Population--A population refers to the entire group or set of all possible observations

Sample--A sample is a subset of the population that is selected for actual observation or measurement. It represents the population and is used to make inferences about the population.

Central Limit Theorem(CLT)-- the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.

Covariance-- is a statistical measure that indicates the extent to which two random variables change together. It tells us whether there is a positive, negative, or no linear relationship between the two variables.

Pearson Correlation Coefficient (often denoted as **r**) is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from −1 to 1.

Spearman's Rank Correlation Coefficient --is a non-parametric measure of the strength and direction of the monotonic relationship between two variables. Unlike Pearson's correlation, which measures linear relationships, Spearman's correlation assesses how well the relationship between two variables can be described by a monotonic function (where one variable consistently increases or decreases with the other).

Outlier-- is a data point that significantly deviates from the other data points in a dataset.

Z-score and IQR are used to identify the outliers

🡪A Z-score greater than ±3 is often considered an outlier.i.e if the data point falls far from the 3rd standard deviation then it is considered as an outlier.

Standardization and normalization are two common techniques used to preprocess data, especially in the context of machine learning and statistical analysis. Both methods are used to adjust the scale of the data, but they do so in different ways and have different purposes.

P value—It is the probability of the null hypothesis to be true.Example-pressing the space key on laptop(usually many of them prefer to press on middle of space bar, so out of 100 p value may be 0.8(80%) and only few press on the sides of the space bar so 10% i.e p=0.1)

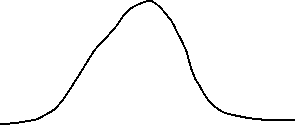
**Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable. It helps in understanding whether a distribution is **symmetrical** or **skewed** (asymmetrical), and in which direction the skewness occurs.

A positive skew (or right-skewed) distribution has a longer tail on the right side of the distribution. This occurs when there are a small number of exceptionally high values that "pull" the tail to the right. Mean > Median > Mode.Eg--Income distribution in a population (many people earn low to moderate incomes, and a few earn very high incomes).

A negative skew (or left-skewed) distribution has a longer tail on the left side. This occurs when there are a small number of exceptionally low values that "pull" the tail to the left. Mean < Median < Mode. **Example**: Age at retirement (many people retire around a typical age, but a few retire much earlier).

Confidence intervals are a range of values used to estimate the true value of a population parameter (such as a mean or proportion). They provide a range within which we expect the parameter to fall, given a certain level of confidence (like 95% or 99%).

🡪**Confidence Level**: The probability that the true population parameter lies within the confidence interval. Common levels are 90%, 95%, and 99%. A 95% confidence level means that if we were to take 100 different samples and compute a confidence interval for each sample, about 95 of the intervals would contain the true population parameter.



Confidence interval( 95%)



Significance level—5%(0.05)

P value and significance value are same

Bernoulli distribution is a discrete probability distribution that describes the outcome of a random experiment with exactly two possible outcomes: "success" (usually coded as 1) and "failure" (usually coded as 0).

🡪Probability Mass Function (PMF**)** is a function that gives the probability that a discrete random variable is exactly equal to some value. It is used to describe the distribution of discrete random variables, such as those that follow a Bernoulli, Binomial, Poisson, or any other discrete probability distribution.

🡪Probability Density Function (PDF) is used to describe the probability distribution of a continuous random variable. Unlike a discrete random variable (where you use a Probability Mass Function, or PMF), a continuous random variable can take on any value within a certain range or interval.

Types of Sampling Techniqes

🡪**Simple Random Sampling**: Every member of the population has an equal chance of being selected.

* *Example*: Drawing names from a hat to select 10 students from a class of 50.

🡪**Stratified Sampling**: The population is divided into subgroups, and random samples are taken from each subgroup. The population is divided into homogeneous subgroups (strata) based on specific characteristics (e.g., age, gender, income), and a random sample is taken from each stratum.

* *Example*: Dividing a company's employees into departments and randomly selecting employees from each department.(also based og age,gender etc)

**🡪Systematic Sampling**: Every n-th member of the population is selected after a random starting point.

* *Example*: Choosing every 5th person from a list of 200 names, starting from a random point.

🡪**Cluster Sampling**: The population is divided into clusters, some clusters are randomly selected, and all members within those clusters are sampled.

* *Example*: Randomly selecting people who have knowledge or experts in AI domain.

Confusion matrix is a tool used in classification problems to evaluate the performance of a classification algorithm. It is a table that compares the actual target values with the values predicted by the model. The matrix helps to visualize the performance of a classification model by showing the counts of correct and incorrect predictions for each class.

**🡪 Structure of a Confusion Matrix:**

For a binary classification problem, the confusion matrix is a 2x2 table with the following components:

1. **True Positives (TP)**: The number of instances correctly classified as the positive class.
2. **False Positives (FP)[Type 1 error]**: The number of instances incorrectly classified as the positive class (i.e., the model predicted positive but the actual class is negative).
3. **True Negatives (TN)**: The number of instances correctly classified as the negative class.
4. **False Negatives (FN) [Type 2 error]**: The number of instances incorrectly classified as the negative class (i.e., the model predicted negative but the actual class is positive).

False Negative--A patient who actually has the disease is tested and the result comes back negative (indicating the disease is absent).

False Positive--A patient who is healthy (does not have the disease) is tested and the result comes back positive (indicating the disease is present).

So based on the domain it is important to see that whether to decrease the FP or FN.Example in medical domain the FN has to be decreased(because if a person actually has disease but it has predicted as negative so it is very bad ,but FP is ok because even if the person is not having the disease but if it shows as positive that person can retest so that not an issue,Basically it depends on the specific domain).



In market collapse prediction it is important to decrease both FN and FP because even though if it predicts like the market collapse but actually it does not collapse,then the people will lose money.So both has to be decreased.



Z-score is a statistical measure that describes a value's position relative to the mean of a group of values.i.e to determine how far the value from the mean .

🡪Applications of Z-score

--Standardization(in preprocessing)

--Comparing scores between different distributions(Eg.Comparing the performance of Indian cricket team of 2 different years scores).

Probability is a branch of mathematics that deals with the likelihood or chance of different outcomes occurring in an experiment or random event.

🡪**Mutually exclusive events** are events that cannot occur simultaneously. In other words, if one event happens, the other cannot. For example, when flipping a coin, the events "getting heads" and "getting tails" are mutually exclusive because they cannot both happen at the same time.

🡪**Not mutually exclusive events** (also known as non-mutually exclusive events) are events that can occur at the same time. In other words, the occurrence of one event does not exclude the possibility of the other event occurring simultaneously. A survey is conducted among a group of people to determine their preferences for two popular movie genres: Action and Comedy.

* **Event A:** A person likes Action movies.
* **Event B:** A person likes Comedy movies.

These events are not mutually exclusive because a person can like both Action and Comedy movies.

Encoding--Encoding in data science and machine learning refers to transforming categorical data (which can be non-numerical or text-based) into numerical format so that it can be fed into machine learning models, which typically require numerical input.

Types of Encoding:

🡪Nominal encoding(Eg-States column—Karnataka,delhi,goa,Mumbai etc)

* One Hot Encoding
* One Hot Encoding for many categorical variables
* Mean Encoding

🡪Ordinal Encoding(Eg-Degree—Phd[with ranking 1],Masters[2],B.E[3],B.Com[4])

* Label Encoding
* Target guided ordinal encoding

Handle missing values in categorical variables

* Delete the rows(not preferable)
* Replace missing values with the most frequent (or mode) category in the feature.
* Apply classifier algorithm to predict(i.e whichever the column has the missing values consider that as the dependent variable and all other as independent variables and then predict for the missing values)
* Apply the unsupervised ML.

Linear Regression-- is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors).

To find the best-fitting line (regression line) through the data points that minimizes the difference between the predicted and actual values of the dependent variable.

* **Simple Linear Regression**: Involves one independent variable to predict the dependent variable. The model can be represented as:

y=b0+b1x+ϵ

y—dependent variable,b0—intercept,b1—slope,x—independent variable

Overfitting-- High accuracy on training data but low accuracy on validation or test data.

Underfitting-- Low accuracy on both training and test data.

**🡪Ridge** and **Lasso regression** are two popular techniques used to prevent overfitting in linear regression models by adding a regularization term to the cost function.

Example-When a OLS(used to find the best fit line) gives 0 which means that line is best fit line but when it is performed for test data we observe overfitting,so to avoid this we add a penalty term to the loss function(y-y^2) so that it gives the larger value i.e greater than zero so we go for the other line ,which ever gives the less value that can be considered as best fit line.

Ridge regression(L2 Regularization) adds an L2 penalty (the square of the magnitude of coefficients) to the loss function. This encourages the model to keep the coefficients small but does not set any of them to zero.(it moves towards zero but it never becomes zero but lasso regression can set to zero where whichever the features does not contribute

Lasso regression(L1 Regularization) adds an L1 penalty (the absolute value of the magnitude of coefficients) to the loss function. This can result in some coefficients being exactly zero, effectively performing feature selection. Lasso regression can shrink some coefficients to zero, making it useful for feature selection in models with a large number of predictors. It helps in simplifying the model and improving interpretability.

* **Multiple Linear Regression** is an extension of simple linear regression that models the relationship between a dependent variable and two or more independent variables. It is a powerful tool for understanding how multiple factors simultaneously impact the target variable.

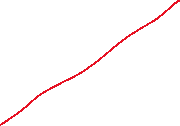
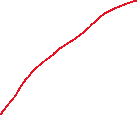
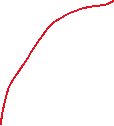
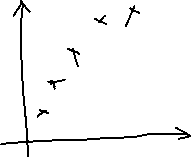
y=b0​+b1​x1​+b2​x2​+…+bn​xn​+ϵ



Multicolinearity in Regression-- refers to a situation in linear regression where two or more independent variables are highly correlated with each other. Multicollinearity can create problems in estimating the coefficients of the model and interpreting their impact on the dependent variable.So it is preferred to consider only one variable when 2 different variables are highly correlated because since both variables shows somewhat equal behaviour since they are correlated,so consider the variable which has high p value for removing and keep the other variable in the data.

Bias and Variance—

* Bias-Bias is considered a systematic error that occurs in the machine learning model itself due to incorrect assumptions in the ML process. Technically, we can define bias as the error between average model prediction and the ground truth.
* Variance-Variance refers to the error due to the model's sensitivity to fluctuations in the training data. It represents the variability of model predictions for a given data point or a set of data points.



---Training data



--Test data

1st graph🡪degree of polynomial is 1(simple linear regression)

Here there is lot of distance between the training data and test data prediction i.e the data points(both training and test data) are very far from the predicted line,so it is not predicting well for both training and test data so it is an underfitting model.So for this there is high bias(high error because wrong prediction for training data) and high variance.

2nd graph-- degree of polynomial is 2

Here the distance between the data points and the predicted line is not so far,so the prediction is quite good for both training and test data.

So it has Low Bias and low variance

3rd graph-- degree of polynomial is 4

Here the training data is predicted very well which means all the training data points lies exactly on the predicted line(so bias is low) but for the test data points the predicted is bad i.e the distance between the test data points(so high variance) and the prediction line is far ,so it is an overfitting model.

It has Low bias and high variance.

So from the above observations we can conclude that the second model which has low bias and low variance give the highest accuracy…

All the above graphs are for regression.

🡪For classification

* Model 1:

Training error—1%(Low bias)

Testing error—20%(High variance)

i.e overfitting model

* Model 2:

Training error—25%(High bias)

Testing error—25%(High variance)

i.e underfitting model

* Model 3:

Training error <10%(Low bias)

Testing error<10%(Low variance)

So this model gives the highest accuracy..

R-Squared--it indicates how well the data fits the regression model.It value is between 0 and 1.The more it is towards 1,it indicates that the more it(data) fits the model.

As the number of variables keeps increasing in the (linear regression equation),the R-squared values keep increasing,it never decreases which means even the co-efficients are increasing it is not making any change or not influencing for the change so to overcome this the ‘Adjusted R-squared’ is used.

🡪Adjusted R-squared--When you add more independent variables to a model, the R-squared value will always increase or remain the same, even if the new variables do not improve the model. **Adjusted R-squared** adjusts for the number of predictors in the model, penalizing the addition of irrelevant variables.(which means if there is no correlation between the independent and dependent variables then the adjusted r-squared value decreases but if there is correlation then the adjusted r-squared remains same or slightly decreases compared to original r-squared)

Diff b/w R-squared and Adjusted R-squared

* Everytime you add a independent variable to a model, the **R-squared** increases even if the independent variable is insignificant,it never declines.Whereas **Adjusted R-squared** increases only when the independent variable is significant

and affects dependent variable.

* Adjusted R-squared value will be always less than or equal to R-squared value.

--------------------------------------------------------------------------------------------------------------

Entropy🡪Entropy is a concept from information theory that measures the **uncertainty or impurity** in a dataset. In the context of machine learning, particularly decision trees, entropy is used to quantify the amount of uncertainty or randomness in the target variable.

Mathematically, for a binary classification problem, entropy H(S) is defined as:

H(S)=−p1log2(p1)−p2log2(p2)

p1​ is the proportion of examples in class 1.

p2​ is the proportion of examples in class 2

**Low entropy** (near 0) means the data is pure, with most examples belonging to a single class.

**High entropy** (near 1) means the data is impure, with the examples being equally distributed among classes.

Information Gain🡪When a feature is used to split the dataset, information gain tells us how much uncertainty was reduced. A higher information gain indicates that the feature provides more information about the target variable and leads to better splits.

Gini impurity 🡪is a metric used in decision trees to measure how **pure** or **impure** a node is.

**Gini Impurity vs. Entropy:**

* Both Gini impurity and entropy measure the degree of impurity in a dataset.
* Gini impurity tends to be computationally faster because it doesn't involve logarithms.
* Entropy is more sensitive to changes in class distribution compared to Gini impurity.
* The two criteria often yield similar splits in practice, though they may lead to slightly different trees.S