BIG DATA ANALYTICS

MC-5101 (*Till Mid-semester[Revised]*)

Dr. Sumit Kumar Tetarave

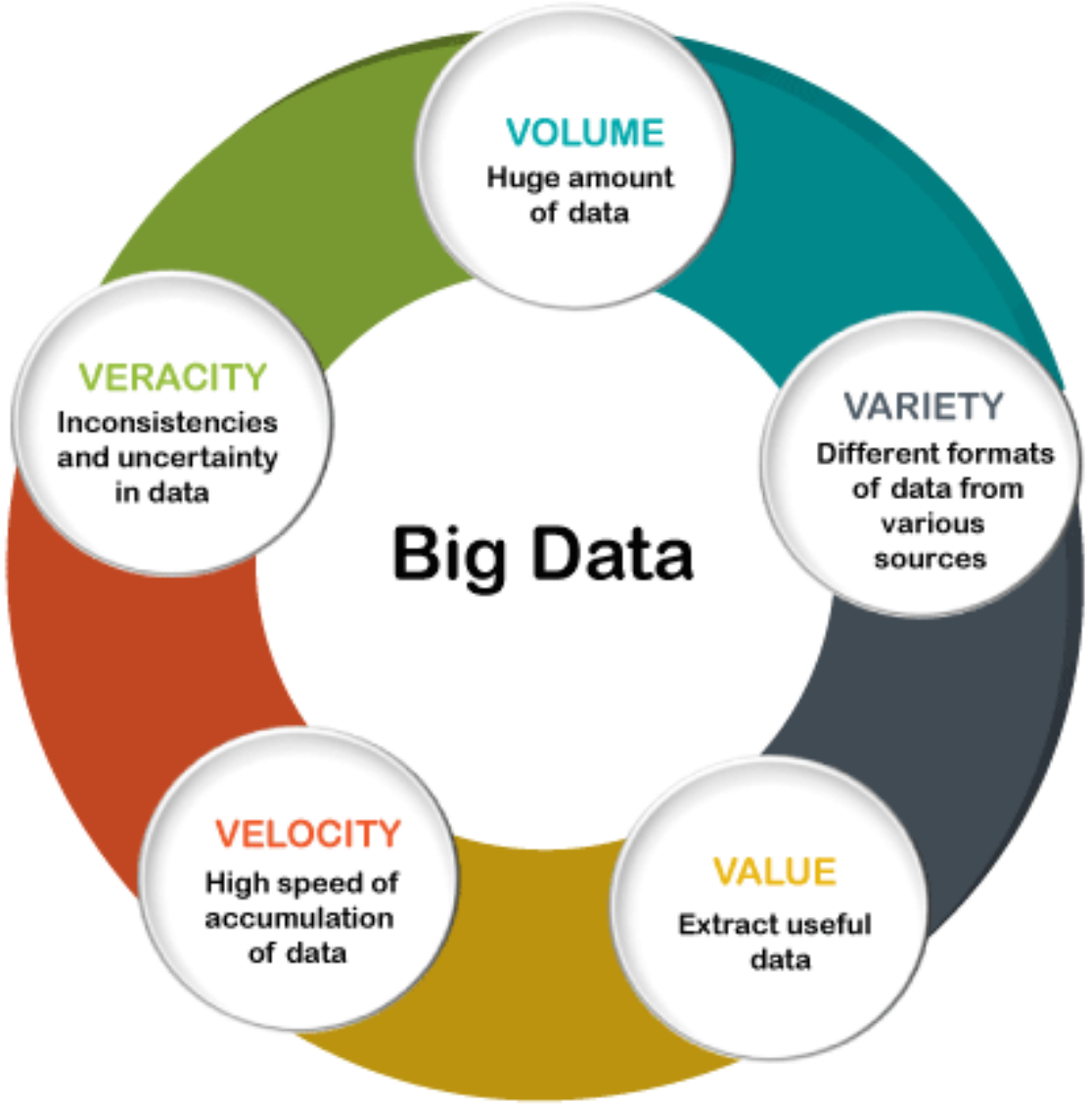
Study Material- 01

Study Content till Mid-Semester

1. Big Data Characteristics
2. Summary of Statistics
3. Data Preprocessing
4. Understanding Effectiveness of ML
5. How to choose the Right Statistical Test

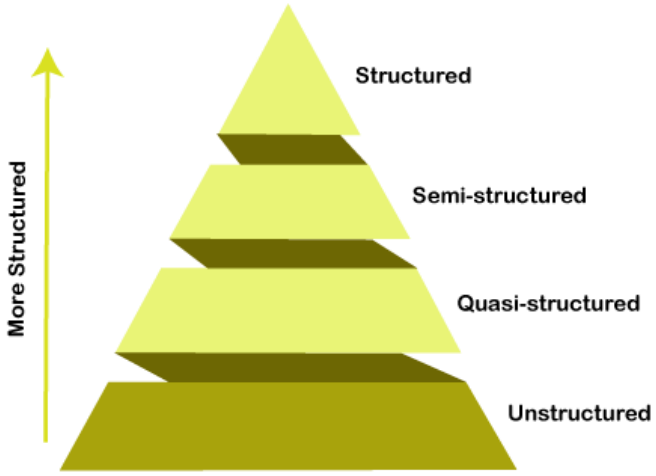
1. Big Data Characteristics

- Big Data contains a large amount of data that is not being processed by traditional data.
- There are five v's of Big Data that explains the characteristics.
  - **Volume**
  - **Veracity**
  - **Variety**
  - **Value**
  - **Velocity**



Big Data Characteristics…

- Volume : Big Data is a vast 'volumes' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more.
- Variety: Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources.



The data is categorized as below:

1.Structured data: In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

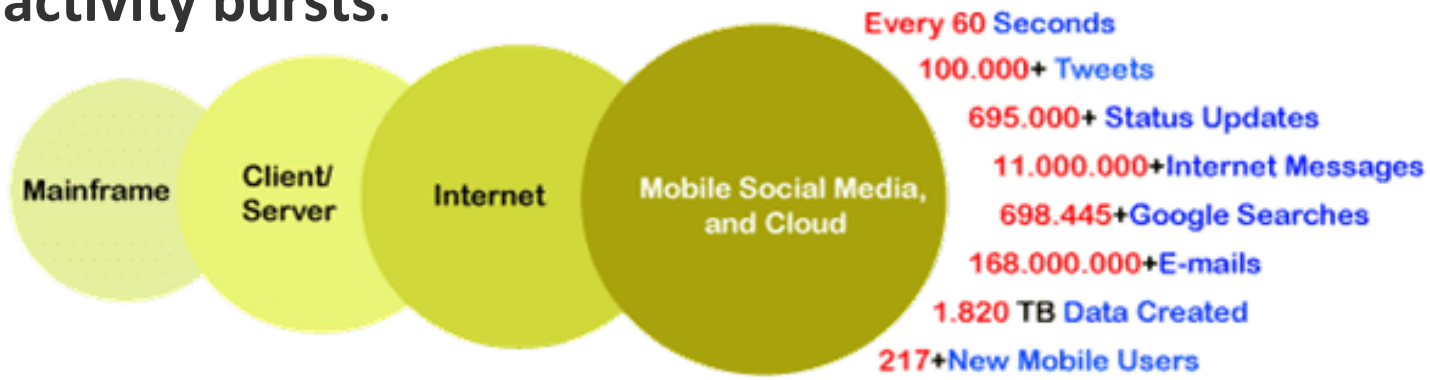
2.Semi-structured: In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

3.Quasi-structured Data: The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

4.Unstructured Data: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

Big Data Characteristics…

- Veracity: Veracity means how much the data is reliable.
  - For example, **Facebook posts** with hashtags.
- Value: It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.
- Velocity: Velocity creates the speed by which the data is created in **real-time**.
  - It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**.



Types of variables:

Quantitative (or Numerical) variables represent amounts of things (e.g. the number of trees in a forest).

- ✓Types of quantitative variables include:
  - **Continuous** (aka ratio variables): represent measures and can usually be divided into units smaller than one (e.g. 0.75 grams).
  - **Discrete** (aka integer variables): represent counts and usually can't be divided into units smaller than one (e.g. 1 tree).

- **Categorical variables** represent groupings of things (e.g. the different tree species in a forest).
  - ✓Types of categorical variables include:
    - **Ordinal**: represent data with an order (e.g. rankings).
    - **Nominal**: represent group names (e.g. brands or species names).
    - **Binary**: represent data with a yes/no or 1/0 outcome (e.g. win or lose).

