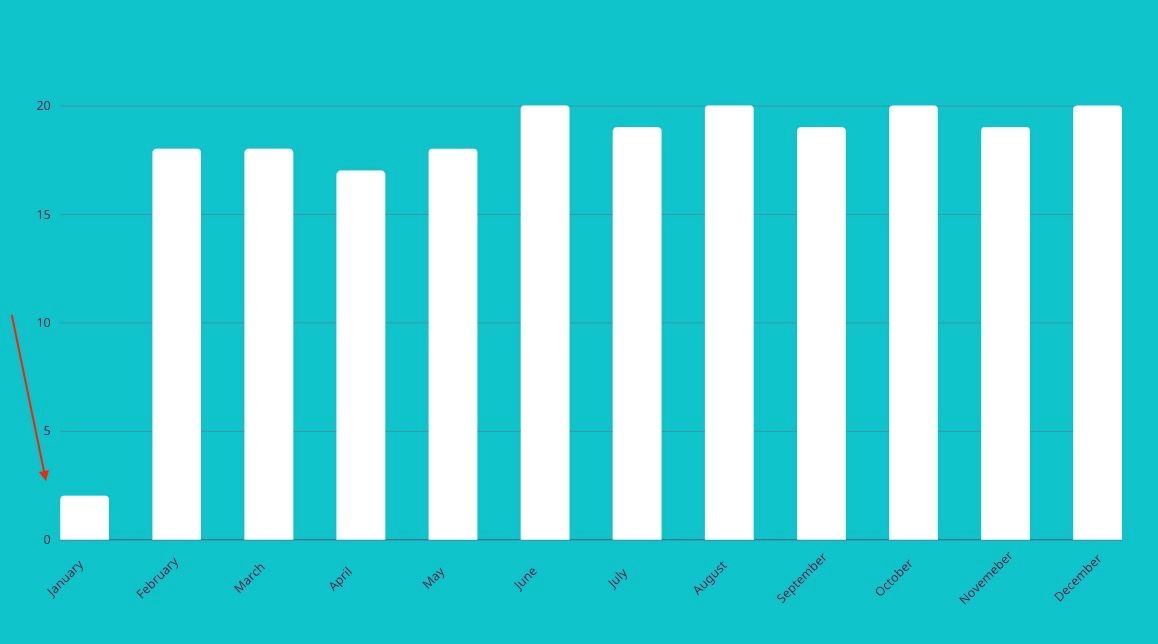
**What is an Outlier in Statistics?**

**Link :** [**https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/**](https://www.freecodecamp.org/news/what-is-an-outlier-definition-and-how-to-find-outliers-in-statistics/)

**Video link :** [**https://www.youtube.com/watch?v=7KeITQajazo&t=425s**](https://www.youtube.com/watch?v=7KeITQajazo&t=425s)

**(or)**

**2.** [**https://www.youtube.com/watch?v=rceisXY9v2s**](https://www.youtube.com/watch?v=rceisXY9v2s)

* **A Definition**
* In simple terms, an outlier is an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you're working with.
* Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.
* Below, on the far left of the graph, there is an outlier.
* The value in the month of January is significantly less than in the other months.
* 
* **How to Identify an Outlier in a Dataset**
* Alright, how do you go about finding outliers?
* An outlier has to satisfy either of the following two conditions:
* outlier < Q1 - 1.5(IQR)
* outlier > Q3 + 1.5(IQR)
* The rule for a low outlier is that a data point in a dataset has to be less than Q1 - 1.5xIQR.
* This means that a data point needs to fall more than 1.5 times the Interquartile range *below* the first quartile to be considered a low outlier.
* The rule for a high outlier is that if any data point in a dataset is more than Q3 - 1.5xIQR, it's a high outlier.
* More specifically, the data point needs to fall more than 1.5 times the Interquartile range *above* the third quartile to be considered a high outlier.
* As you can see, there are certain individual values you need to calculate first in a dataset, such as the IQR. But to find the IQR, you need to find the so called first and third quartiles which are Q1 and Q3 respectively.
* So, let's see what each of those does and break down how to find their values in both an odd and an even dataset.
* **How to Find the Upper and Lower Quartiles in an Odd Dataset**
* To get started, let's say that you have this dataset:
* 25,14,6,5,5,30,11,11,13,4,2
* The first step is to **sort the values in ascending numerical order**,from smallest to largest number.
* 2,4,5,5,6,11,11,13,14,25,30
* The lowest value (**MIN**) is 2 and the highest (**MAX**) is 30.
* **How to calculate**Q2**in an odd dataset**
* The next step is to find the **median** or *quartile 2 (Q2)*.
* This particular set of data has an odd number of values, with a total of 11 scores all together.
* To find the median in a dataset means that you're finding the middle value – the single middle number in the set.
* In odd datasets, there in only one middle number.
* Since there are 11 values in total, an easy way to do this is to split the set in two equal parts with each side containing 5 values.
* The median value will have 5 values on one side and 5 values on the other.
* (2,4,5,5,6), 11 ,(11,13,14,25,30)
* The median is 11 as it is the number that separates the first half from the second half.
* An alternative way to double check if you're right is to do this:
* (total\_number\_of\_scores + 1) / 2.
* This is (11 + 1) /2 = 6, which means you want the number in the 6th place of this set of data – which is 11.
* So Q2 = 11.
* **How to calculate**Q1**in an odd dataset**
* Next, to find the *lower quartile*, Q1, we need to find the median of the first half of the dataset, which is on the left hand side.
* As a reminder, the initial dataset is:
* (2,4,5,5,6), 11 ,(11,13,14,25,30)
* The first half of the dataset, or the *lower half*, does not include the median:
* 2,4,5,5,6
* This time, there is again an odd set of scores – specifically there are 5 values.
* You want to again split this half set into another half, with an equal number of two values on each side. You'll get a unique number, which will be the number in the middle of the 5 values.
* Pick the middle value that stands out:
* (2,4),5,(5,6)
* In this case it's Q1 = 5.
* To double check, you can also do total\_number\_of\_values + 1 / 2, similar to the previous example:
* (5 + 1) /2 = 3.
* This means you want the number in the 3rd place, which is 5.
* **How to calculate**Q3**in an odd dataset**
* To find the *upper quartile*, Q3, the process is the same as for Q1 above. But in this case you take the second half on the right hand side of the dataset, above the median and without the median itself included:
* (2,4,5,5,6), 11 ,(11,13,14,25,30)
* 11,13,14,25,30
* You split this half of the odd set of numbers into another half to find the median and subsequently the value of Q3.
* You again want the number in the 3rd place like you did for the first half.
* (11,13),14,(25,30)
* So Q3 = 14.
* **How to calculate**IQR**in an odd dataset**
* Now, the next step is to calculate the IQR which stands for Interquartile Range.
* This is the difference/distance between the lower quartile (Q1) and the upper quartile (Q3) you calculated above.
* As a reminder, the formula to do so is the following:
* IQR = Q3 - Q1
* To find the IQR of the dataset from above:
* IQR= 14 - 5
* IQR = 9
* **How to find an outlier in an odd dataset**
* To recap so far, the dataset is the one below:
* 2,4,5,5,6,11,11,13,14,25,30
* and so far, you have calucalted the five number summary:
* MIN = 2
* Q1 = 5
* MED = 11
* Q3 = 14
* MAX = 30
* Finally, let's find out if there are any outliers in the dataset.
* As a reminder, an outlier must fit the following criteria:
* outlier < Q1 - 1.5(IQR)
* Or
* outlier > Q3 + 1.5(IQR)
* To see if there is a lowest value outlier, you need to calculate the first part and see if there is a number in the set that satisfies the condition.
* Outlier < Q1 - 1.5(IQR)
* Outlier < 5 - 1.5(9)
* Outlier < 5 - 13.5
* outlier < - 8.5
* There are no lower outliers, since there isn't a number less than -8.5 in the dataset.
* Next, to see if there are any higher outliers:
* Outlier > Q3 + 1.5(IQR)=
* Outlier > 14 + 1.5(9)
* Outlier > 14 + 13.5
* Outlier > 27,5
* And there is a number in the dataset that is more than 27,5:
* 2,4,5,5,6,11,11,13,14,25,30
* In this case, 30 is the outlier in the existing dataset.
* **How to Find the Upper and Lower Quartiles in an Even Dataset**
* What happens when you have a dataset that consists of an even set of data?
* There isn't just one stand-out median (Q2), nor is there a standout upper quartile (Q1) or standout lower quartile (Q3).
* So the process of calculating quartiles and then finding an outlier is a bit different.
* **How to calculate**Q2**in an even dataset**
* Say that you have this dataset with 8 numbers:
* 10,15,20,26,28,30,35,40
* This time, the numbers are already sorted from lowest to highest value.
* To find the **median** number in an even dataset, you need to find the value that would be in between the *two* numbers that are in the middle. You add them together and divide them by 2, like so:
* 10,15,20,26,28,30,35,40
* 26 + 28 = 54
* 54 / 2 = 27
* **How to calculate**Q1**in an even dataset**
* To calculate to upper and lower quartiles in an even dataset, you keep all the numbers in the dataset (as opposed to in the odd set you removed the median).
* This time, the dataset is cut in half.
* 10,15,20,26 | 28,30,35,40
* To find Q1, you split the first half of the dataset into another half which leaves you with a remaining even set:
* 10,15 | 20,26
* To find the median of this half, you take the two numbers in the middle and divide them by two:
* Q1 = (15 + 20)/2
* Q1 = 35 / 2
* Q1 = 17,5
* **How to calculate**Q3**in an even dataset**
* To find Q3, you need to focus on the second half of the dataset and split that half into another half:
* 28,30,35,40 -> 28,30 | 35,40
* The two numbers in the middle are 30 and 35.
* You add them and divide them by two, and the result is:
* Q3 = (30 + 35)/2
* Q3 = 65 / 2
* Q3 = 32,5
* **How to calculate the**IQR**in an even dataset**
* The formula for calculating IQR is exactly the same as the one we used to calculate it for the odd dataset.
* IQR = Q3 - Q1
* IQR = 32,5 - 17,5
* IQR = 15
* **How to find an outlier in an even dataset**
* As a recap, so far the five number summary is the following:
* MIN = 10
* Q1 = 17,5
* MED = 27
* Q3 = 32,5
* MAX = 40
* To calculate any outliers in the dataset:
* outlier < Q1 - 1.5(IQR)
* Or
* outlier > Q3 + 1.5(IQR)
* To find any lower outliers, you calcualte Q1 - 1.5(IQR) and see if there are any values less than the result.
* outlier < 17,5 - 1.5(15)=
* outlier < 17,5 - 22,5
* outlier < -5
* There aren't any values in the dataset that are less than -5.
* Finally, to find any higher outliers, you calculate Q3 - 1.5(IQR) and see if there are any values in the dataset that are higher than the result
* outlier > 32.5 + 1.5(15)=
* outlier > 32.5 + 22.5
* outlier > 55
* There aren't any values higher than 55 so this dataset doesn't have any outliers.
* **Conclusion**
* In this article you learned how to find the interquartile range in a dataset and in that way calculate any outliers.

**How do we clean data**

**Link:** [**https://www.v7labs.com/blog/data-cleaning-guide**](https://www.v7labs.com/blog/data-cleaning-guide) **video :** [**https://www.youtube.com/watch?v=UWANxcHkKo8**](https://www.youtube.com/watch?v=UWANxcHkKo8) **(English) .**

**(or)**

**2.** [**https://www.youtube.com/watch?v=e-NxQPzkJxI**](https://www.youtube.com/watch?v=e-NxQPzkJxI) **(tamil)**

**Steps :**

[**https://www.youtube.com/watch?v=mwEPXevpqls**](https://www.youtube.com/watch?v=mwEPXevpqls)

*Garbage in, garbage out.*

This saying should become your mantra if you are serious about [building accurate machine learning models](https://www.v7labs.com/training) that find real-world applications.

And here's some food for thought—

Quality data beats even the most sophisticated algorithms.

Without clean data, your models will deliver misleading results and seriously harm your decision-making processes. You'll end up frustrated (been there, done that!), and it's simply not worth it.

Instead, let us walk you step-by-step through the data cleaning process.

Here’s what we’ll cover:

1. [What is data cleaning?](https://www.v7labs.com/blog/data-cleaning-guide#h1)
2. [The importance of data cleaning](https://www.v7labs.com/blog/data-cleaning-guide#h2)
3. [Data cleaning vs. data transformation](https://www.v7labs.com/blog/data-cleaning-guide#h3)
4. [5 characteristics of quality data](https://www.v7labs.com/blog/data-cleaning-guide#h4)
5. [How to clean data for Machine Learning?](https://www.v7labs.com/blog/data-cleaning-guide#h5)
6. [Data Cleaning best practices: Key Takeaways](https://www.v7labs.com/blog/data-cleaning-guide#h6)

hat is data cleaning?

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect.

This is generally data that can have a negative impact on the model or algorithm it is fed into by reinforcing a wrong notion.

Data cleaning not only refers to removing chunks of unnecessary data, but it’s also often associated with fixing incorrect information [within the train-validation-test dataset](https://www.v7labs.com/blog/train-validation-test-set) and reducing duplicates.

The importance of data cleaning

Data cleaning is a key step before any form of analysis can be made on it.

[Datasets in pipelines](https://www.v7labs.com/blog/best-free-datasets-for-machine-learning) are often collected in small groups and merged before being fed into a model. Merging multiple datasets means that redundancies and duplicates are formed in the data, which then need to be removed.

Also, incorrect and poorly collected datasets can often lead to models learning incorrect representations of the data, thereby reducing their decision-making powers.

It's far from ideal.

The reduction in model accuracy, however, is actually the *least* of the problems that can occur when unclean data is used directly.

Models trained on raw datasets are forced to take in noise as information and this can lead to accurate predictions when the noise is uniform within [the training and testing set](https://www.v7labs.com/blog/train-validation-test-set)—only to fail when new, cleaner data is shown to it.

Data cleaning is therefore an important part of any [machine learning pipeline](https://www.v7labs.com/blog/machine-learning-guide), and you should not ignore it.

Data cleaning vs. data transformation

As we’ve seen, data cleaning refers to the removal of unwanted data in the dataset before it’s fed into the model.

Data transformation, on the other hand, refers to the conversion or transformation of data into a format that makes processing easier.

In data processing pipelines, the incoming data goes through a data cleansing phase before any form of transformation can occur.  The data is then transformed, often going through stages like normalization and standardization before further processing takes place.

5 characteristics of quality data

Data typically has five characteristics that can be used to determine its quality.

These five characteristics are referred to within the data as:

* Validity
* Accuracy
* Completeness
* Consistency
* Uniformity

Besides checking up on these generic characteristics, there are still other specialized methods that [data scientists](https://www.v7labs.com/blog/data-science-interview-questions-and-answers) and data engineers use to check the quality of their data.

**Validity**

Data collection often involves a large group of people presenting their details in various forms (including phone numbers, addresses, and birthdays) in a document that is stored digitally.

Modern methods of data collection find validity an easy-to-maintain characteristic as they can control the data that is being entered into digital documents and forms.

Typical constraints applied on forms and documents to ensure data validity are:

* **Data-type constraints:** Data-type constraints help prevent inconsistencies arising due to incorrect data types in the wrong fields. Typically, these are found in fields like age, phone number, and name where the original data is constrained to contain only alphabetical or numerical values.
* **Range constraints:** Range constraints are applied in fields where prior information about the possible data is already present. These fields include—but are not limited to—date, age, height, etc.
* **Unique constraints:** Unique constraints are ones that update themselves each time a participant enters data into the document or form. This type of constraint prevents multiple participants from entering the same details for parameters that are supposed to be unique. Generally, these constraints are activated at fields like social security number, passport number, and username.
* **Foreign key constraints:**Foreign Key constraints are applicable to fields where the data can be limited to a set of previously decided keys. These fields are typically country and state fields where the range of information that can be provided is easily known beforehand.
* **Cross-field validation:** Cross-field validation is more of a check than a constraint to make sure that multiple fields in the document correspond to each other. For example, if the participant enters a group of values that should come to a particular number or amount, that amount serves as a validator that stops the participant from entering the wrong values.

**Accuracy**

Accuracy refers to how much the collected data is both feasible and accurate. It’s almost impossible to guarantee perfectly accurate data, thanks to the fact that it contains personal information that’s only available to the participant. However, we can make near-accurate assumptions by observing the feasibility of that data.

Data in the form of locations, for example, can easily be cross-checked to confirm whether the location exists or not, or if the postal code matches the location or not. Similarly, feasibility can be a solid criterion for judging. A person cannot be 100 feet tall, nor can they weigh a thousand pounds, so data going along these lines can be easily rejected.**Completeness**

Completeness refers to the degree to which the entered data is present in its entirety.

Missing fields and missing values are often impossible to fix, resulting in the entire data row being dropped. The presence of incomplete data, however, can be appropriately fixed with the help of proper constraints that prevent participants from filling up incomplete information or leaving out certain fields.

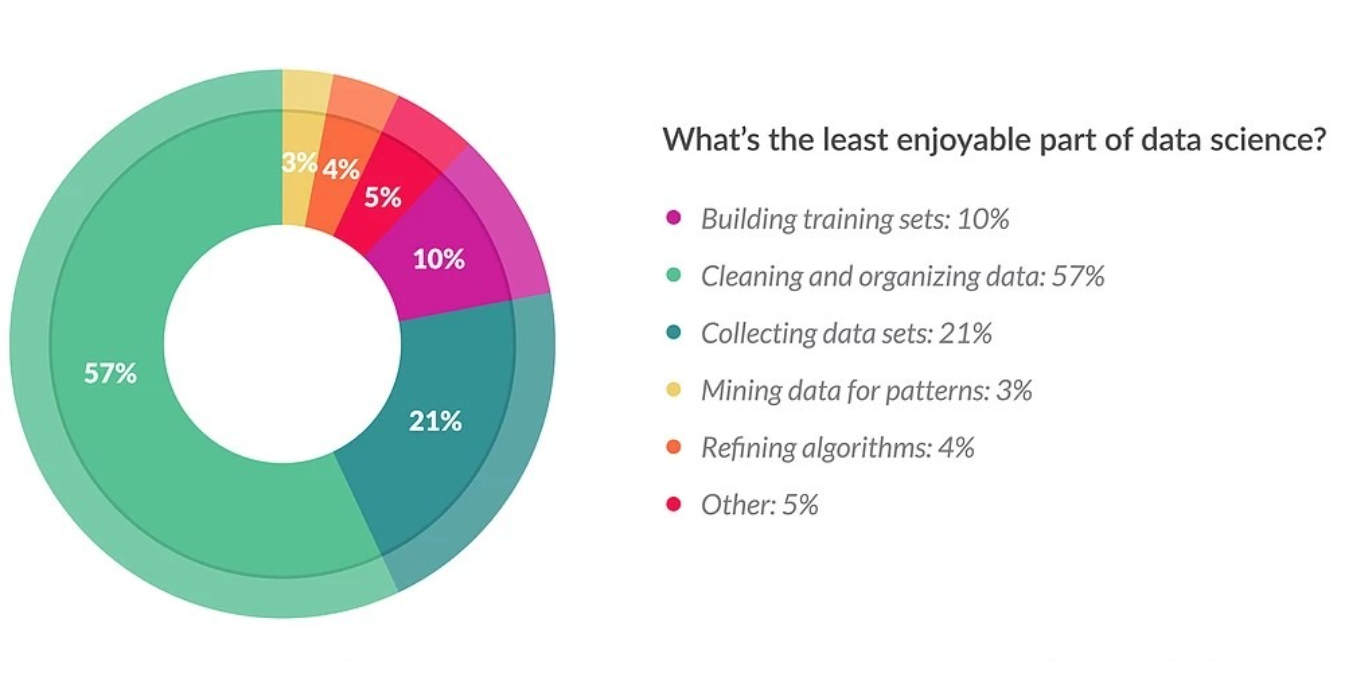
**Consistency**

Consistency refers to how the data responds to cross-checks with other fields. Studies are often held where the same participant fills out multiple surveys which are cross-checked for consistency. Cross checks are also included for the same participant in more than a single field.

How to clean data for Machine Learning?

As research suggests—

Data cleaning is often the least enjoyable part of data science—and also the *longest*.



Indeed, cleaning data is an arduous task that requires manually combing a large amount of data in order to:

a) reject irrelevant information.

b) analyze whether a column needs to be dropped or not.

Automation of the cleaning process usually requires a an extensive experience in dealing with dirty data. It’s kinda tricky to implement in a manner that doesn’t bring about data loss.

Re**move duplicate or irrelevan**t data

Data that’s processed in the form of data frames often has duplicates across columns and rows that need to be filtered out.

Duplicates can come about either from the same person participating in a survey more than once or the survey itself having multiple fields on a similar topic, thereby eliciting a similar response in a large number of participants.

While the latter is easy to remove, the former requires investigation and algorithms to be employed. Columns in a data frame can also contain data highly irrelevant to the task at hand, resulting in these columns being dropped before the data is processed further.

Fix syntax errors

Data collected over a survey often contains syntactic and grammatical issues, due mainly to the fact that a huge demographic is represented through it. Common syntax issues like date, birthday and age are simple enough to fix, but syntax issues involving spelling mistakes require more effort.

Algorithms and methods which find and fix these errors have to be employed and iterated through the data for the removal of typos and grammatical and spelling mistakes.

Syntax errors, meanwhile, can be prevented altogether by structuring the format in which data is collected, before running checks to ensure that the participants have not wrongly filled in known fields. Setting strict boundaries for fields like State, Country, and School goes a long way to ensuring quality data.

Filter out unwanted outliers

Unwanted data in the form of outliers has to be removed before it can be processed further. Outliers are the hardest to detect amongst all other inaccuracies within the data.

Thorough analysis is generally conducted before a data point or a set of data points can be rejected as an outlier. Specific models that have a very low outlier tolerance can be easily manipulated by a good number of outliers, therefore bringing down the prediction quality.

Handle missing data

Unfortunately, missing data is unavoidable in poorly designed data collection procedures. It needs to be identified and dealt with as soon as possible. While these artifacts are easy to identify, filling up missing regions often requires careful consideration, as random fills can have unexpected outcomes on the model quality.

Often, rows containing missing data are dropped as it’s not worth the hassle to fill up a single data point accurately. When multiple data points have missing data for the same attributes, the entire column is dropped. Under completely unavoidable circumstances and in the face of low data, data scientists have to fill in missing data with calculated guesses.

These calculations often require observation of two or more data points similar to the one under scrutiny and filling in an average value from these points in the missing regions.

Validate data accuracy

Data accuracy needs to be validated via cross-checks within data frame columns to ensure that the data which is being processed is as accurate as possible. Ensuring the accuracy of data is, however, hard to gauge and is possible only in specific areas where a predefined idea of the data is known.

Fields like countries, continents, and addresses can only have a set of predefined values that can be easily validated against. In data frames constructed from more than a single source/survey, cross-checks across sources can be another procedure to validate data accuracy.

Covariance vs Correlation: What’s the difference?

Link : <https://www.mygreatlearning.com/blog/covariance-vs-correlation/>

Video link : <https://www.youtube.com/watch?v=Ifrr-bB4bbo> (tamil)

2. <https://www.youtube.com/watch?v=wDAd_QHKoOg> (mean,variance,sd)

3. <https://www.youtube.com/watch?v=jmhtmVB8Hao> (TAMIL)

In statistics, covariance and correlation are two mathematical notions. Both phrases are used to describe the relationship between two variables. This blog talks about covariance vs correlation: what’s the difference? Let’s get started!

* [Introduction](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#introduction)
* [What is covariance?](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#what-is-covariance)
* [What is correlation?](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#what-is-correlation)
* [Applications of covariance](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#applications-of-covariance)
* [Applications of correlation](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#applications-of-correlation)
* [Methods of calculating the correlation](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#methods-of-calculating-the-correlation)
* [Variance](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#variance)
* [Standard Deviation](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#standard-deviation)
* [Differences between Covariance and Correlation](https://www.mygreatlearning.com/blog/covariance-vs-correlation/#differences-between-covariance-and-correlation)

**Introduction**

Covariance and correlation are two mathematical concepts used in statistics. Both terms are used to describe how two variables relate to each other. Covariance is a measure of how two variables change together. The terms covariance vs correlation is very similar to each other in probability theory and statistics. Both the terms describe the extent to which a random variable or a set of random variables can deviate from the expected value. But what is the difference between covariance vs correlation? Let’s understand this by going through each of these terms.

It is calculated as the covariance of the two variables divided by the product of their standard deviations. Covariance can be positive, negative, or zero. A positive covariance means that the two variables tend to increase or decrease together. A negative covariance means that the two variables tend to move in opposite directions.

A zero covariance means that the two variables are not related. Correlation can only be between -1 and 1. A correlation of -1 means that the two variables are perfectly negatively correlated, which means that as one variable increases, the other decreases. A correlation of 1 means that the two variables are perfectly positively correlated, which means that as one variable increases, the other also increases. A correlation of 0 means that the two variables are not related.  
  
In [statistics](https://www.mygreatlearning.com/blog/inferential-statistics-an-overview/), it is frequent that we come across these two terms known as covariance and correlation. The two terms are often used interchangeably. These two ideas are similar, but not the same. Both are used to determine the linear relationship and measure the dependency between two random variables. But are they the same? **Not really.**

Despite the similarities between these mathematical terms, they are different from each other.

Covariance is when two variables vary with each other, whereas Correlation is when the change in one variable results in the change in another variable.

In this article, we will try to define the terms correlation and covariance matrices, talk about covariance vs correlation, and understand the application of both terms.

**What is covariance?**

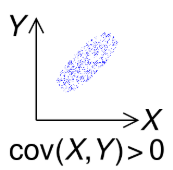
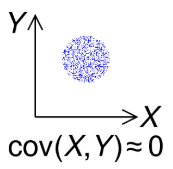
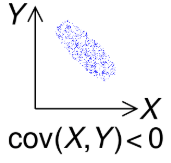
Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The values of covariance can be any number between the two opposite infinities. Also, it’s important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:



The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.



**Example:**

|  |  |
| --- | --- |
| **X** | **Y** |
| **10** | **40** |
| **12** | **48** |
| **14** | **56** |
| **8** | **32** |

**Step 1: Calculate Mean of X and Y**

Mean of X ( μx ) : 10+12+14+8 / 4 =  11

Mean of Y(μy) = 40+48+56+32 = 44

**Step 2: Substitute the values in the formula**



|  |  |
| --- | --- |
| **xi –x̅** | **yi – ȳ** |
| 10 – 11 = -1 | 40 – 44 = – 4 |
| 12 – 11 = 1 | 48  – 44 = 4 |
| 14 – 11 = 3 | 56 – 44 = 12 |
| 8 – 11 = -3 | 32 – 44 = 12 |

**Substitute the above values in the formula**

Cov(x,y) = (-1) (-4) +(1)(4)+(3)(12)+(-3)(12)

                  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

                                            4

**Cov(x,y) =** 8/2 =**4**

**Hence, Co-variance for the above data is 4**

**What is correlation?**

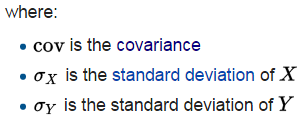
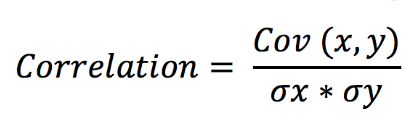
Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

To do so we have to normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

The main result of a correlation is called the correlation coefficient.



The correlation coefficient is a dimensionless metric and its value ranges from -1 to +1.

The closer it is to +1 or -1, the more closely the two variables are related.

If there is no relationship at all between two variables, then the correlation coefficient will certainly be 0. However, if it is 0 then we can only say that there is no linear relationship. There could exist other functional relationships between the variables.

When the correlation coefficient is positive, an increase in one variable also increases the other. When the correlation coefficient is negative, the changes in the two variables are in opposite directions.

**Example:**

|  |  |
| --- | --- |
| **X** | **Y** |
| **10** | **40** |
| **12** | **48** |
| **14** | **56** |
| **8** | **32** |

**Step 1: Calculate Mean of X and Y**

Mean of X ( μx ) : 10+12+14+8 / 4 =  11

Mean of Y(μy) = 40+48+56+32/4 = 44

**Step 2: Substitute the values in the formula**



|  |  |
| --- | --- |
| **xi –x̅** | **yi – ȳ** |
| 10 – 11 = -1 | 40 – 44 = – 4 |
| 12 – 11 = 1 | 48  – 44 = 4 |
| 14 – 11 = 3 | 56 – 44 = 12 |
| 8 – 11 = -3 | 32 – 44 = 12 |

**Substitute the above values in the formula**

Cov(x,y) = (-1) (-4) +(1)(4)+(3)(12)+(-3)(12)

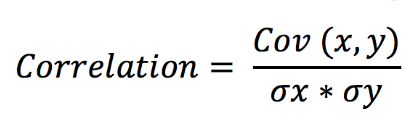
                  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

                                            4

**Cov(x,y) =** 8/2 =**4**

**Hence, Co-variance for the above data is 4**

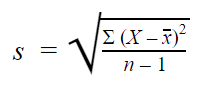
**Step 3: Now substitute the obtained answer in Correlation formula**



Before substitution we have to find standard deviation of x and y

Lets take the data for X as mentioned in the table that is 10,12,14,8

To find standard deviation



**Step 1: Find the mean of x that is x̄**

 10+14+12+8 /4 = 11

**Step 2: Find each number deviation: Subtract each score with mean to get mean deviation**

|  |
| --- |
| 10 – 11 = -1 |
| 12 – 11 = 1 |
| 14 – 11 = 3 |
| 8 – 11 = -3 |

**Step 3: Square the mean deviation obtained**

|  |  |
| --- | --- |
| **-1** | **1** |
| **1** | **1** |
| **3** | **9** |
| **-3** | **9** |

**Step 4: Sum the squares**

1+1+9+9 = 20

**Step5: Find the variance**

**Divide the sum of squares with n-1 that is 4-1 = 3**

20 /3 = 6.6

**Step 6: Find the square root**

Sqrt of 6.6 = 2.581

**Therefore, Standard Deviation of x = 2.581**

**Find for Y using same method**

The Standard Deviation of y = 10.29

Correlation = 4 /(**2.581** x10.29 )

Correlation = 0.15065

So, now you can understand the difference between Covariance vs Correlation.

**Applications of covariance**

1. Covariance is used in Biology – Genetics and Molecular Biology to measure certain DNAs.
2. Covariance is used in the prediction of amount investment on different assets in financial markets
3. Covariance is widely used to collate data obtained from astronomical /oceanographic studies to arrive at final conclusions
4. In Statistics to analyze a set of data with logical implications of principal component we can use covariance matrix
5. It is also used to study signals obtained in various forms.

**Applications of correlation**

1. Time vs Money spent by a customer on online e-commerce websites
2. Comparison between the previous records of weather forecast to this current year.
3. Widely used in pattern recognition
4. Raise in temperature during summer  v/s water consumption amongst family members is analyzed
5. The relationship between population and poverty is gauged

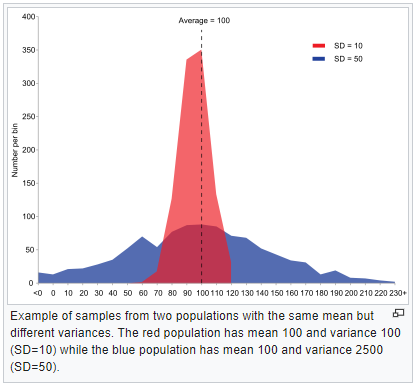
**Methods of calculating the correlation**

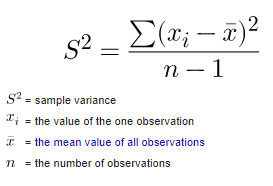
1. The graphic method
2. The scatter method
3. Co-relation Table
4. Karl Pearson  Coefficient of Correlation
5. Coefficient of Concurrent deviation
6. Spearman’s rank correlation coefficient

Before going into the details, let us first try to understand variance and standard deviation.

**Variance**

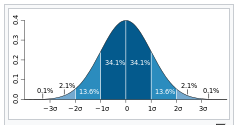
Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers are spread out from their average value.

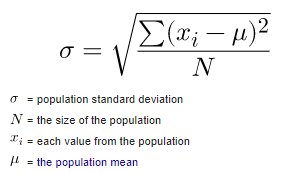




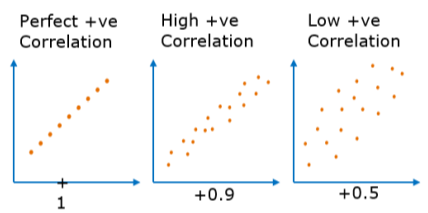
**Standard Deviation**

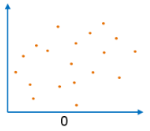
Standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range. It essentially measures the absolute variability of a random variable.



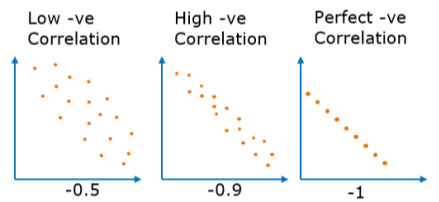


Covariance and correlation are related to each other, in the sense that covariance determines the type of interaction between two variables, while correlation determines the direction as well as the strength of the relationship between two variables.





|  |  |
| --- | --- |
| Covariance | Correlation |
| Covariance is a measure to indicate the  extent to which two random  variables change in tandem. | Correlation is a measure used to represent how strongly two random variables are related to each other. |
| Covariance is nothing but a  measure of correlation. | Correlation refers to the scaled form of covariance. |
| Covariance indicates the direction  of the linear relationship between variables. | Correlation on the other hand measures both the strength and direction of the linear relationship between two variables. |
| Covariance can vary between -∞ and +∞ | Correlation ranges between -1 and +1 |
| Covariance is affected by the change in scale.  If all the values of one variable  are multiplied by a constant  and all the values of another variable are multiplied, by a similar or different constant, then the covariance is changed. | Correlation is not influenced by the change in scale. |
| Covariance assumes the units from the product of the units of the two variables. | Correlation is dimensionless, i.e. It’s a unit-free measure of the relationship between variables. |
| Covariance of two dependent variables measures how much in real quantity (i.e. cm, kg, liters) on average they co-vary. | Correlation of two dependent variables measures the proportion of how much on average these variables vary w.r.t one another. |
| Covariance is zero in case of independent variables (if one variable moves and the other doesn’t) because then the variables do not necessarily move together. | Independent movements do not contribute to the total correlation. Therefore, completely independent variables have a zero correlation. |



**Differences between Covariance and Correlation**

Both the Covariance and Correlation metrics evaluate two variables throughout the entire domain and not on a single value. The differences between them are summarized in a tabular form for quick reference. Let us look at Covariance vs Correlation.

**Conclusion**

Both Correlation and Covariance are very closely related to each other and yet they differ a lot.

When it comes to choosing between Covariance vs Correlation, the latter stands to be the first choice as it remains unaffected by the change in dimensions, location, and scale, and can also be used to make a comparison between two pairs of variables. Since it is limited to a range of -1 to +1, it is useful to draw comparisons between variables across domains. However, an important limitation is that both these concepts measure the only linear relationship.

**Types of data distributions**

**What are the Types of Data Distribution/Statistical Distribution Models?**

**Link :** [**https://www.analyticssteps.com/blogs/10-types-statistical-data-distribution-models**](https://www.analyticssteps.com/blogs/10-types-statistical-data-distribution-models)

**Video link :** [**https://www.youtube.com/watch?v=b9a27XN\_6tg**](https://www.youtube.com/watch?v=b9a27XN_6tg)

1. **Bernoulli’s Distribution**

This is one of the simplest distributions that can be used as an initial point to derive more complex distributions. Bernoulli’s distribution has possibly two outcomes (success or failure) and a single trial.

For example, tossing a coin, the success probability of an outcome to be heads is p, then the probability of having tail as outcome is (1-p). Bernoulli’s distribution is the special case of binomial distribution with a single trial.

The density function can be given as

**f(x) = px  (1-p)(1-x)   where x € (0,1)**

It can also be written as;

The graph of Bernoulli's distribution is shown below where the probability of success is less than probability of failure.

*Bernoulli’s Distribution*

The distribution has following characteristics;

* The number of trials, to be performed, need to be predefined for a single experiment.
* Each trial has only two possible outcomes-success or failure.
* The probability of success of each event/experiment must be the same.
* Each event must be independent of each other.

