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

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

**1NF**.

The most basic form of data normalization is 1NFm 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 and each record must be unique.

**2NF**.

To ensure no repeating entries, to be in the 2NF rule, the data must first apply to all the 1NF requirements. Following that, 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.

**3NF**.

For data to be in this rule, it must first comply with all the 2NF requirements. Following that, 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.

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

A statistical distribution, or probability distribution, describes how values are distributed for a field. The statistical distribution shows which values are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution. The most common representation of the distribution is Histogram.

A frequency distribution table is a chart that summarizes values and their frequency. It's a useful way to organize data if you have a list of numbers that represent the frequency of a certain outcome in a sample.

A Probability distribution can be use in either of the following ways. By using a discrete distribution, you can define the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has only countable values, for example, a list of non-negative integers.

1. **What is a decision? How's 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.

1. **Google- what is Gini in probability and explain in your own terms.**

The Gini coefficient is a statistic which quantifies the amount of inequality that exists in a population. The Gini coefficient 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.

Chart, line chart, histogram

Description automatically generated

1. **What is entropy?**

An 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. Entropy is a scientific concept as well as a measurable physical property that is most commonly associated with a state of disorder, randomness, or uncertainty.

Entropy can be calculated for a random variable X with k in K discrete states as follows: H(X) = -sum(each k in K p(k) \* log(p(k)))

1. **What is Euclidean distance?**

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.

The Euclidean distance formula is given by:

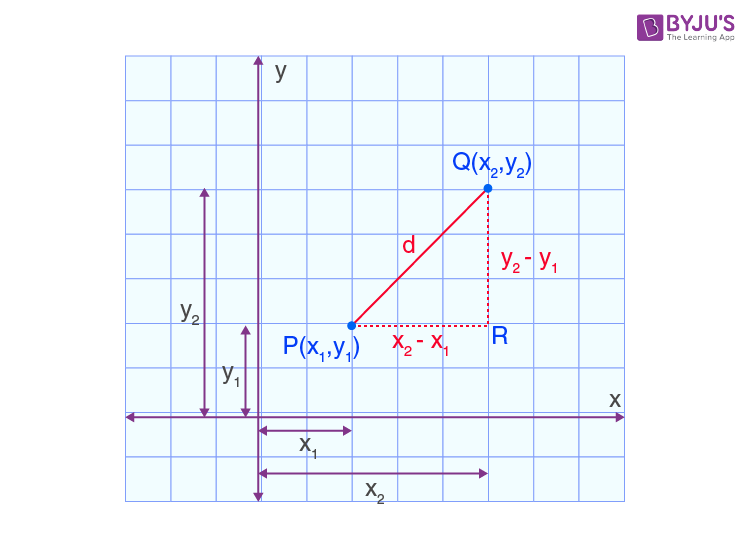
d =√[(x2– x1)2 + (y2– y1)2]

Where,

“d” is the Euclidean distance

(x1, y1) is the coordinate of the first point

(x2, y2) is the coordinate of the second point.



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

Suppose you are study two quantitative variable and change in one variable has a reciprocate effect on the other variable(directly or indirectly), this is called correlation. Correlation is a measure that determines the degree to which two or more random variables move in sequence.

Covariance is a statistical term that refers to a systematic relationship between two random variables in which a change in the other reflects a change in one variable.

Covariance and correlation are two terms that are opposed and are both used in [statistics](https://www.simplilearn.com/what-is-statistical-analysis-article) and [regression analysis.](https://www.simplilearn.com/tutorials/excel-tutorial/regression-analysis) Covariance shows you how the two variables differ, whereas correlation shows you how the two variables are related.

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

The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

There is no correct value for MSE. Simply put, the lower the value the better and 0 means the model is perfect.

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

variance is used in calculating standard deviation, which is a measure of how spread out a set of data is. Covariance is used in calculating correlation, which is a measure of how two variables relate to each other.  whereas the MSE measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated.