1. **What is data normalization? How is it different from data normalization?**

Normalization is a pre-processing stage of any type of problem statement.

Data transformation operations, such as normalization and aggregation, are additional data preprocessing procedures that would contribute toward the success of the data extract process.

**Data normalization**

Data normalization consists of remodeling numeric columns to a standard scale. Data normalization is generally considered the development of clean data. Diving deeper, however, the meaning or goal of data normalization is twofold:

* Data normalization is the organization of data to appear similar across all records and fields.
* It increases the cohesion of entry types, leading to cleansing, lead generation, segmentation, and higher quality data.

**Importance of Data Normalization**

* Data Normalization disposes of various anomalies that can make an examination of the information more complicated. A portion of those irregularities can manifest from erasing information, embedding more data, or refreshing existing data. Once those mistakes are worked out and eliminated from the framework, further advantages can be acquired through different jobs in the data and data examination.
* It is for the most part through data normalization that the data inside a data set can be designed so that it can be visualized and examined.

## **Database Normalization**

**Database normalization** is a technique for [creating database tables](https://phoenixnap.com/kb/how-to-create-a-table-in-mysql) with suitable columns and keys by decomposing a large table into smaller logical units. The process also considers the demands of the environment in which [the database](https://phoenixnap.com/kb/what-is-a-database) resides.

Normalization is an **iterative process.** Commonly, normalizing a database occurs through a series of tests. Each subsequent step decomposes tables into more manageable information, making the overall database logical and easier to work with.

**Importance of data normalization**

Normalization helps a database designer optimally distribute attributes into tables. The technique eliminates the following:

* Attributes with **multiple** values.
* Doubled or **repeated** attributes.
* **Non-descriptive** attributes.
* Attributes with **redundant** information.
* Attributes created from **other features**.

Although total database normalization is not necessary, it provides a well-functioning information environment. The method systematically ensures:

* **A database structure** suitable for generalized queries.
* **Minimized data redundancy**, increasing memory efficiency on a [database server](https://phoenixnap.com/servers/database).
* **Maximized data integrity** through the reduced insert, update, and delete anomalies.

### 1. First Normal Form –

If a relation contain composite or multi-valued attribute, it violates first normal form or a relation is in first normal form if it does not contain any composite or multi-valued attribute. A relation is in first normal form if every attribute in that relation is **singled valued attribute**.

### 2. Second Normal Form –

To be in second normal form, a relation must be in first normal form and relation must not contain any partial dependency. A relation is in 2NF if it has **No Partial Dependency,**i.e.**,**no non-prime attribute (attributes which are not part of any candidate key) is dependent on any proper subset of any candidate key of the table.

**Partial Dependency –** If the proper subset of candidate key determines non-prime attribute, it is called partial dependency.

**3. Third Normal Form –**

A relation is in third normal form, if there is **no transitive dependency** for non-prime attributes as well as it is in second normal form.  
A relation is in 3NF if **at least one of the following condition holds** in every non-trivial function dependency X –> Y

* 1. X is a super key.
  2. Y is a prime attribute (each element of Y is part of some candidate key).

**Transitive dependency –** If A->B and B->C are two FDs then A->C is called transitive dependency.

1. **What is a distribution? What are the uses for frequency and probability distribution?**

A distribution in statistics is a function that shows the possible values for a variable and how often they occur.

Frequency distribution in statistics represents the number of times a value or an outcome repeats itself over a period of time during an event. It is featured either as a graph or table. This representation helps data analysts assess the frequency of a value over an interval during an instance and determine the possibility of such outcomes occurring again under similar circumstances.

distributions are important for statistics because we need to collect the sample and estimate the parameters of the population distribution. Hence distribution is necessary to make inferences about the overall population.

A probability distribution is an essential feature of statistics and provides a way to**describe random variables**. As such, it is used in every field that uses statistics. In finance, for example, they are used to calculate expected returns of stocks and hedge risks.

 In social sciences, they are widely used in experiments, surveys, and modeling. In physics and chemistry, they are used to describe physical objects’ properties and in astronomy to describe waves.

Probability distributions are also used in several other branches of mathematics. For example, stochastic differential equations are the area of mathematics that studies differential equations with random variables. In game theory, Poisson games study situations with a random number of players.

1. **What is decision? How is it different from inference?**

In statistics, a set of quantitative methods for reaching optimal decisions. A solvable decision problem must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their consequences.

The ladder of inference is a tool that explains how we make decisions. In computability theory and computational complexity theory, a decision problem is a problem that can be posed as a yes–no question of the input values. An example of a decision problem is deciding whether a given natural number is prime.

**4. what is Gini in probability and explain in your own terms?**

The Gini coefficient (Gini index or Gini ratio) is a statistical measure of economic inequality in a population. The coefficient measures the dispersion of [income](https://corporatefinanceinstitute.com/resources/careers/compensation/remuneration/) or distribution of wealth among the members of a [population](https://corporatefinanceinstitute.com/resources/knowledge/economics/demographics/).

The Gini coefficient is one of the most frequently used measures of economic inequality. The coefficient can take any values between 0 to 1 (or 0% to 100%). A coefficient of zero indicates a perfectly equal distribution of income or [wealth](https://corporatefinanceinstitute.com/resources/knowledge/economics/gross-domestic-product-gdp/) within a population. A coefficient of one represents a perfect inequality when one person in a population receives all the income, while other people earn nothing. In addition, in some rare cases, the coefficient can exceed 100%. This may theoretically occur when the income or wealth of a population is negative.

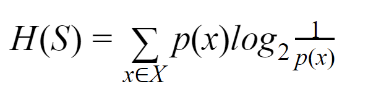
However, the above-mentioned scenarios are extremely rare in the real world. The data shows that the coefficient generally ranges from 24% to 63%.

Please note that the Gini coefficient is not an absolute measure of a country’s income or wealth. The coefficient only measures the dispersion of income or wealth within a population.

1. **What is entropy?**

entropy is defined as a measure of randomness or disorder of a system.

For a finite set S, Entropy, also called Shannon Entropy, is the measure of the amount of randomness or uncertainty in the data. It is denoted by H(S).

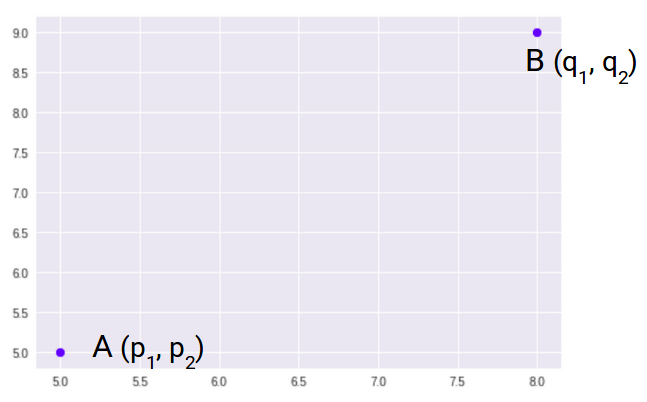


In simple terms, it predicts a certain event by measuring the purity.

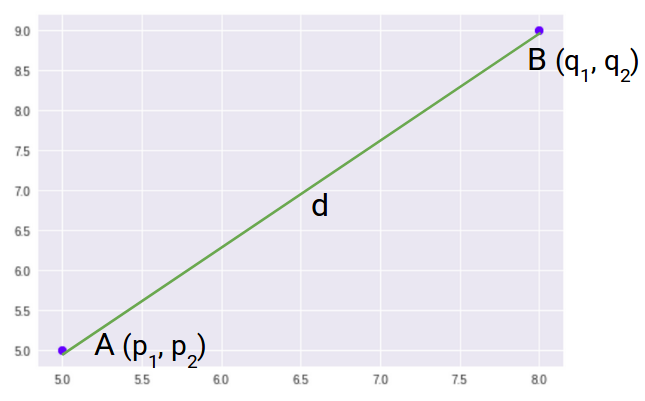
1. **What is Euclidean Distance?**

Euclidean Distance represents the shortest distance between two points.

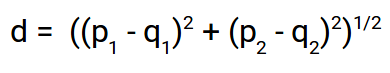
Most machine learning algorithms including K-Means use this distance metric to measure the similarity between observations. Let’s say we have two points as shown below:



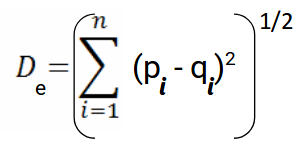
So, the Euclidean Distance between these two points A and B will be:



Here’s the formula for Euclidean Distance:



We use this formula when we are dealing with 2 dimensions. We can generalize this for an n-dimensional space as:



Where,

* n = number of dimensions
* pi, qi = data points

**7. What's the difference between correlation and covariance?**

**Covariance** and **Correlation** are two mathematical concepts which are quite commonly used in business statistics. Both of these two determine the relationship and measures the dependency between two random variables. Despite, some similarities between these two mathematical terms, they are different from each other. Correlation is when the change in one item may result in the change in another item.

Correlation is considered as the best tool for for measuring and expressing the quantitative relationship between two variables in formula. On the other hand, covariance is when two items vary together.

Other differences are as follows:

1. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
2. Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
3. The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between -∞ and +∞.
4. Covariance is affected by the change in scale, i.e. if all the value of one variable is multiplied by a constant and all the value of another variable are multiplied, by a similar or different constant, then the covariance is changed. As against this, correlation is not influenced by the change in scale.
5. Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

**8. What is mean squared error?**

The mean squared error measures the average of the squares of the errors. What this means, is that it returns the average of the sums of the square of each difference between the estimated value and the true value. The MSE is always positive, though it can be 0 if the predictions are completely accurate.

**9. What is the difference between covariance, standard deviation and mean squared error?**

Covariance measures the direction of the relationship between two variables. A positive covariance means that both variables tend to be high or low at the same time. A negative covariance means that when one variable is high, the other tends to be low.

Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).