**Data Analysis and Preprocessing**

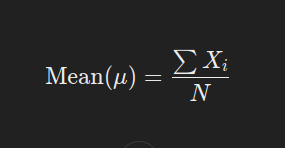
20/02/2025 Ihit R Acharya

**What** **is descriptive statistics?**

It is the method of organizing and processing the statistical data into a meaningful observation, this can be done by measuring the central tendency of data showing that there are low number of outliers in a data set, which gives us an idea about the reliability of the data set provided to us. They can help us to compare between two data sets by checking their relative compatibilities thus helping in statistical data analysis between two similar data sets. Usage of different types of graphs and plots can help us gain insightful look into the way data are interrelated to each other. Measures like variance, mean, median, etc give meaning to data since a mean which is close to the extremes shows the data set’s large variance which can add unneeded bias/error to the conclusion/observations from the data set. Large variance could indicate unreliable data, hence showing that the following data set could be excluded due to inconsistent or unreliable results.

**Explain mean, median, mode, standard deviation, and variance?**

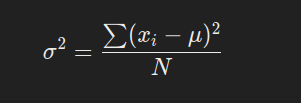
Mean: It is a numerical quantity that denotes the average of all the data points. It helps us identify the point where most of the data points are centered about.



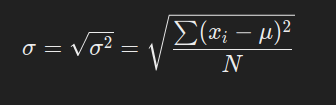
Median: It is a numerical quantity that represents the middle data point from the entire data set. If odd number of data points are there then it is middle most data point in even dataset it is avg of the 2 middle points.

Mode: It is a numerical quantity that shows the most frequently occurring data point which can help us get an idea of the point which could affect data the most. It can also show reliability in some cases since it shows firmness in a specific value of the data point.

Variance: It is a numerical quantity that gives us information about how spread out the data is. It gives us the squared error for the data set which can help magnify the net spread of the set of data points.



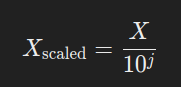
Standard Deviation: It is a numerical quantity that also gives us information about the variability/spread in data. However, unlike variance it gives the net error/difference leading to easy interpretation in comparison to squared values which could exaggerate the spread of data points.



**What is normalization? What are the different methods of Normalization?**

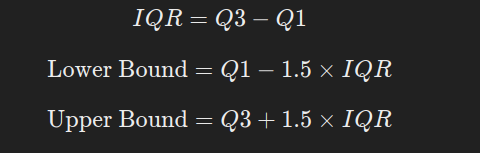
Normalization is a method that is utilized to fit the data to a certain range/scale. Normalization is a process to accordingly help convert the current scale of the data points to a different one. There are many methods for normalization which can be implemented to fit the data namely being feature scaling, mean normalization, Z-score normalization. Feature scaling is a method that brings the data points to a range between 0 and 1 and can be universally applied. It is done by dividing the limits by the upper limit.eg: a<x<b and c<y<d is normalized to give a/b<x<1 and c/d<y<1.Feature scaling has benefits that feature scaling has the best time complexity among the normalizations since it doesn’t need to compute mean or variance, however it can’t deal with data with high variance and in such cases will give inaccurate data. Mean normalization is the method of normalization in which the limit is subtracted by the mean and divided by the difference between upper and lower limit. It gives us the data points scaled down in a range of –1 and 1 .e.g.: a<x<b and c<y<d , is normalized to a-u1 /(b-a) < x < b-u1 /(b-a) and a-u2/(b-a) < y < b-u2 /(b-a) , u1 and u2 are means of x and y respectively. It has the benefits being that it transforms the data to have mean = 0 which allows the data to be centered around 0,it also has benefits over min-max normalization being that it can be utilized over highly spread out data, it can also have another benefit being that it help counter with the tilt/skewness of the data, however it has disadvantages being time complexity in comparison to feature scaling normalization and is inferior to Z score normalization of data .Z score normalization is a method that subtracts the limits by their mean and divides by their standard deviation .e.g.: a<x<b and c<y<d , is normalized to a-u1 / σ1 < x < b-u1 / σ1 and a-u2/ σ2 < y < b-u2 / σ2 , u1 , σ1 and u2 , σ2 are means and standard deviation of x and y respectively. This method provides data taking variance into account thus helps in better dealing with outliers in the data. It also helps deal with independent variables scales giving it another benefit over rest of the normalizations. Log Scaling is a method in which the following formula is applied:

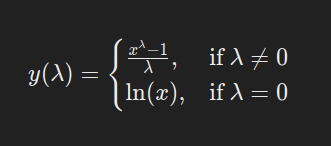
 c is a constant

Log scaling is useful for dealing with skewed data allowing it to deal with value biases, however it is unable to deal with negative values, hence a prerequisite algorithm/normalization is required to utilize log scaling. Decimal scaling is used like this: j is the largest value for which Xscaled is lying between 0 to 1. This method similar to feature scaling in a way that the data point retain their original distribution so it doesn’t abstract the net properties off t

**What are outliers, and different methods of outlier removal?**

Outliers are a set of anomalous data points that don’t agree with the rest of the data, they may be extreme values or maybe highly scattered around the plot. Outliers can be dealt with by utilizing Z-score normalization and IQR range method. IQR range method is a method where data points between a range/percentile are only considered.Following is the formula:



There are also data cleaning libraries like pandas which can be utilized to remove outlier values with the means of the data set. Capping Outliers is a method where outlier data are replaced by the closest extreme point of the data that is considered valid, so it suppresses the effects of the outliers. Transformations is a method where the effect of skewness on causing outliers can be reduced using log or squaring or box-cox. Box-cox is a method where a power/lambda is used to transform the data, different values of lambda can cuase different transforms like square,root,inverse or log transforms. Following is formula: 

**What are common errors found in raw data and what are the methods used to fix them?**

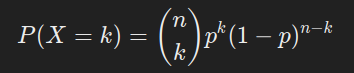
Common error includes NAN values, negative values or missing data which are dealt using pandas which is a data handling and cleaning library which can help deal with these values, the methods include pd.fillna or pd.replace which can help deal with these values and accordingly replace these outlier values with mean of the data set so as to deal with these problems in data. There are also different scales of the data for different columns or rows which are dealt with using normalization methods like Z-score normalization.

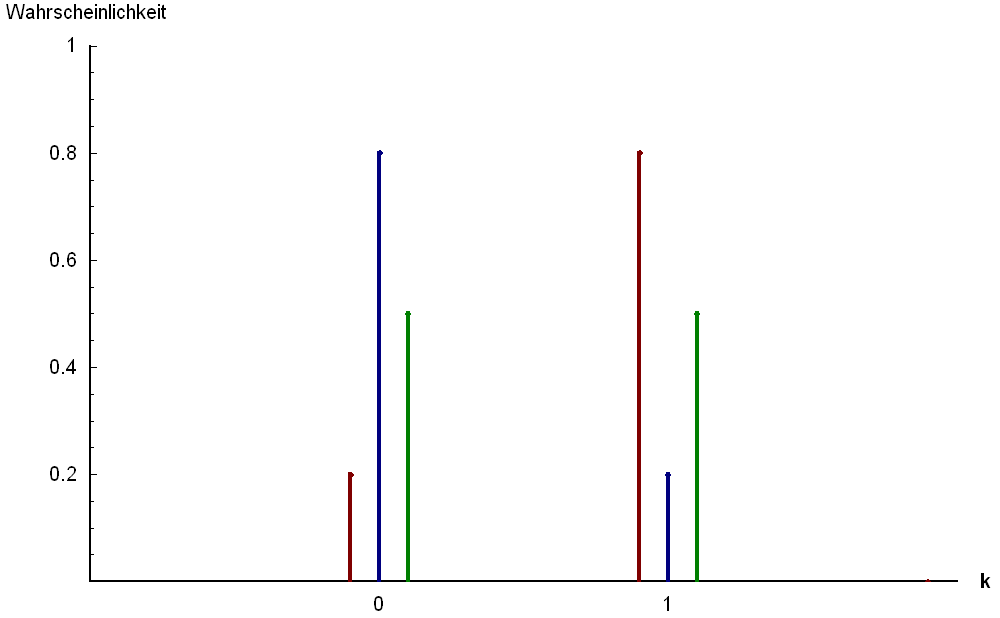
**What is probability distribution and its different types?**

Probability distributions give us an idea about the probability of the possibility of occurrence of an event. There are many different probability distributions that exist and that are dependent on the type of data set provided like discrete or continuous:

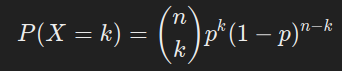
Discrete data is dealt with using:

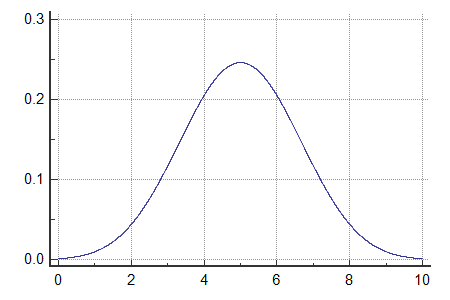
A) Bernoulli Distribution only accepts that there are only two discrete values 0 or 1 and is done one by one only on a singular event or trial



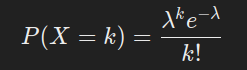


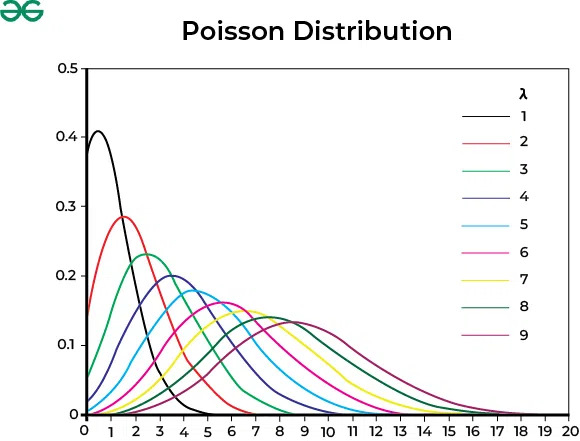
B) Binomial Distribution which is also utilized when dealing with discrete data, but “k” can take any positive integer value. Binomial distribution is a method which acts as a series of Bernoulli trials for n events:





C) Poisson Distribution also allows us to deal with multiple series of events/trails just like binomial distribution. However, it used in cases where the probability of an event occurrence is rare:

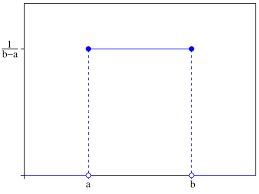


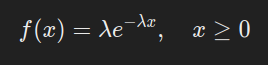


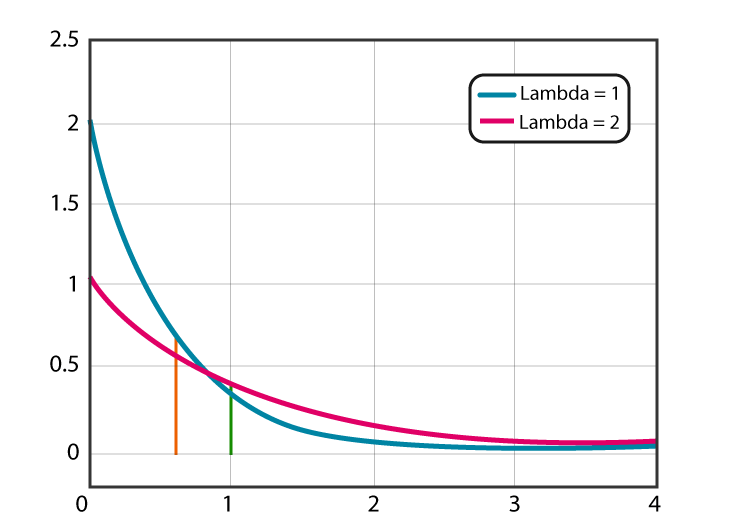
Continuous Probability Distributions: They are used for cases where variables can take a range of values which is infinite number of values due to decimals.

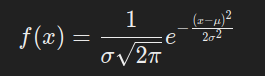
A) Uniform distribution is a distribution used to deal with variant data in a range[a,b] and it provides equal probability to each event/case.

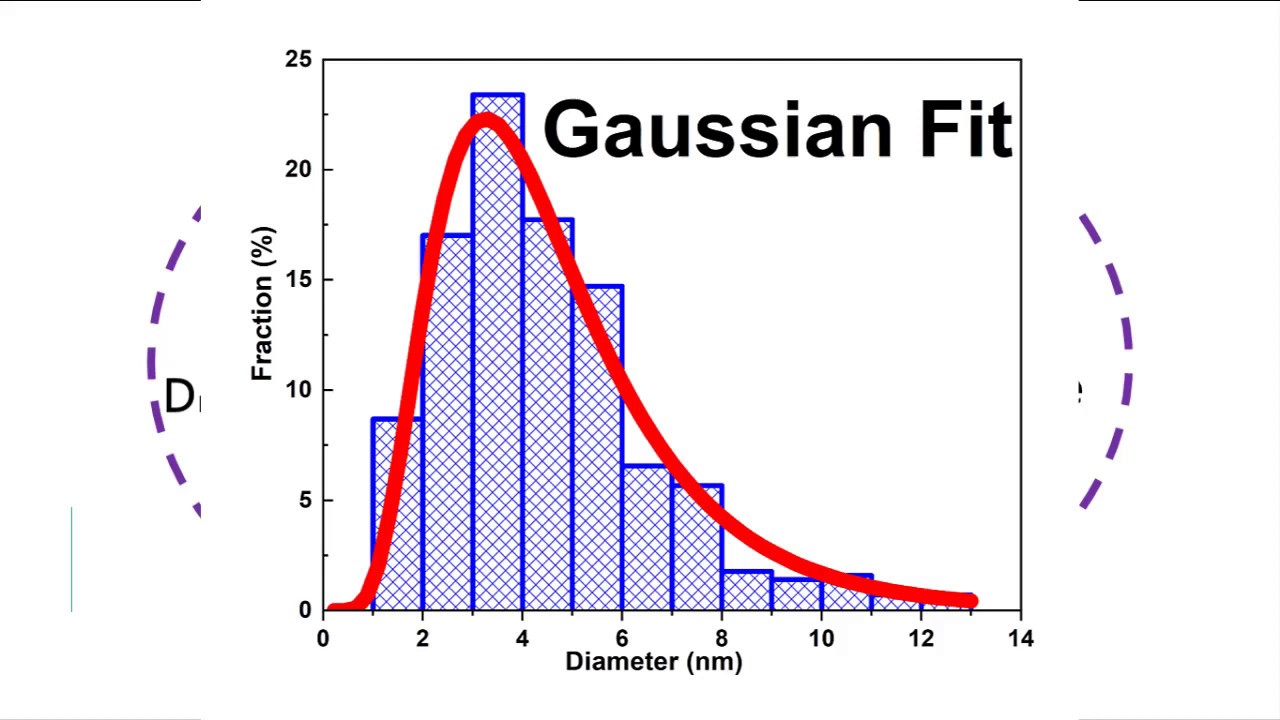




B) Exponential Distribution is used for use cases where x can be any positive decimal value and used in scenarios where we are dealing with a non-negative variable :

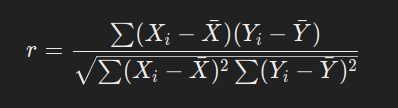


C) Gaussian Distribution it can accept any decimal value of x between –infinity and +infinity allowing it to be used in many algorithms since it has a universally accepting range: .



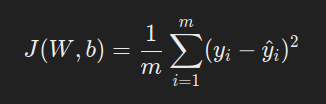
**What is correlation analysis, its significance, and the applications?**

Correlation analysis is a method to get the correlation between two quantities, ranging between –1 and 1 respectively showing negative and positive correlation and 0 showing no correlation between the two variable quantities. This is important in data analysis since it gives us an idea whether two quantities are related to each other or not and accordingly applying the relevant machine learning algorithms on it. It helps us grasp information about the data as to trends and spikes in the data. It is utilized in many sectors like finance and artificial intelligence since knowledge of the relationship between two quantities can help improve predictions from machine learning models. This relationship also helps to optimize the cost function in Neural network path planning algorithms since establishing the relationship between the necessary data can help optimize cost and increase efficiency. Correlation coefficient can be found out between two variables using the Pearson’s coefficient formula:



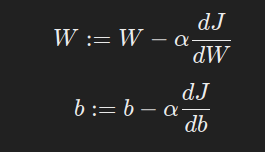
**What is regression in detail?**

Regression is a statistical technique used to model and analyze the relationship between dependent and independent variables. It predicts a continuous output based on input features. It works by trying to get a line/curve of best fit around the available data points, it does so by trying to optimize the line/curve by varying the weights and bias assigned to them using a process called gradient descent. The method works by calculating the cost function and accordingly making changes to the weights and biases to accordingly to minimize the cost function. Gradient descent is applied on the weights and biases with an appropriate number of epochs and learning rate which is important for minimalization of the cost function. The cost function is a measure of the total error between the predicted and actual value. There are many different types of cost functions with their specific use cases. Mean squared loss function is used since it is a simple cost function that is easy to implement with variety of use cases in machine learning. A dependent variable is fitted using a combination of independent variables weights and biases. The loss function between the predicted and actual value is calculated and the cost is calculated, and weights and biases are updated by subtracting their respective partial derivatives of cost function with respect to the variables and the cost function is rechecked and this process is repeated until cost function drops below a threshold. The following is the squared mean cost function being minimized



J is the cost function, w are the weights, and b is the bias, and m is the number of training examples.

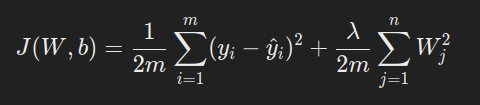
The following is updating methods of weights and biases:



**What is overfitting and underfitting? How to avoid it?**

Overfitting and underfitting is a problem that is encountered as a product line/curve fitting, Overfitting means that line is fitted overly towards the given training data sets which leads to inaccuracy when new model is used against a new test data set. Underfitting is when not enough data is provided causing the cost function to not be properly fitted i.e. line of best fit isn’t optimized for the data set which causes rise of inaccuracy. Overfitting and underfitting can be dealt with by changing the degree of features or number of features being taken into consideration. If overfitting we should try to ignore higher order exponential features or choose only those features which have the highest correlation coefficient. For underfitting we could take into consideration more terms of higher powers and increase the learning rate or the number of epochs. The problem of overfitting and underfitting can be avoided by getting more data so that there are more training examples for the machine learning model. We can also use a method called Regularization during gradient descent, Regularization is a process by which we can decrease the coefficients of the features hence forcing the model to keep on reducing magnitude of weights for the line of best fit, thus helping mellow out the problem of overfitting or underfitting.

Regularization: We can apply it on both w and b to force them to take smaller values. However applying regularization on w is sufficient for training machine learning models. The following is the updated cost function:



Regularization of only w also causes change in the formula for updating of the weights. The following is the updated partial derivative of cost function with respect to the weight.