Dataset Basics: How to represent mathematically

Train and Test using a Dataset

	Age(X_1)	Weight(X_2)	Blood Pressure(X_3)	Sugar(X_4)	Cholesterol (X_5)	Having Diabetes or not (Y)
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	yes
2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	no
	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	yes
	⋮	⋮	⋮	⋮	⋮	⋮
1000	$x_{1000,1}$	$x_{1000,2}$	$x_{1000,3}$	$x_{1000,4}$	$x_{1000,5}$	no

New Patient	$x_{new,1}$	$x_{new,2}$	$x_{new,3}$	$x_{new,4}$	$x_{new,5}$	?

Example of *2-Class Classification Problem*

- Because our answer consists of two options either yes or no.
- Recognising a disease (Diabetes) from past data of patients.

Terminologies

- $Y \rightarrow$ Response (Statistics)/ Target (Machine Learning)
- $X_1, X_2, \ldots, X_p \rightarrow$ Predictors (Statistics)/ Features (Machine Learning)/ Attributes (Database)
  - $x_1, x_2, \ldots, x_p \; and \; Y \rightarrow$Sample/ Example/ Pattern (Machine Learning)
- $X_j, j = 1,2,\ldots,p$ denote ***generic variables*** representing a predictor/ feature. X_j is not a vector rather just a variable.
- Y denotes generic variable representing response/ target. Y is not a vector rather just a variable.
- $x_i$is a column vector containing feature values of ith sample.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}, i = 1,2,\cdots,N.$$

$$x_i \in \mathbb{R}^p$$

- The values of X_j (jth feature) in the set of samples are denoted as x_{ij}, where $x_{ij}$is the value of jthfeature in ithsample. that is in x_{ij}, i represent the sample number and j represent the feature number.

$$i = 1,2,\cdots,N, j = 1,2,\cdots,p.$$

Note

$$x_i \in \mathbb{R}^p, y_i \in \mathbb{R} \; and \; x_{ij} \in \mathbb{R}.$$

Dataset Basics…

	Image of Handwritten Digits (Matrices or Column Vectors)	Actual Digit (Y)
1	⋮	0
2	⋮	1
⋮	⋮	0
⋮	⋮	5
⋮	⋮	⋮
⋮	⋮	8
N	⋮	9
New Image	⋮	?

Example of *10-Class Classification Problem*

- Because our answer consists of ten options that is digits from 0 to 9.
- Recognizing hand written digits.

Definition
Classification Problem: M-Class Classification Problem
Given a data set
$$D = \{(x_i, y_i)\}_{i=1}^{N}$$
or
$$D = \{(x_i, y_i)\} \text{ where } i = 1, 2, 3, \cdots, N$$
where, $x_i \in \mathbb{R}^p, y_i \in \{1, 2, 3, \cdots, M\}$ and a new sample $x_0 \in \mathbb{R}^p$ determine the class to which x_0 is associated with based on D.
Input: D, x_0
Output: y_0

Risk free Interest Rate	Price of Gold	Price of Petrol	USD/ INR Exchange Value	Stock Index (Number)
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	?

Example of "*Stock Index Prediction*"

- We get the value of Stock Index as a number that is a Real Number.
- So, it is a Regression Problem.

Definition
Regression Problem:
Given a data set
$$D = \{(x_i, y_i)\}_{i=1}^{N}$$
or
$$D = \{(x_i, y_i)\} \text{ where } i = 1, 2, 3, \cdots, N$$
of N patterns, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and a new predictor vector $x_0 \in \mathbb{R}^p$ determine the response $y_0 \in \mathbb{R}$ associated with the pattern x_0 based on D.

Dataset Basics…

Movie Name	*Persons*				I want or not (Y)
	$user_1$	$user_2$	⋯	$user_m$	
m_1	⋮	⋮	⋯	⋮	yes
m_2	⋮	⋮	⋯	⋮	yes
m_3	⋮	⋮	⋯	⋮	no
⋮	⋮	⋮	⋯	⋮	⋮
m_r	⋮	⋮	⋯	⋮	yes
New Movie	⋮	⋮	⋯	⋮	?

Example of *2-Class Classification Problem*

- Because our answer consists of two options either yes or no.
- Movie Recommendation.

	Advertising Budget on			Sales (Y)
	TV (X_1)	Radio (X_2)	News Paper (X_3)	
x_1	x_{11}	x_{12}	x_{13}	y_1
x_2	x_{21}	x_{22}	x_{23}	y_2
⋮	⋮	⋮	⋮	⋮
x_i	x_{i1}	x_{i2}	x_{i3}	y_i
⋮	⋮	⋮	⋮	⋮
x_N	x_{N1}	x_{N2}	x_{N3}	y_N

Example of *Regression Problem*

- Predicting Sales based on advertising budget on three media:
- TV, Radio, News Paper.

Image of Cloud		Quantity of Rainfall (in cm)
3rd Week of June	4th Week of June	
2020	2020	?
2019	2019	1.5 cm
2018	2018	2.00 cm
2017	2017	2.5 cm
⋮	⋮	⋮
2000	2000	1.003 cm

➢ Predicting Rainfall in a certain week.

Example of "*Time Series Prediction Problem*"

- Input is an Image and Output is a Real Number.
- So, it is a Regression Problem.

2. Summary of Statistics

What is Statistics?

Statistics is the science of collecting data and analysing them to infer proportions (sample) that are representative of the population. In other words, statistics is interpreting data in order to make predictions for the population.

Branches of Statistics:

There are two branches of Statistics.

DESCRIPTIVE STATISTICS : Descriptive Statistics is a statistics or a measure that describes the data.

INFERENTIAL STATISTICS : Using a random sample of data taken from a population to describe and make inferences about the population is called Inferential Statistics.

Descriptive Statistics

Descriptive Statistics is summarising the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalisation or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

1. Measures of Central Tendency
2. Measures of Dispersion (or Variability)

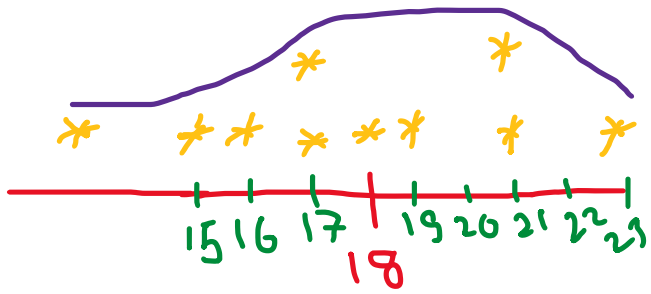
Measures of Central tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the centre of the data. This one number summary is of three types. a) Mean b) Median c) Mode

a. **Mean :** Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23



$$Mean = \frac{17+16+21+18+15+17+21+19+11+23}{10} = \frac{178}{10} = 17.8$$

$$\text{New average} = \frac{(\text{Old average} \times \text{Number of data points}) + \text{New data point}}{\text{Number of data points} + 1}$$

Q.34	The sample average of 50 data points is 40. The updated sample average after including a new data point taking the value of 142 is _____.

Solution: To find the updated sample average after including a new data point, you need to use the formula for calculating the sample average:

$$\text{New average} = \frac{(\text{Old average} \times \text{Number of data points}) + \text{New data point}}{\text{Number of data points} + 1}$$

$$\text{New average} = \frac{(40 \times 50) + 142}{50 + 1} = \frac{2142}{51} \approx 42$$

Measures of Central tendency (Median)

b. **Median** : Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

- If the number of observations is odd, the median is given by the middle observation in the sorted form.
- If the number of observations are even, median is given by the mean of the two middle observations in the sorted form.

An important point to note is that the order of the data (ascending or descending) does not affect the median.

Measures of Central tendency (Median) …

To calculate Median, let's arrange the data in ascending order.

11, 15, 16, 17, 17, 18, 19, 21, 21, 23

50% 50%

Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17+18}{2} = 17.5$$

Measures of Central tendency (Mode)

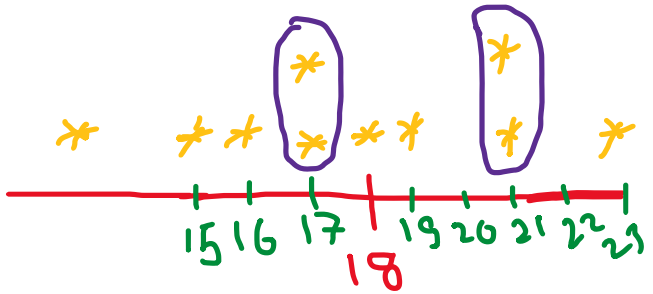
c. **Mode :** Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.

- If there is only one number that appears the maximum number of times, the data has one mode, and is called Uni-modal.
- If there are two numbers that appear the maximum number of times, the data has two modes, and is called Bi-modal.
- If there are more than two numbers that appear the maximum number of times, the data has more than two modes, and is called Multi-modal.

Measures of Central tendency (Mode)…

Consider the following data points.

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

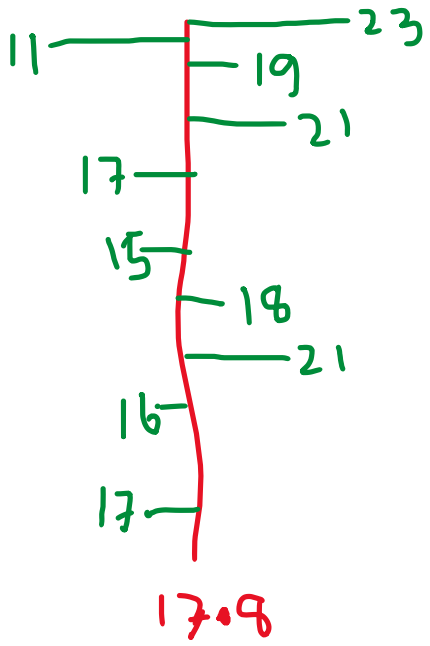


Mode is given by the number that occurs the maximum number of times. Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Mean is the accurate measure to describe the data when we do not have any outliers present. Median is used if there is an outlier in the dataset. Mode is used if there is an outlier AND about ½ or more of the data is the same.

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)



1. **Absolute Deviation from Mean** — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describes the variation in the data set, in the sense that it tells the average absolute distance of each data point in the set. It is calculated as

$$\textit{Mean Absolute Deviation} = \frac{1}{N}\sum_{i=1}^{N} |X_i - \bar{X}|$$

17, 16, 21, 18, 15, 17, 21, 19, 11, 23

0.8, 1.8, 3.2, 0.2, 2.8, 3.2, 1.2, 6.8, 5.2 = 25.2/10 = 2.52

Measures of Dispersion (or Variability)…

2. **Variance** — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set. It is calculated as

$$Variance = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

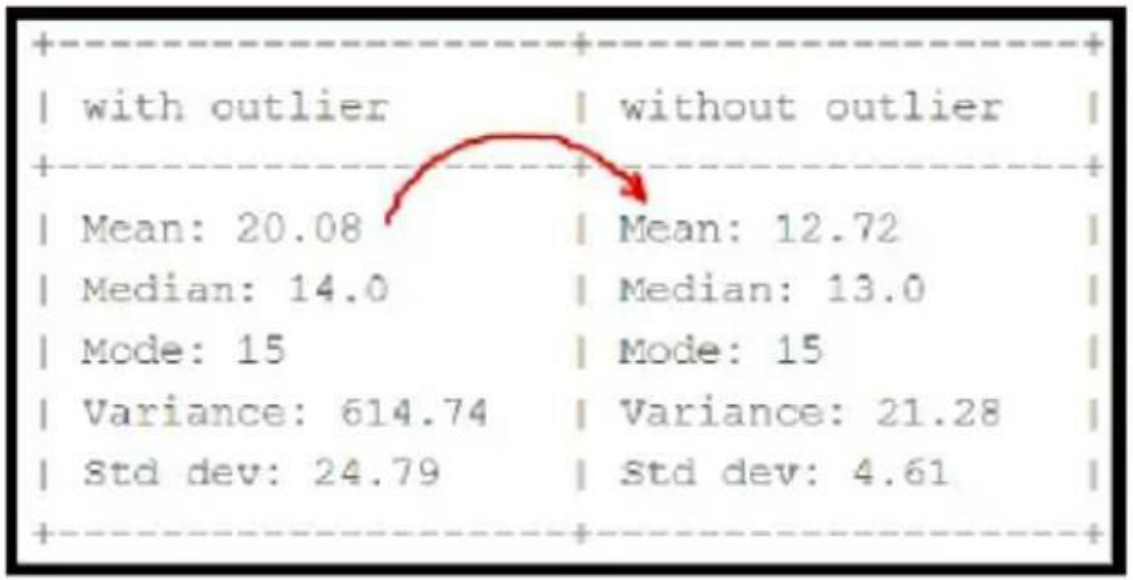
3. **Standard Deviation** — The square root of Variance is called the Standard Deviation. It is calculated as

$$Std\ Deviation = \sqrt{Variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

'Mean' is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.

Example:

Consider a small dataset, sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]. By looking at it, one can quickly say '101' is an outlier that is much larger than the other values.



with outlier	without outlier
Mean: 20.08	Mean: 12.72
Median: 14.0	Median: 13.0
Mode: 15	Mode: 15
Variance: 614.74	Variance: 21.28
Std dev: 24.79	Std dev: 4.61

fig. Computation with and without outlier

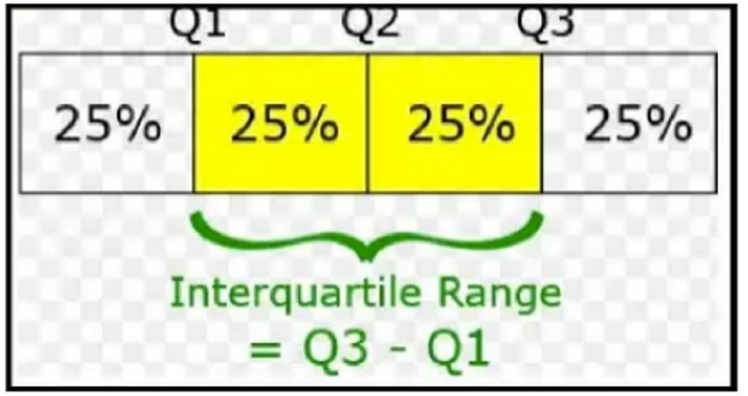
From the above calculations, we can clearly say the Mean is more affected than the Median.

Measures of Dispersion (or Variability)…

4. **Range** — Range is the difference between the Maximum value and the Minimum value in the data set. It is given as

$$Range = Maximum - Minimum$$

5. **Quartiles** — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.



- 25% of the data points lie below Q1 and 75% lie above it.
- 50% of the data points lie below Q2 and 50% lie above it. Q2 is nothing but Median.
- 75% of the data points lie below Q3 and 25% lie above it.

Interquartile Range and Box and Whisker Plot

Q1, Q2, Q3 and Inter-quartile range ? Draw the Box for it.

4, 17, 7, 14, 18, 12, 3, 16, 10, 4, 4, 11

Put them in order:

3, 4, 4, 4, 7, 10, 11, 12, 14, 16, 17, 18

Cut it into quarters:

3, 4, 4 | 4, 7, 10 | 11, 12, 14 | 16, 17, 18

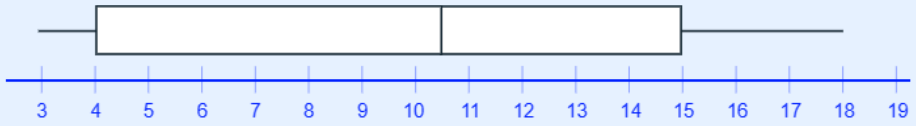
In this case all the quartiles are between numbers:

- Quartile 1 (Q1) = (4+4)/2 = **4**
- Quartile 2 (Q2) = (10+11)/2 = **10.5**
- Quartile 3 (Q3) = (14+16)/2 = **15**

Also:

- The Lowest Value is **3**,
- The Highest Value is **18**

So now we have enough data for the **Box and Whisker Plot**:



And the **Interquartile Range** is:

Q3 − Q1 = 15 − 4 = **11**

Q1, Q2, Q3 and Inter-quartile range ? Draw the Box for it.

Example2: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

•Quartile 1 (Q1) = **4**
•Quartile 2 (Q2), which is also the Median, = **5**
•Quartile 3 (Q3) = **7**

To define the outlier base value is defined above and below datasets normal range namely Upper and Lower bounds, define the upper and the lower bound (1.5*IQR value is considered) :

upper = Q3 +1.5*IQR

lower = Q1 – 1.5*IQR

In the above formula as according to statistics, the 0.5 scale-up of IQR (new_IQR = IQR + 0.5*IQR) is taken.

Measures of Dispersion (or Variability)…

6. **Skewness** — The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

$$Skewness = \frac{3\,(Mean - Median)}{Std\ Deviation}$$

Positive Skew — This is the case when the tail on the right side of the curve is bigger than that on the left side. For these distributions, mean is greater than the mode.

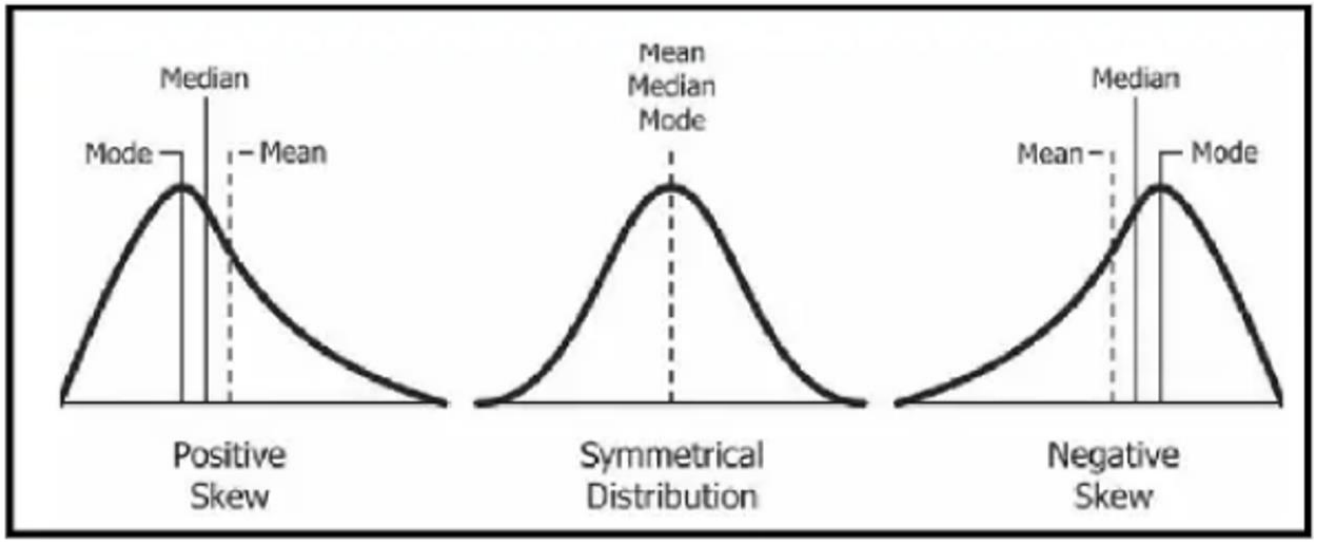
Negative Skew — This is the case when the tail on the left side of the curve is bigger than that on the right side. For these distributions, mean is smaller than the mode.

Calculating Skewness

$$Skewness = \frac{3\,(Mean - Median)}{Std\ Deviation}$$

The most commonly used method of calculating Skewness is

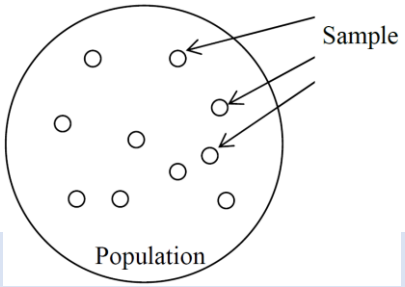
If the skewness is zero, the distribution is symmetrical. If it is negative, the distribution is Negatively Skewed and if it is positive, it is Positively Skewed.



Some of Basic Methods of Data Analysis:

- Descriptive Analysis
- Exploratory Analysis
- Inferential Analysis
- Predictive Analysis

Term	Used for	Example
Statistic	Statistic is computed for the Sample	Sample mean ($\bar{x}$), Sample Standard deviation (s), Sample size (n)
Parameter	Parameters are predicted from sample and are about the Population	Population mean (μ), Population Standard deviation (σ), Population size (N)



- Descriptive Analysis- is used to present basic summaries about data.

- Exploratory Analysis- Instead of performing the final analysis,
  - we may like to explore the data for possible relationships using exploratory data analysis.

- Inferential Analysis- Inferential analysis is performed to answer the question that
  - what is the probability of that the results obtained from an analysis can be applied to the entire population.

- Predictive Analysis- analysis. The prescriptive analysis aims to take predictions one step forward and
  - suggest solutions to present and future issues.

Sampling:

- In general the size of data that is to be processed today is quite large.
- One of the key objectives of statistics, which uses sample data, is
  - to determine the statistic of the sample and
  - find the probability that the statistic developed for the sample
    - ✓ would determine the parameters of population with a specific percentage of accuracy.
- a good sample is representative of its population.

Example. *You are given the position in a factory before and after the settlement of an industrial dispute. Comment on the gains or losses from the point of view of workers and that of management.*

	Before	*After*
No. of workers	*3,000*	*2,900*
Mean wage (in Rs.)	*220*	*230*
Median wage (in Rs.)	*250*	*240*
Standard deviation (in Rs.)	*30*	*26*

Solution. On the basis of the above data we are in a position to make the following comments:

(*a*) The number of workers after the settlement has decreased from 3000 to 2900. This is a definite loss to the persons thrown out or retrenched. It may also be a loss to the management if their retrenchment affects the efficiency of work adversely.

Solution:

(*b*) We know that:

$\Rightarrow$ **Total wages paid = (Average wage) × (Total No. of workers)**

Total wages paid by the management before the settlement = 3000 × 220 = **Rs. 660000**

Total wages paid by the management after the settlement = 2900 × 230 = **Rs. 667000**

Thus the total wages paid by the management have gone up after the dispute (the additional wage bill being Rs. 7000), although the number of workers has been reduced from 3000 to 2900. Thus the average wage per worker has increased after the settlement , **which is a distinct advantage to the workers.**

It may be pointed out that the increased wages paid by the management (Rs. 7000) should not be viewed as a disadvantage to the management unless we have definite reasons to believe that the efficiency and productivity have not gone up after the settlement. However, the loss to the management due to higher wage bill, will be more than compensated if after the settlement, there is a increase in the efficiency of the workers or/and increase in productivity.

Solution:

(*c*) Though the number of workers has decreased from 3000 to 2900 after the settlement, the average wage per worker has gone up from Rs. 220 to Rs. 230. This might probably be a consequence of the retrenchment of casual labour or temporary labour working on daily wages or so with relatively lower wages.

(*d*) The median wage after the settlement has come down from Rs. 250 to Rs. 240. This implies that before the dispute upper 50% of the workers were getting wages above Rs. 250 whereas after the settlement they get wages only above Rs. 240.

(*e*) Using the empirical relation between mean, median and mode (for a moderately asymmetrical distribution), *viz*..

$$\textbf{Mode} = 3\,\textbf{Median} - 2\,\textbf{Mean},$$

we get:

$$\textbf{Mode (before settlement)} = 3 \times 250 - 2 \times 220 = \textbf{Rs. 310.}$$

$$\textbf{Mode (after settlement)} = 3 \times 240 - 2 \times 230 = \textbf{Rs. 260.}$$

Thus, the modal wage has come down from Rs. 310 (before settlement) to Rs. 260 (after settlement). Hence, **after the settlement there is concentration of wages around a much smaller value.**

Solution:

(*f*) C.V. (before settlement) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{100 \times 30}{220} = 13.64\%.$

C.V. (after settlement) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{100 \times 26}{230} = 11.30\%.$

Since C.V. has decreased from 13.64 % to 11.30 %, so the distribution of wages has become less variable *i.e.*, more consistent or uniform after the settlement of the dispute. Thus after the settlement there are less disparities in wages and from management point it will result in greater satisfaction to the workers.

(*g*) Since we are given mean and median, we can calculate Karl Pearson's coefficient of skewness for studying the symmetry of the distribution of wages before the settlement and after the settlement,

$$SK \text{ (before settlement)} = \frac{3(\text{Mean} - \text{Median})}{\sigma} = \frac{3(220 - 250)}{30} = -3.$$

$$SK \text{ (after settlement)} = \frac{3(\text{Mean} - \text{Median})}{\sigma} = \frac{3(230 - 240)}{26} = -1.15.$$

SK (before settlement) is highly negative but *SK* (after settlement) is moderately negative

$\Rightarrow$ after the settlement, there is a less longer tail towards the left. **Thus the number of workers getting lower wages has increased.**

Measures of Dispersion (or Variability)

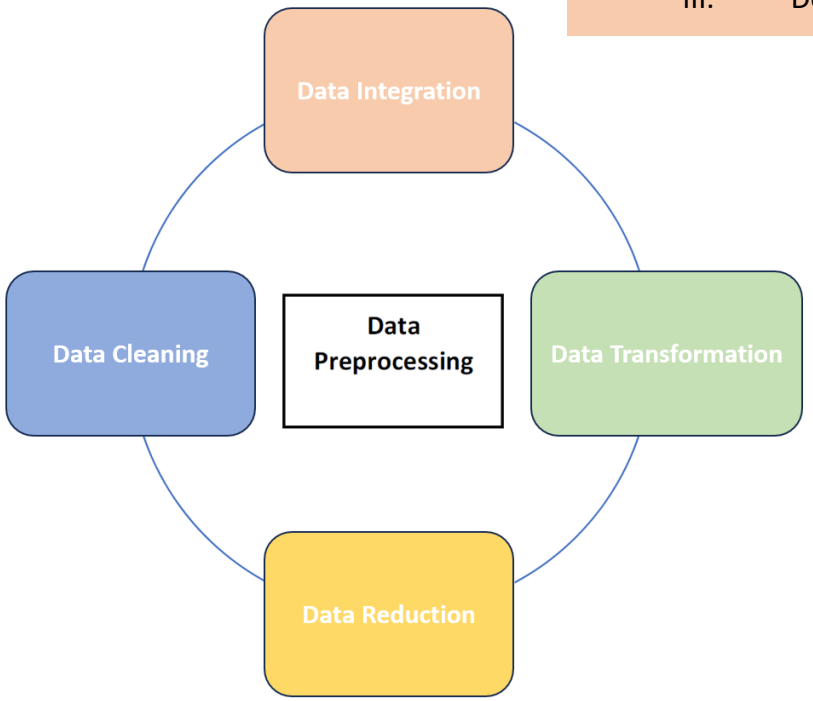
1. Absolute Deviation from Mean
2. Variance
3. SD
4. Range
5. Quartiles
6. Skewness

Next:
3. Data Preprocessing

- Data Cleaning accounting for:
  - Missing values
    - Use a global constant to fill in the missing values
    - Use the attribute mean to fill in the missing value
    - Use the mean attribute for all samples belonging to the same class as the given tuple
    - Using the most likely meaning to fill in the missing value
  - Noisy Data
    - An outlier can be treated as noise
    - Correcting inconsistent data points, and Smoothing noisy data.
      - For numeric values, we can use a scatter plot or box plot to identify outliers
    - Methods for solving it
      - Binning
      - Regression
      - Clustering

- Some approaches to integrate data:
  - Data consolidation
  - Data virtualization
  - Data propagation

- During data migration, there are certain issues to be considered namely:
  - I. Schema Integration and Object Matching
  - II. Redundancy
  - III. Detection and Resolution of data value conflicts.



- Techniques used for data transformation:
  - Smoothing, Aggregation,
  - Normalization
    - Min-max normalization
    - Z-score normalization
    - Normalization by decimal scaling
  - Discretization, Generalization,
  - Feature Construction and Concept hierarchy generation

- Following are some of the strategies for Data Reduction:
  - ✓***1. Data cube aggregation***, which applies aggregation of data during data cube construction.
  - ✓2. ***Set of sub-set attributes***, which may be detected and removed by obsolete, weakly significant or redundant measurements.
    - ✓**Attribute Subset Selection Methods**-
      - ✓1. Stepwise Forward Selection
      - ✓ii. Stepwise Backward Elimination
      - ✓iii. Combination of Forward Selection and Backward Elimination
      - ✓iv. Decision Tree Introduction
  - ✓3. ***Dimensionality reduction*** where encoding approaches are used to reduce the data collection size to a minimum.
    - ✓It can be divided into
      - ✓i. ***feature selection*** and
      - ✓ii. ***feature extraction***
    - ✓Dimensionality Reduction Methods
      - ✓Principal component analysis (PCA)
      - ✓High correlation filter
      - ✓Missing values ratio
      - ✓Low variance filter
      - ✓Random forest
  - ✓4. ***Reduction of numbers when alternative***, smaller data representation replaces or predicts information.
    - ✓There are two ways to minimize numbers
      - ✓i. Parametric Methods
      - ✓ii. Non-Parametric Methods
  - ✓**5. *Discretization***
    - ✓Following are the categories of Discretization:
      - ✓i. Supervised Discretization
      - ✓ii. Unsupervised Discretization or Splitting
      - ✓iii. Top-Down Discretization or Splitting
      - ✓iv. Bottom – up Discretization or merging
  - ✓***6. Category hierarchy generation*** where raw data values for the attributes are replaced by ranges or higher logical levels.
    - ✓Five methods for concept hierarchy generation
      - ✓i. Binning
      - ✓ii. Histogram Analysis
      - ✓iii. Entropy – Based Discretization
      - ✓iv. Interval merging by χ^2– Analysis
      - ✓v. Cluster Analysis
      - ✓vi. Discretization by Intuitive Partitioning

Purpose of Data Transformation

- Data transformation is the process of converting raw data into a format or structure that would be more suitable for model building and also data discover in general.
- The process of data transformation can also be referred as extract/transform/load (ETL).
  - The extract phase involves identifying and pulling data from the various source systems that create data and then moving data to a single repository.
  - Next, the raw data is cleaned, if needed.
  - It is then transformed into a target format that can be fed into operational system or into a data warehouse or another repository
    - to use in business intelligence and analytics applications.
- The data are transformed in ways that are ideal for analyzing the data.

Steps of Data Transformation

- The data transformation involves steps that are:

1. Smoothing
2. Aggregation
3. Generalization
4. Discretization
5. Normalization
   - Min-max normalization
   - Z-score normalization
   - Normalization by decimal scaling
6. Attribute and feature construction
   - New attributes constructed from the given ones.

Steps of Data Transformation…

- **Smoothing:** It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns
- **Aggregation**: Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used.
- **Generalization**:It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

Steps of Data Transformation…

- Discretization: It is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals, also called bins,
  - that span the range of the variable values.
- Discretization is used to change the distribution of skewed variables and to minimize the influence of outliers, and
  - hence improve the performance of some machine learning models.
- For the variable X, the interval width may be the following binning methods:

1. **Group-wise:** $$Width = \frac{Max(X) - Min(X)}{Bins}$$

For example:
If the values of the variable vary between 0 and 100 and we want 5 Bins, then,
Width = (100-0) / 5 = 20

Bin 1: 1 - 20
Bin 2: 21 - 40
Bin 3: 41 - 60
Bin 4: 61 - 80
Bin 5: 81 - 100

or

2. **Equal-frequency** : $$Data\ in\ a\ Bin = \frac{Number\ of\ Data}{Bins}$$

For example:
Data = 2, 4, 6, 8, 10, 12, 14, 16, 19
Bins = 3, then,
Bin Edge = (1.9- 6], (6-12], (12- 19]

Bin 1: 2, 4, 6
Bin 2: 8, 10, 12
Bin 3: 14, 16, 19

Discretization: Use

- Use1: The first and final bins ([0-20 and 80-100] or Fixed Bin =3) can be expanded to accommodate outliers, that is,
  - values under 0 or greater than 100 would be placed in those bins as well, by extending the limits to minus and plus infinity.
- Use2: Binning uses the "neighborhood" to smooth the storage value of the records. That's the value around it by following methods:
  i. Smoothing by bin mean
  ii. Bin borders smoothing

For example:

Data: 2, 4, 6, 8, 10, 12, 14, 16, 19, 20

Bins: 3, then do a Data Discretization.

✓ Bin Edge = (1.9-8], (8-14], (14-20]

Using Equal-frequency:

Bin 1: 2, 4, 6, 8

Bin 2: 10, 12, 14

Bin 3: 16, 19, 20

i) Smoothing by bin mean

Bin 1: 5, 5, 5, 5

Bin 2: 12, 12, 12

Bin 3: 18.3, 18.3, 18.3

✓ Each bin values are replaced by the respective bin mean value.

ii) Bin borders smoothing

Bin 1: 2, 2, 8, 8

Bin 2: 10, 10, 14

Bin 3: 16, 19, 20

✓ The maximum and minimum values are defined as the boundary values when the bin boundary is smoothed.

✓ Each bin value is then replaced by the closest limit value.

Steps of Data Transformation… (Normalization)

- **Normalization:** Data normalization involves converting all data variables into a given range. Some of the techniques that are used for accomplishing normalization are:

  i) **Min–max normalization**: This transforms the original data linearly.

  ii) **Z-score normalization**: In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation.

  iii) **Normalization by decimal scaling**: It normalizes the values of an attribute by changing the position of their decimal points

Min – Max Normalization

- Which translates the initial data linearly.
- Suppose min X is the minimum and max X is the maximum limit of the variable.
- We've got the formula

$$v' = \frac{v - minX}{maxX - minX}(\text{new\_}max_X\text{-new\_}min_X) + \text{new\_}min_X$$

- where **v** is the value that you want to plot in the new set.
- v' is the fresh meaning you get when you normalize the old value.

Min – Max Normalization: Example

- Assume the attribute revenue minimum and maximum values are 1000 and 16000 respectively. Diagram income in the [0.0, 1.0] range. For salaries a value of 14000 is converted into a minimum – maximum standardization.

$$= \frac{14000-1000}{16000-1000}\ (1.0 - 0) + 0$$

$$=0.866$$

Z-Score Normalization

- The value of a variable (X) in the Z-Score Normalization or Zero-Median Normalization is uniform according to the mean of the X and its standard deviation.
- The value $\boldsymbol{v}$ of the X attribute is computer-standardized to $\boldsymbol{v'}$

$$\boldsymbol{v'} = \frac{\boldsymbol{v} - \overline{\boldsymbol{X}}}{\boldsymbol{\sigma X}}$$

Where $\overline{X}$ and $\boldsymbol{\sigma}$ are the mean and standard deviations respectively of the X attribute.

This method is helpful if the true minimum and maximum attribute X are not certain or if outliers surpass the minimum – maximum normalization.

Z-Score Normalization: Example

- Suppose that the average and standard deviation in attribute revenue number are respectively 22,000 and 7,000. In the case of Z-Score standardization, the revenue of 42,000 is transferred to Z-Score.

$$\frac{42000 - 22000}{7000} = 2.85$$

Q.27	Let the minimum, maximum, mean and standard deviation values for the attribute *income* of data scientists be ₹46000, ₹170000, ₹96000, and ₹21000, respectively. The z-score normalized *income* value of ₹106000 is closest to which **ONE** of the following options?
(A)	0.217
(B)	0.476
(C)	0.623
(D)	2.304

Sol:

Z-score = (106000-96000)/21000

= 10000/21000= 0.476

Decimal Scaling Normalization

- Standardizes the variable values by changing their decimal position.
- We can measure the number of points that the decimal point is passed by the absolute maximum value of the X attribute.
- The value of v of the X attribute is computer-standardized to as:

$$v' = \frac{v}{10^j}$$

J is such an integer that maximum.

Example 1:

CGPA	Formula	CGPA Normalized after Decimal scaling
2	2/10	0.2
3	3/10	0.3

Here maximum value of CGPA is 3 so we can convert it to a decimal by dividing by 10. Why 10?

- we will count total numbers in our maximum value and then put 1 and after 1 we can put zeros equal to the length of the maximum value.

Example 2:

Salary bonus	Formula	CGPA Normalized after Decimal scaling
400	400 / 1000	0.4
310	310 / 1000	0.31

Dimensionality Reduction

- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and
  - the process to reduce these features is called dimensionality reduction.
- Handling the high-dimensional data is very difficult in practice, commonly known as
  - the *curse of dimensionality*.
- If the machine learning model is trained on high-dimensional data, it becomes overfitted and results in poor performance.
- Hence, it is often required to reduce the number of features, which can be done with dimensionality reduction.
- Some benefits of applying dimensionality reduction technique to the given dataset are given below:
  - By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
  - Less Computation training time is required for reduced dimensions of features.
  - Reduced dimensions of features of the dataset help in visualizing the data quickly.
  - It removes the redundant features (if present) by taking care of multi-collinearity.

Techniques for Dimensionality Reduction

- Dimensionality reduction is accomplished based on either ***feature selection*** or ***feature extraction***.
- ***Feature selection*** is based on omitting those features from the available measurements which do not contribute to class separability. In other words, redundant and irrelevant features are ignored.
    a) **Variance Thresholds**
    b) **Correlation Thresholds**
    c) **Genetic Algorithms**
    d) **Stepwise Regression-** This has two types: ***forward*** and ***backward***.
- ***Feature extraction***, Feature extraction is for creating a new, smaller set of features that still captures most of the useful information. This can come as supervised (e.g. LDA) and unsupervised (e.g. PCA) methods.
    a) **Principal Component Analysis (PCA)**
    b) **Linear Discriminant Analysis (LDA)**

Feature selection:

a) **Variance Thresholds:** This technique looks for the variance from one observation to another of a given feature and then
   - if the variance is not different in each observation according to the given threshold,
     - feature that is responsible for that observation is removed.

b) **Correlation Thresholds**: We first calculate all pair-wise correlations. Then, if
   - the correlation between a pair of features is above a given threshold,
     - we remove the one that has larger mean absolute correlation with other features.

- Like the previous technique, this is also based on intuition and hence the burden of tuning the thresholds in such a way that the useful information will not be neglected, will fall upon the user.
- Because of those reasons, algorithms with built-in feature selection or algorithms like PCA(Principal Component Analysis) are preferred over this one.

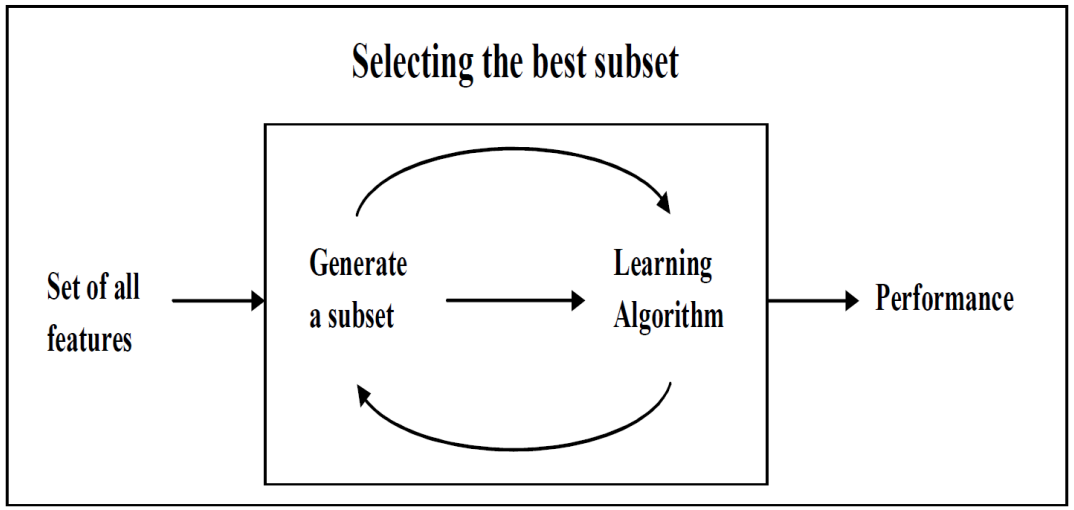
a) **Genetic Algorithms**: They are search algorithms that are inspired by evolutionary biology and natural selection, combining mutation and cross-over to efficiently traverse large solution spaces.

- Genetic Algorithms are used to find an optimal binary vector, where each bit is associated with a feature.
  - ✓ If the bit of this vector equals 1, then the feature is allowed to participate in classification.
  - ✓ If the bit is a 0, then the corresponding feature does not participate.

Feature selection…

d) Stepwise Regression: This is a greedy algorithm and commonly has a lower performance than the supervised methods such as regularizations etc.

- This has two types: ***forward*** and ***backward***.
  - For forward stepwise search, we start without any features. Then,
    - We train a 1-feature model using each of our candidate features and keep the version with the best performance.
    - We would continue adding features, one at a time, until our performance improvements stall.
  - Backward stepwise search is the same process, just reversed:
    - start with all features in our model and
    - then remove one at a time until performance starts to drop substantially.



Feature Extraction: **Principal Component Analysis (PCA)**

- PCA (unsupervised learning) is a dimensionality reduction that
  - identifies important relationships in our data,
  - transforms the existing data based on these relationships, and then
  - quantifies the importance of these relationships so we can keep the most important relationships.

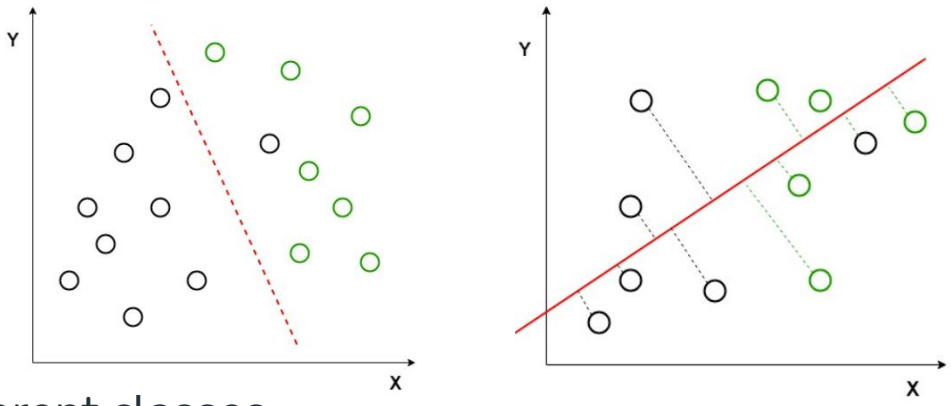
Objectives of PCA:

1. Reduces attribute space: It is basically a non-dependent procedure:
   - From a large number of variables to a smaller number of factors.
   - But there is no guarantee that the dimension is interpretable.
2. Identifying patterns: PCA can help identify patterns or relationships between variables.
3. Feature extraction: PCA can be used to extract features from a set of variables
   - that are more informative or relevant than the original variables.
4. Data compression: PCA can be used to compress large datasets by reducing the number of variables
   - while retaining as much information as possible.
5. Noise reduction: PCA can be used to reduce the noise in a dataset by
   - Identifying and removing the principal components that correspond to the noisy parts of the data.
6. Visualization: PCA can be used to visualize high-dimensional data in a lower-dimensional space,
   - making it easier to interpret and understand.

```
# Applying PCA function on training
# and testing set of X component
from sklearn.decomposition import PCA

pca = PCA(n_components = 2)

X_train = pca.fit_transform(X_train)
X_test = pca.transform(X_test)

explained_variance = pca.explained_variance_ratio_
```

Feature Extraction: **Linear Discriminant Analysis (LDA)**

- Linear Discriminant Analysis (LDA) is a supervised learning algorithm
  - used for classification tasks in machine learning.
- It is a technique used to find a linear combination of features that
  - best separates the classes in a dataset.



Example:

- Suppose we have two sets of data points belonging to two different classes
  - that we want to classify.
- As shown in the given 2D graph, when the data points are plotted on the 2D plane,
  - there's no straight line that can separate the two classes of the data points completely.
- Hence, in this case, LDA (Linear Discriminant Analysis) is used
  - which reduces the 2D graph into a 1D graph
  - in order to maximize the separability between the two classes.

