1. **What is data normalization? How is it different from database normalization (1st/2nd/3rd)?**

Ans: Data normalization is generally considered the development of clean data. 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.

Simply put, this process includes eliminating unstructured data and redundancy (duplicates) to ensure logical data storage. When data normalization is done correctly, you will end up with standardized information entry. For example, this process applies to how URLs, contact names, street addresses, phone numbers, and even codes are recorded. These standardized information fields can then be grouped and read swiftly.

**Whereas database normalization is the** design technique that reduces data redundancy and eliminates undesirable characteristics like Insertion, Update and Deletion Anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The purpose of Normalization in SQL is to eliminate redundant (repetitive) data and ensure data is stored logically.

**1. First Normal Form (1NF):**

The most basic form of data normalization is 1NF which ensures there are no repeating entries in a group. To be considered 1NF,

* each entry must have only one single value for each cell
* each record must be unique.

For example, you are recording the name, address, gender of a person, and if they bought cookies.

**2. Second Normal Form (2NF):**

Again, working to ensure no repeating entries, to be in the 2NF rule,

* the data must first apply to all the 1NF requirements.
* data must have only one primary key.
* To separate data to only have one primary key, all subsets of data that can be placed in multiple rows should be placed in separate tables. Then, relationships can be created through new foreign key labels.

For example, you are recording the name, address, gender of a person, if they bought cookies, as well as the cookie types. The cookie types are placed into a different table with a corresponding foreign key to each person’s name.

**3. Third Normal Form (3NF):**

For data to be in this rule,

* it must first comply with all the 2NF requirements.
* data in a table must only be dependent on the primary key. If the primary key is changed, all data that is impacted must be put into a new table.

For example, you are recording the name, address, and gender of a person but go back and change the name of a person. When you do this, the gender may then change as well. To avoid this, in 3NF gender is given a foreign key and a new table to store gender.

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

Ans: A Statistical Distribution of data represents how often each value occurs in the dataset. Distributions of data may be frequency distributions or probability distributions. Distributions gives an idea of the central tendency (typical value), dispersion (deviations from the typical value) and most common and uncommon values present in the dataset.

**Frequency Distribution:** It is a graph or dataset organized to show the frequency(number of times) of occurrence of each possible outcome of a repeated event observed many times. It is usually represented using a histogram.

**Probability Distribution:** It is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

**3) What is a decision? How's it different from inference?**

Ans: A decision problem is a computational problem that can be posed as a yes–no question of the input values.

For example, deciding whether a number is prime or not

Statistical inference is the process of analyzing the result and making conclusions from data subject to random variation. It is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data.

**4) Google- what is Gini in probability, and explain in your own terms.**

Ans: The Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. It is a number between 0 and 1, with 0 representing perfect equality and 1 perfect inequality. Sometimes these statistics are reported in terms of percentages, with numbers between 0 and 100.

It is typically used to quantify income inequality in human populations, and in that case a Gini index of 0 would mean that everyone earns the same, while 1 would mean that one person earns all the money there is.

**5) What is entropy?**

Ans: Entropy is a scientific concept as well as a measurable physical property that is mostly associated with a state of disorder, randomness, or uncertainty.

Information Entropy or Shannon's entropy quantifies the amount of uncertainty (or surprise) involved in the value of a random variable or the outcome of a random process. Its significance in the decision tree is that it allows us to estimate the impurity or heterogeneity of the target variable.

**6) What is Euclidean Distance?**

Ans: The Euclidean distance is defined as the distance between two points. In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.

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

Ans: Covariance and correlation both are used in regression analysis to measure associativity of two variables, the basic differences are that

* covariance measures how the two variables differ, while correlation measures how the two variables are related.
* Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance.
* The value of correlation takes place between -1 and +1. Conversely, the value of covariance lies between -∞ and +∞.

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

Ans: The mean squared error is the sum of the squared difference of each value of the dataset from the mean. It is given as:

It is used to see how well a model fits a dataset. The greater the MSE, the more unfit is the model for that data.

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

Ans: **Covariance** measures the direction of the relationship between two variables. A positive covariance implies that both the variables have high or low values together. A negative covariance implies that if one variable has high value then the other has low values. This is given by:

**Standard Deviation** is a measure of the dispersion of the actual values in the data from the central tendency. It conveys the spread of the data. It is given as the square root of variance.

**Mean squared error** measures the amount of error in a statistical model and is given as: