**Technical Session**

**09-09-2022**

**2211586-Vaishali.A-CDB22DW022**

**Questions:**

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

2) What is a distribution? What are the uses for frequency and probability distribution?

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

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

5) What is entropy?

6) What is Euclidean distance?

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

8) What is mean squared error?

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

**Answers:**

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

The production of clean data is generally referred to as **Data Normalization**. However, when you dig a little deeper, 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.

**First Normal Form (1NF)**

The simplest kind of Data Normalization is**1NF**, which assures that no two entries in a group are identical. The following rules must be followed for a table to be in the first normal form:

* Each cell should only have one value.
* Each record should be distinct from the others.

**Second Normal Form (2NF)**

All subsets of data that can be placed in multiple rows are placed in distinct tables in a **2NF** table. The following rules must be followed for a table to be in the second normal form:

* It should be in the 1NF.
* Any subset of Candidate keys should not be functionally dependent on the Primary key.

**Third Normal Form (3NF)**

The following rules must be followed for a table to be in the 3NF:

* It should be in the 2NF.
* It should not have any functional dependencies that are transitive.

A Transitive Functional dependency occurs when a change in one column (that isn’t a primary key) affects all of the other columns.

**2) 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**. Think about a die. It has six sides, numbered from 1 to 6.

Frequency Distribution

Frequency distribution shows the frequency of occurrence of every possible outcome as collected from a sample group. An example would be an election, where there’s a running tally of votes for every party and its candidate.

* It helps analyse the data collected to ensure it could help in effective decision-making.
* Grouped, ungrouped, cumulative, relative, and relative cumulative distributions are some of its widely used types.

Probability Distribution

The probability distribution gives the possibility of each outcome of a random experiment. It provides the probabilities of different possible occurrences.

The probability distribution is one of the important concepts in statistics. It has huge applications in business, engineering, medicine and other major sectors. It is majorly used to make future predictions based on a sample for a random experiment. For example, in business, it is used to predict if there will be profit or loss to the company using any new strategy or by proving any hypothesis test in the medical field, etc.

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

Decision theory, 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 inference and decision processes are therefore**different and solved by different methods** - Bayes' theorem and maximization of expected utility respectively.

Inferences represent a step in the process from collecting data or information to rendering a final judgment on that information, and to making a decision about how to respond.

Statistical inference is the act of using observed data to infer unknown properties and characteristics of the probability distribution from which the data have been extracted.

Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics. Hypothesis testing and [confidence intervals](https://byjus.com/maths/confidence-interval/)are the applications of the statistical inference. Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation. It allows us to provide a probable range of values for the true values of something in the population.

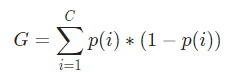
**4) Google- what is Gini in probability, and explain in your own terms**

The Gini Index or Gini Impurity is calculated by subtracting the sum of the squared probabilities of each class from one. It favours mostly the larger partitions and are very simple to implement. In simple terms, it calculates the probability of a certain randomly selected feature that was classified incorrectly.

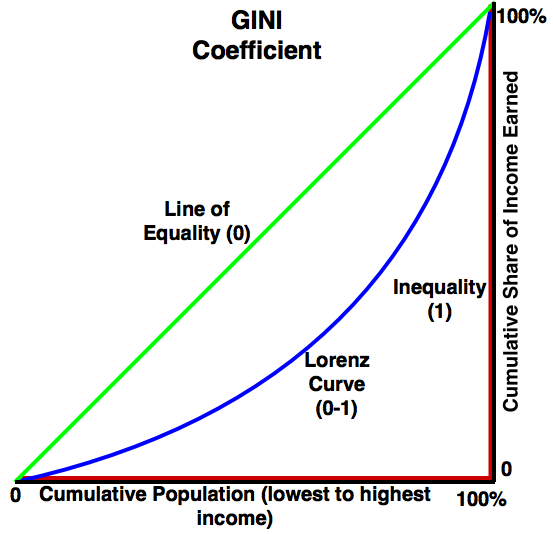
The Gini Index varies between 0 and 1, where 0 represents purity of the classification and 1 denotes random distribution of elements among various classes.

A Gini Index of 0.5 shows that there is equal distribution of elements across some classes.

Mathematically, The Gini Index is represented by



The Gini Index works on categorical variables and gives the results in terms of “success” or “failure” and hence performs only binary split.

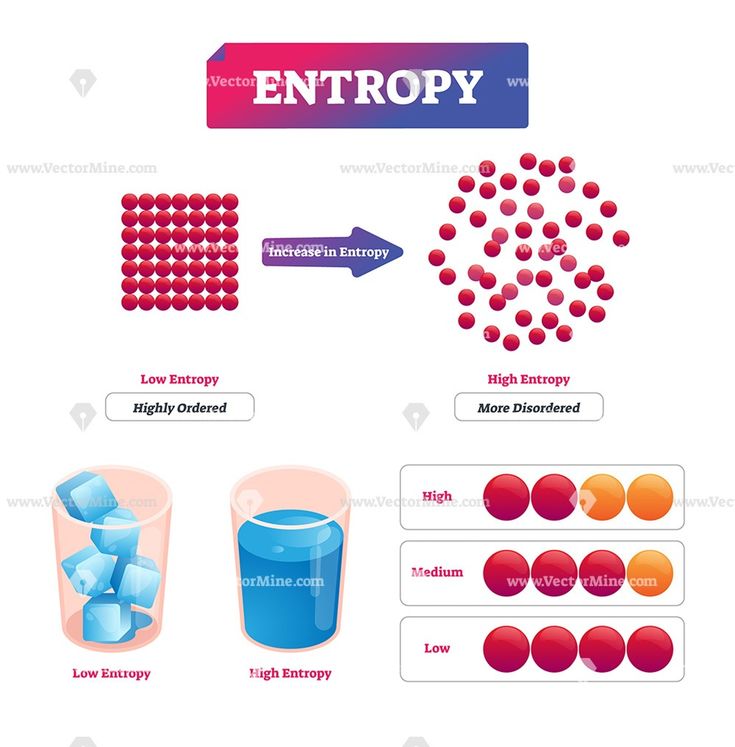


**5) What is entropy?**

Entropy is a measure of the randomness or disorder of a system.

The value of entropy depends on the mass of a system. It is denoted by the letter S and has units of joules per kelvin.

Entropy can have a positive or negative value. According to the second law of thermodynamics, the entropy of a system can only decrease if the entropy of another system increases.



**6) 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. As the Euclidean distance can be found by using the coordinate points and the [**Pythagoras theorem**](https://byjus.com/maths/pythagoras-theorem/), it is occasionally called the Pythagorean distance.

The Euclidean distance formula helps to find the distance of a line segment. Let us assume two points, such as (x1, y1) and (x2, y2) in the two-dimensional coordinate plane.

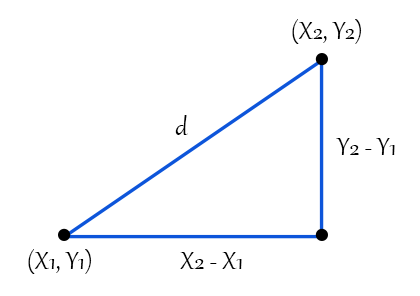
Thus, the Euclidean distance formula is given by:

**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. 

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

### Definition of Correlation

Correlation is described as a measure in statistics, which determines the degree to which two or more random variables move in tandem. During the study of two variables, if it has been observed that the movement in one variable, is reciprocated by an equivalent movement another variable, in some way or the other, then the variables are said to be correlated.

Correlation is of two types, i.e. positive correlation or negative correlation. The variables are said to be positively or directly correlated when the two variables move in the same direction. On the contrary, when the two variables move in opposite direction, the correlation is negative or inverse.

### Definition of Covariance

Covariance is a statistical term, defined as a systematic relationship between a pair of random variables wherein a change in one variable reciprocated by an equivalent change in another variable.

Covariance can take any value between -∞ to +∞, wherein the negative value is an indicator of negative relationship whereas a positive value represents the positive relationship. Further, it ascertains the linear relationship between variables. Therefore, when the value is zero, it indicates no relationship. In addition to this, when all the observations of the either variable are same, the covariance will be zero.

In Covariance, when we change the unit of observation on any or both the two variables, then there is no change in the strength of the relationship between two variables, but the value of covariance is changed.

**Differences:**

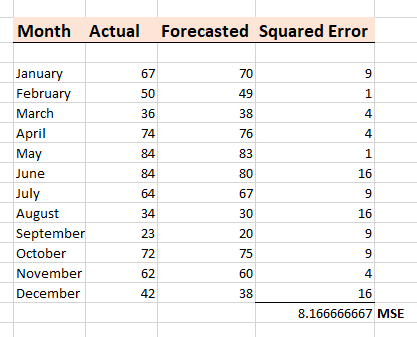
**Covariance** and **Correlation** are two mathematical concepts which are quite commonly used in business statistics. Both of these two determine the relationship and measures the dependency between two random variables. Despite, some similarities between these two mathematical terms, they are different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

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. It incorporates the variance of the estimator (how widely spread the estimates are) and its bias (how different the estimated values are from their true values).



from sklearn.metrics import mean\_squared\_error

actual\_values = [3, -0.5, 2, 7]

predicted\_values = [2.5, 0.0, 2, 8]

mean\_squared\_error(actual\_values, predicted\_values)

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

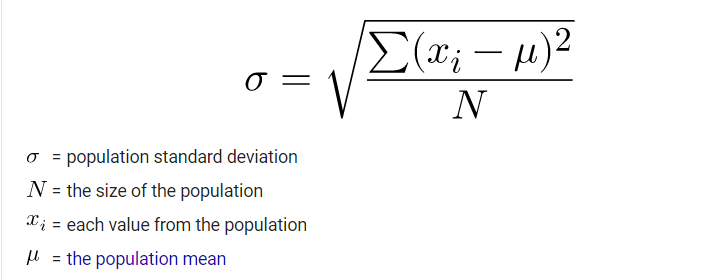
Covariance-

In mathematics and [statistics](https://corporatefinanceinstitute.com/resources/knowledge/basic-statistics-concepts/), covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables.

Standard Deviation :

It is a measure of dispersion of observation within dataset relative to their mean.It is square root of the variance and denoted by Sigma (σ) .

Standard deviation is expressed in the same unit as the values in the dataset so it measure how much observations of the data set differs from its mean.



MSE-

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and [predicted values](https://statisticsbyjim.com/glossary/fitted-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).