1. **Binomial Distribution**

The binomial distribution is applied in binary outcomes events where the probability of success is equal to the probability of failure in all the successive trials. Its example includes tossing a biased/unbiased coin for a repeated number of times.

As input, the distribution considers two parameters, and is thus called as bi-parametric distribution. The two parameters are;

* The number of times an event occurs, n, and
* Assigned probability, p, to one of the two classes

For n number of trials, and success probability, p, the probability of successful event (x) within n trials can be determined by the following formula

The graph of binomial distribution is shown below when the probability of success is equal to probability of failure.

*Binomial distribution*

The binomial distribution holds the following properties;

* For multiple trials provided, each trial is independent to each other, i.e, the result of one trial cannot influence other trials.
* Each of the trials can have two possible outcomes, either success or failure, with probabilities p, and (1-p).
* A total number of n identical trials can be conducted, and the probability of success and failure is the same for all trials.

1. **Normal (Gaussian) Distribution**

Being a continuous distribution, the normal distribution is most commonly used in data science. A very common process of our day to day life belongs to this distribution- income distribution, average employees report, average weight of a population, etc.

The formula for normal distribution;

Where μ = Mean value,

σ = Standard probability distribution of probability,

x = random variable

According to the formula,  the distribution is said to be normal if mean (μ) = 0 and standard deviation (σ) = 1

The graph of normal distribution is shown below which is symmetric about the centre (mean).

*Normal distribution*

Normal distribution has the following properties;

* Mean, mode and median coincide with each other.
* The distribution has a bell-shaped distribution curve.
* The distribution curve is symmetrical to the centre.
* The area under the curve is equal to 1.

1. **Poisson Distribution**

Being a part of discrete probability distribution, poisson distribution outlines the probability for a given number of events that take place in a fixed time period or space, or particularized intervals such as distance, area, volume.

For example, conducting risk analysis by the insurance/banking industry, anticipating the number of car accidents in a particular time interval and in a specific area.

Poisson distribution considers following assumptions;

* The success probability for a short span is equal to success probability for a long period of time.
* The success probability in a duration equals to zero as the duration becomes smaller.
* A successful event can’t impact the result of another successful event

A poisson distribution can be modeled using the formula below,

Where 𝝺 represents the possible number of events take place in a fixed period of time, and X is the number of events in that time period.

The graph of poisson distribution is shown below;

*Poisson distribution*

Poisson distribution has the following characteristics;

* The events are independent of each other, i.e, if an event occurs, it doesn’t affect the probability of another event occurring.
* An event could occur any number of times in a defined period of time.
* Any two events can’t be occurring at the same time.
* The average rate of events to take place is constant.

1. **Exponential Distribution**

Like the poisson distribution, exponential distribution has the time element; it gives the probability of a time duration before an event takes place.

Exponential distribution is used for survival analysis, for example, life of an air conditioner, expected life of a machine,and length of time between metro arrivals.

A variable X is said to possess an exponential distribution when

Where  λ stands for rate and always has value greater than zero.

The graph of exponential distribution is shown below;

*Exponential distribution*

The exponential distribution has following characteristics;

* As shown in the graph, the higher the rate, the faster the curve drops, and lower the rate, flatter the curve.
* In survival analysis, λ is termed as a failure rate of a machine at any time t with the assumption that the machine will survive upto t time.

1. **Multinomial Distribution**

The multinomial distribution is used to measure the outcomes of experiments that have two or more variables. It is the special type of binomial distribution when there are two possible outcomes such as true/false or success/failure.

The distribution is commonly used in biological, geological and financial applications.

A very popular Mendel experiment where two strains of peas (one green and wrinkled seeds and other is yellow and smooth seeds) are hybridized that produced four different strains of seeds-green and wrinkled, green and round, yellow and round, and yellow and wrinkled. This resulted in multinomial distribution and led to the discovery of the basic principles of genetics.

The density function for multinomial distribution is

Where n= number of experiments.

Px= probability of occurrence of an experiment.

The graph of exponential distribution is shown below;

*Multinomial Distribution*

The following are properties of multinomial distribution;

* An experiment can have a repeated number of trials, for example, rolling of  a dice multiple times.
* Each trial is independent of each other.
* The success probability of each outcome must be the same (constant) for all trials of an experiment.

1. **Beta Distribution**

Beta distribution comes under continuous probability distributions having the interval [0,1] with two shape parameters that can be expressed by alpha (ɑ) and beta(ꞵ). These two parameters are the exponent of a random variable and control the shape of the distribution.

The distribution shows the family of probabilities and is a suitable model to depict random behaviour of percentages or proportions. It is used for the data models that hold uncertainties of the success probabilities in a random experiment.

The probability density function for the beta distribution is

Where 𝝱 is second shape parameter and B( ɑ, ꞵ) is normalizing constant that makes sure area under the curve is one.

The graph of beta distribution is shown below;

*Beta Distribution*

The general formulation of beta distribution is also known as the beta distribution of first kind and beta distribution of second kind is another name of beta prime distribution.

Beta distribution has many applications in statistical description of allele frequencies in genetic population, time allocation in project management, sunshine data, proportions of minerals in rocks, etc.

(Referred blog: [Conditional Probability](https://www.analyticssteps.com/blogs/conditional-probability-definition-properties-examples))

1. **Beta-binomial distribution**

A data distribution is said to be beta-binomial if the

* Probability of success, p, is greater than zero.
* And, shape of beat binomial parameter, α > 0, as well as β > 0

Being the simplest form of Bayesian mode, beta-binomial distribution has extensive applications in intelligence testing, epidemiology, and marketing.

The graph of beta-binomial distribution looks as below;

*Beta-binomial distribution*

The parametric shape can be defined in the form of  the probability of success such that

* A distribution tends to a binomial distribution for the greater value of α and β.
* The value of discrete uniform distribution is equivalent to the distribution between 0 to n, if both the values α = β = 1.
* For n = 1, the beta-binomial distribution is approximately the same as Bernoulli distribution.

Talking about the key difference amid a beta-distribution and binomial distribution, the success probability, p, is always fixed for a set of trials whereas it is not fixed for beta-binomial distribution and changes trail to trail.

1. **T- Distributions**

In [statistics](https://www.analyticssteps.com/blogs/what-statistics-types-variance-bayesian-statistics), t-distribution is the most important distribution, also known as student’s t-distribution. It is employed to estimate population parameters when the sample size is small, and the [standard deviation](https://www.analyticssteps.com/blogs/what-standard-deviation) is unknown.

It is widely used for hypothesis testing and built confidence intervals for mean values. The graph of t-distribution distribution is shown below;

*T-distribution*

T-distribution has the following properties;

* Similar to normal distribution, the t-distribution has bell-shaped curve distribution and is symmetric when mean is zero.
* The shape of distribution doesn’t alter with degrees of freedom, and has the range – ∞ to ∞.
* The variance is always more than one.
* As the sample size, n, increases, t-distribution acts as normal distribution where the considered sample size is greater than 30.

(Must check: [T-test vs Z-test](https://www.analyticssteps.com/blogs/what-are-differences-between-z-test-and-t-test))

1. **Uniform distribution**

Uniform distribution can either be discrete or continuous where each event is equally likely to occur. It has a constant probability constructing a rectangular distribution.

In this type of distribution, an unlimited number of outcomes will be possible and all the events have the same probability, similar to Bernoulli’s distribution. For example, while rolling a dice, the outcomes are 1 to 6 that have equal probabilities of ⅙ and represent a uniform distribution.

A variable X is said to have uniform distribution if the probability density function is

The graph of a uniform distribution looks as below

*Uniform distribution*

The uniform distribution has the following properties;

* The probability density function combines to unity.
* Every input function has an equal weightage.

**Multicollinearity in Data Science**

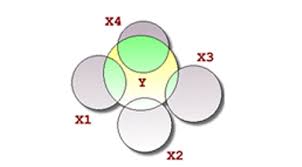
**Link :** [**https://www.analyticsvidhya.com/blog/2021/03/multicollinearity-in-data-science/**](https://www.analyticsvidhya.com/blog/2021/03/multicollinearity-in-data-science/)

**Video link :** [**https://www.youtube.com/watch?v=tcaruVHXZwE**](https://www.youtube.com/watch?v=tcaruVHXZwE)

**Introduction**

Multicollinearity is a topic in Machine Learning of which you should be aware. I know this topic since from past years I have dive into the Statistics concept which is important for all those who are do something in the field of Data Science. I have seen a lot of Data Science peoples who are professional but they don’t know some stuff related to multicollinearity.

This is especially important for all those peoples who are coming from a non-mathematical background or those who have not more knowledge of Statistical concepts. It is not just learning a topic in Data Science but It is important when you are trying to crack the Data Science Interviews and finding insights from the data on which we have to apply the ML Algorithms.



So, in this article, we will see what is multicollinearity, why it is a problem, what causes multicollinearity, and then try to understand what is the illness in the model where multicollinearity exists and finally we see some techniques to remove the multicollinearity.

I highly recommend that before diving further into much deeper concepts, it is good if you have a basic understanding of a Regression model and some statistical terms.

**Table of Contents**

1. What is Multicollinearity?

2. The problem with having multicollinearity

3. What causes multicollinearity?

4. How to detect multicollinearity by using VIF?

5. Test for detection for multicollinearity

6. Solutions for multicollinearity

**What is Multicollinearity?**

Multicollinearity occurs when two or more independent variables(also known as predictor) are highly correlated with one another in a regression model.

This means that an independent variable can be predicted from another independent variable in a regression model. For Example, height, and weight, student consumption and father income, age and experience, mileage and price of a car, etc.

Let us take a simple example from our everyday life to explain this. Assume that we want to fit a regression model using the independent features such as pocket money and father income, to find the dependent variable, Student consumption here we cannot find an exact or individual effect of all the independent variables on the dependent variable or response since here both independent variables are highly correlated means as father income increases pocket money also increases and if father income decreases pocket money also decreases.

This is the multicollinearity problem!

**The problem with having multicollinearity**

Since in a regression model our research objective is to find out how each predictor is impacting the target variable individually which is also an assumption of a method namely Ordinary Least Squares through which we can find the parameters of a regression model.

So finally to fulfill our research objective for a regression model we have to fix the problem of multicollinearity which is finally important for our prediction also.

Let say we have the following linear equation

**Y=a0+a1\*X1+a2\*X2**

Here X1 and X2 are the independent variables. The mathematical significance of a1 is that if we shift our X1 variable by 1 unit then our Y shifts by a1 units keeping X2 and other things constant. Similarly, for a2 we have if we shift X2 by one unit means Y also shifts by one unit keeping X1 and other factors constant.

But for a situation where multicollinearity exists our independent variables are highly correlated, so if we change X1 then X2 also changes and we would not be able to see their Individual effect on Y which is our research objective for a regression model.

“ This makes the effects of X1 on Y difficult to differentiate  from the effects of X2 on Y. ”

**Note:**

Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of individual independent features on the dependent feature in your model and that can be a problem when we want to interpret your model.

**What causes multicollinearity?**

Multicollinearity might occur due to the following reasons:

1*.*Multicollinearity could exist because of the problems in the dataset at the time of creation. These problems could be because of poorly designed experiments, highly observational data, or the inability to manipulate the data. (This is known as **Data related multicollinearity**)

For example, determining the electricity consumption of a household from the household income and the number of electrical appliances. Here, we know that the number of electrical appliances in a household will increase with household income. However, this cannot be removed from the dataset.

2. Multicollinearity could also occur when new variables are created which are dependent on other variables(Basically when we do the data preprocessing or feature engineering to make the new feature from the existing features . This is known as **Structure related multicollinearity**)

For example, creating a variable for BMI from the height and weight variables would include redundant information in the model. (Since BMI depend on the height and the weight itself)

3. When there are identical variables in the dataset.

For example, including the weight of a person in kilograms and another for weight in grams or some other units.

4. When we want to encode the categorical features to numerical features for applying the machine learning algorithms since ML algorithms only understand numbers not text. So for this task, we use the concept of the Dummy variable. Inaccurate use of Dummy variables can also cause multicollinearity. (This is known as Dummy Variable Trap)

For example, in a dataset containing the status of marriage variable with two unique values: ‘married’, ’single’. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for ‘married’/’single’ status.

1. Insufficient data in some cases can also cause multicollinearity problems.