**TECHNICAL SESSION – ASSIGNMENT (09.09.2022)**

**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. This procedure entails**removing Unstructured Material**as well as Redundancy (duplicates) to ensure logical data storage.

**Database normalization** is the process of structuring a [relational database](https://en.wikipedia.org/wiki/Relational_database) in accordance with a series of so-called [normal forms](https://en.wikipedia.org/wiki/Database_normalization#Normal_forms) in order to reduce [data redundancy](https://en.wikipedia.org/wiki/Data_redundancy) and improve [data integrity](https://en.wikipedia.org/wiki/Data_integrity). Normalization entails organizing the [columns](https://en.wikipedia.org/wiki/Column_(database)) (attributes) and [tables](https://en.wikipedia.org/wiki/Relation_(database)) (relations) of a database to ensure that their [dependencies](https://en.wikipedia.org/wiki/Dependency_theory_(database_theory)) are properly enforced by database integrity constraints.

**First Normal Form (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. For example, you are recording the name, address, gender of a person, and if they bought cookies.

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

**Third Normal Form (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.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.

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

The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space. This distribution describes the grouping or the density of the observations, called the probability density function. We can also calculate the likelihood of an observation having a value equal to or lesser than a given value. A summary of these relationships between observations is called a cumulative density function.

**Frequency distribution** is easier than working with raw statistical data as it’s based on data that’s already been categorized, which, in turn, makes it easier to calculate data for a data set like the mean, the median and average of occurrences. While most areas of research can use a frequency distribution to observe data.

Some specific instances in which we use charts, graphs and tables: Statistical analysis, Sales and marketing trend research, medical research, Behavioural studies, Sociological experiments, Customer, employee or student satisfaction, Weather research and predictions.

**Probability distribution** can be used in 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.

By using a continuous distribution, you can define the probabilities of the possible values of a continuous random variable. A continuous random variable is a random variable that has a set of possible infinite and uncountable values.

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

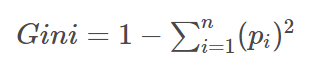
Statistical decisions are decisions made based on observations of a phenomenon that obeys probabilistic laws that are not completely known.

A solvable [decision problem](https://www.britannica.com/topic/decision-problem) must be capable of being tightly formulated in terms of initial conditions and choices or courses of action, with their [consequences](https://www.britannica.com/dictionary/consequences). In general, such consequences are not known with certainty but are expressed as a set of probabilistic outcomes. Each outcome is assigned a “utility” value based on the preferences of the decision maker. An optimal decision, following the logic of the theory, is one that maximizes the [expected utility](https://www.britannica.com/topic/expected-utility). Thus, the ideal of decision theory is to make choices rational by reducing them to a kind of routine calculation.

**Inference is** the process of drawing conclusions about a [parameter](https://www.merriam-webster.com/dictionary/parameter) one is seeking to measure or estimate. One principal approach of statistical [inference](https://www.merriam-webster.com/dictionary/inference) is [Bayesian estimation](https://www.britannica.com/topic/Bayess-theorem), which incorporates reasonable expectations or prior judgments (perhaps based on previous studies), as well as new observations or experimental results. Another method is the [likelihood](https://www.britannica.com/science/likelihood) approach, in which “prior probabilities” are [eschewed](https://www.merriam-webster.com/dictionary/eschewed) in favour of calculating a value of the parameter that would be most “likely” to produce the observed distribution of experimental outcomes. In parametric inference, a particular mathematical form of the [distribution function](https://www.britannica.com/science/distribution-function) is assumed. Nonparametric inference avoids this assumption and is used to estimate parameter values of an unknown distribution having an unknown functional form.

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

The Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one.  It favors larger partitions.  Information Gain multiplies the probability of the class times the log (base=2) of that class probability.  Information Gain favors smaller partitions with many distinct values.  The Gini index (or coefficient) is a synthetic indicator that captures the level of inequality for a given variable and population. It varies between 0 (perfect equality) and 1 (extreme inequality). Between 0 and 1, the higher the Gini index, the greater the inequality.



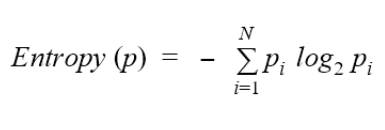
It is equal to 0 in a situation of perfect equality where the variable takes an identical value over the whole population. At the other extreme, it is equal to 1 in the most unequal situation possible, where the variable is equal to 0 over the entire population except for one individual. Inequality measured in this way can relate to variables such as income, wages, standard of living.

**5) What is entropy?**

The entropy measures the “amount of information” present in a variable. In information theory, the entropy of a random variable is the average level of “information“, “surprise”, or “uncertainty” inherent in the variable’s possible outcomes.

That is, the more certain or the more deterministic an event is, the less information it will contain. In a nutshell, the information is an increase in uncertainty or entropy.

Claude E. Shannon had expressed this relationship between the probability and the heterogeneity or impurity in the mathematical form with the help of the following equation:



The uncertainty or the impurity is represented as the log to base 2 of the probability of a category (pi). The index (i) refers to the number of possible categories. Here, i = 2 as our problem is a binary classification.

**6) What is Euclidean distance?**

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 says:

**d = √ [ (x22 – x11)2 + (y22 – y11)2]** where,

* (x11, y11) are the coordinates of one point
* (x22, y22) are the coordinates of the other point
* d is the distance between (x11, y11) and (x22, y22).

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

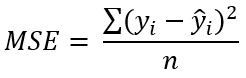
Statistical analysis is an excellent method for mathematically exploring the relationships between different variables. Two important statistical methods for analyzing data include covariance and correlation. In general terms, covariance describes the difference between two variables, while correlation looks at the relationship between two variables.

Covariance and standard deviation are easy to be confused, as are correlation and standard deviation. However, each term - covariance, correlation, and standard deviation- represents distinct statistical concepts. Covariance specifically examines whether a change in one variable produces a change in a second variable. While covariance shows how two variables are related to each other, it does not show the strength of the relationship between the two variables. This is where correlation is important. Correlation looks at the strength of the relationship between two variables. To calculate the correlation between two variables, the covariance value is divided by the standard deviation of both variables. Standard deviation is a measure of the amount of variation that is found in a data set.

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

The **mean squared error** (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the [mean](https://www.statisticshowto.com/mean/)squared error as you’re finding the average of a set of errors.

Formula for Mean squared error:



**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 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:

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

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.

σ=√[1N∑Ni=1(Xi−μ)]2