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

# apply Linear Discriminant Analysis
lda = LinearDiscriminantAnalysis(n_components=2)
X_train = lda.fit_transform(X_train, y_train)
X_test = lda.transform(X_test)
```

Two criteria are used by LDA to create a new axis:

1. Maximize the distance between means of the two classes.
2. Minimize the variation within each class.

We have studied:

1. Dimensionality Reduction

Next:

4. Understanding **Effectiveness** of ML

Understanding AI/ML

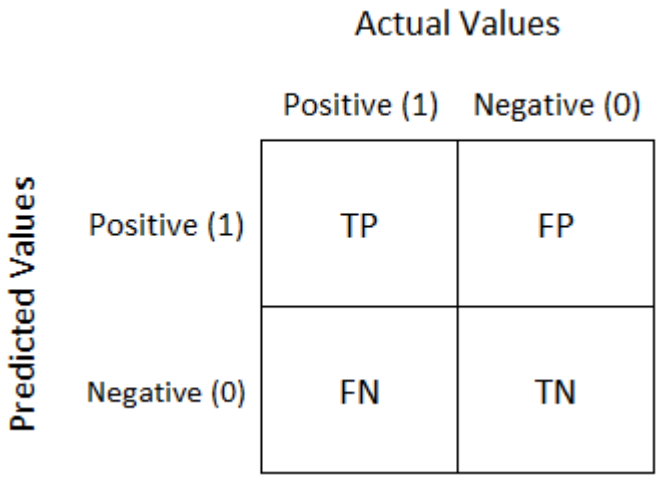
- Machine Learning is a branch of Artificial intelligence that focuses on the development of algorithms and statistical models
  - that can learn from and make predictions on data.
- Supervised algorithm is a type of machine learning where the algorithm learns from labeled data.
  - Labeled data means the dataset whose respective target value is already known.
- Supervised learning has two types:
  - **Regression**:
    - It predicts the continuous output variables based on the independent input variable.
      - like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.
  - **Classification**:
    - It predicts the class of the dataset based on the independent input variable.
      - Class is the categorical or discrete values. like the image of an animal is a cat or dog?
- Unsupervised Machine Learning is the technique where models are not provided with the labeled data and
- They have to find the patterns and structure in the data to know about the data.
  - **Clustering and Association** algorithms are a part of Unsupervised ML.

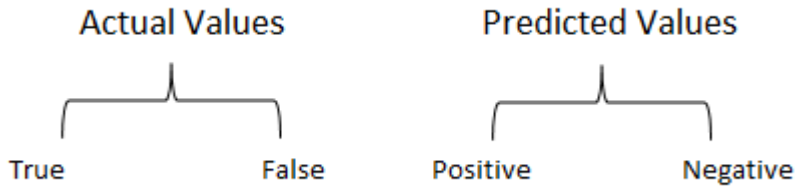
How to measure the effectiveness of AI/ML models?

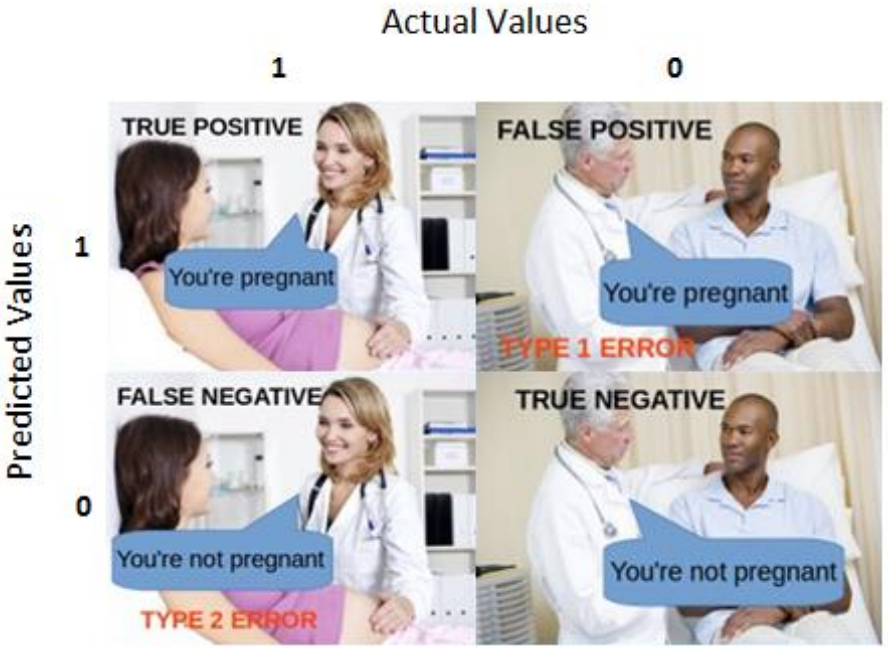
- When we get the data, (after data cleaning, pre-processing, and wrangling), the first step we do is
  - to feed it to an outstanding model and
  - get output in probabilities.
- But, How can we measure the effectiveness of our model?
  - Better the effectiveness, better the performance.
- Supervised:
  - **Regression** model's performance:
    - MAE, MSE, RMSE, R-squared, and Adjusted R-squared**.**
  - **Classification** model's performance:
    - Confusion Matrix and AUROC (Area Under the Receiver Operating Characteristics)
- Unsupervised:
  - **Clustering** model's performance:
    - Silhouette Coefficient
    - Calinski-Harabasz Index
    - Davies-Bouldin Index
  - **Association** model's performance: To measure the associations between thousands of data items, there are several metrics. Some of them are:
    - Support
    - Confidence
    - Lift

What is Confusion Matrix and why we need it?

- It is a performance measurement for machine learning classification problem
  - where output can be two or more classes.
- It is a table with 4 different combinations of predicted and actual values.
- It is extremely useful for measuring
  - Recall,
  - Precision,
  - Specificity,
  - Accuracy,
  - F1-Score and most importantly AUC-ROC curves.





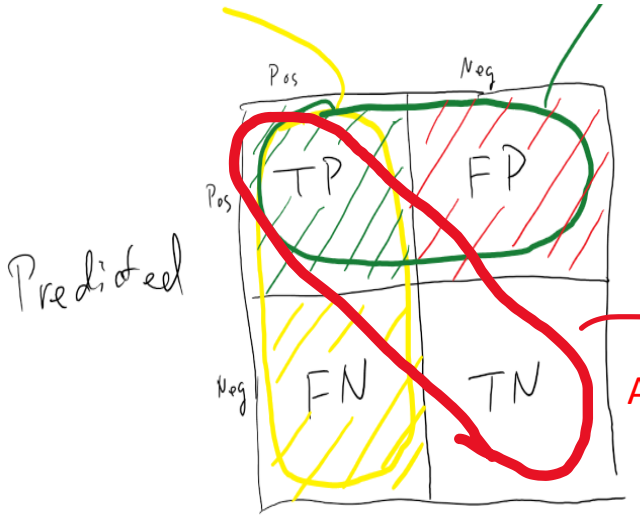


How to Calculate Confusion Matrix for a 2-class (Binary) classification problem?

y	y pred	output for threshold 0.6		Recall	Precision	Accuracy
0	0.5	0	TN	$= \frac{2}{2+2}$	$= \frac{2}{2+1}$	$= \frac{2+2}{7}$
1	0.9	1	TP			
0	0.7	1	FP			
1	0.7	1	TP			
1	0.3	0	FN			
0	0.4	0	TN			
1	0.5	0	FN			

$\text{Recall} = \frac{TP}{TP+FN}$ Actual

$\text{Precision} = \frac{TP}{TP+FP}$



$\text{Accuracy} = \frac{TP + TN}{Total}$

- **Accuracy**: From all the classes (positive and negative), how many of them we have predicted correctly.
  - ✓ Accuracy should be high as possible.
- **Recall**: From all the positive classes, how many we predicted correctly.
  - ✓ Recall should be high as possible.
- **Precision**: From all the classes we have predicted as positive, how many are actually positive.
  - ✓ Precision should be high as possible.

What is F-Measure or F1-Score?

- It is difficult to compare two models with low precision and high recall or vice versa.
- So to make them comparable, we use F-Score.
- F-score helps to measure Recall and Precision at the same time.
- It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

More:
Confusion Matrix for Multiple Classes

- Let us elaborate on the features of the multi-class confusion matrix with an example.
- Suppose we have the test set (consisting of 191 total samples) of a dataset with the distribution (Left Figure).
- The confusion matrix obtained by training a classifier and evaluating the trained model on this test set (Right Figure).
- Let that matrix be called "*M*," and each element in the matrix be denoted by "*M_ij*,"
  - where "*i*" is the row number (predicted class), and
  - "*j*" is the column number (expected class), e.g., *M_11*=52, *M_42*=1.

Class	№ of Samples
1	60
2	34
3	43
4	54

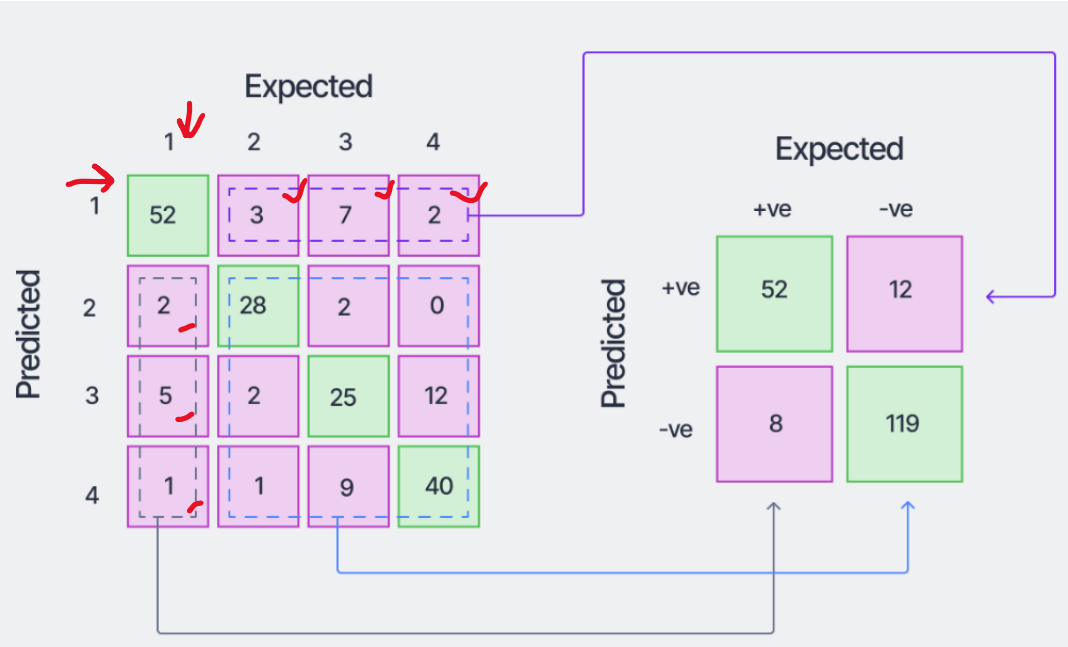
Confusion Matrix for Multiple Classes…

This confusion matrix gives a lot of information about the model's performance:

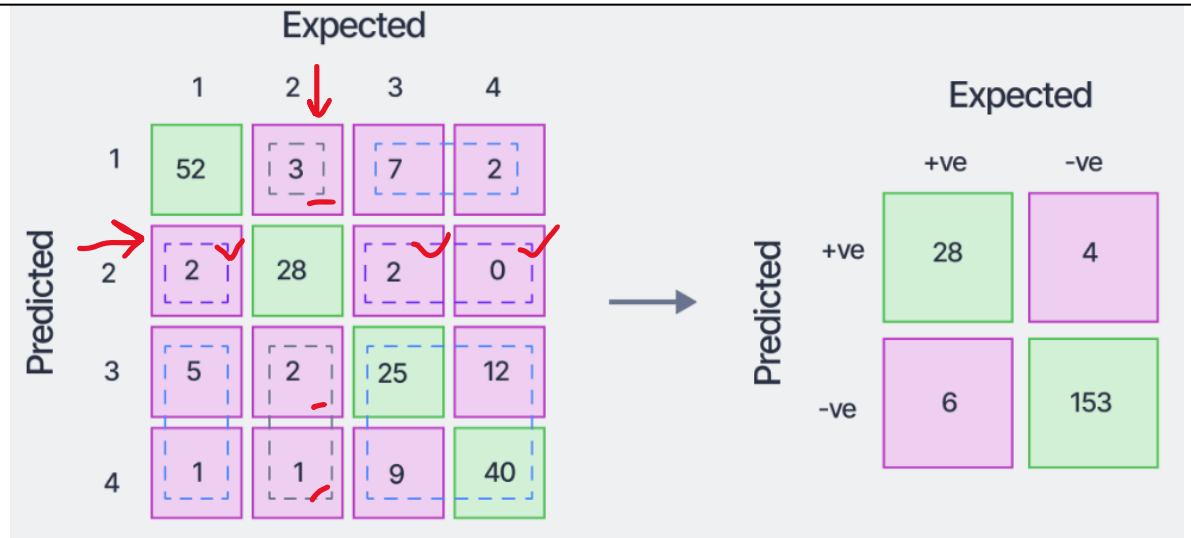
- As usual, the diagonal elements are the correctly predicted samples.
- A total of 145 samples were correctly predicted out of the total 191 samples.
  - ✓ Thus, the overall accuracy is 75.92%.
- *M_24*=0 implies that the model does not confuse samples originally belonging to class-4 with class-2, i.e.,
  - the classification boundary between classes 2 and 4 was learned well by the classifier.
- To improve the model's performance,
  - one should focus on the predictive results in class-3.
  - A total of 18 samples (adding the numbers in the red boxes of column 3) were misclassified by the classifier, which is the highest misclassification rate among all the classes.
  - Accuracy in prediction for class-3 is, thus, 58.14% only.

Converting the Confusion matrix

- The confusion matrix can be converted into a one-vs-all type matrix (binary-class confusion matrix)
  - for calculating class-wise metrics like accuracy, precision, recall, etc.
- Converting the matrix to a one-vs-all matrix for class-1 of the data looks like as shown below.
- Here, the positive class refers to class-1, and the negative class refers to “NOT class-1”.
- Now, the formulae for the binary-class confusion matrices can be used for calculating the class-wise metrics.



Similarly, for class-2, the converted one-vs-all confusion matrix will look like the following:



Converting the Confusion matrix…

- Calculate the class-wise
  - accuracy, precision, recall, and f1-scores

Class	Precision (%)	Recall (%)	F1-Score (%)
1	81.25	86.67	83.87
2	87.50	82.35	84.85
3	56.82	58.14	57.47
4	78.43	74.07	76.19

Class - 1

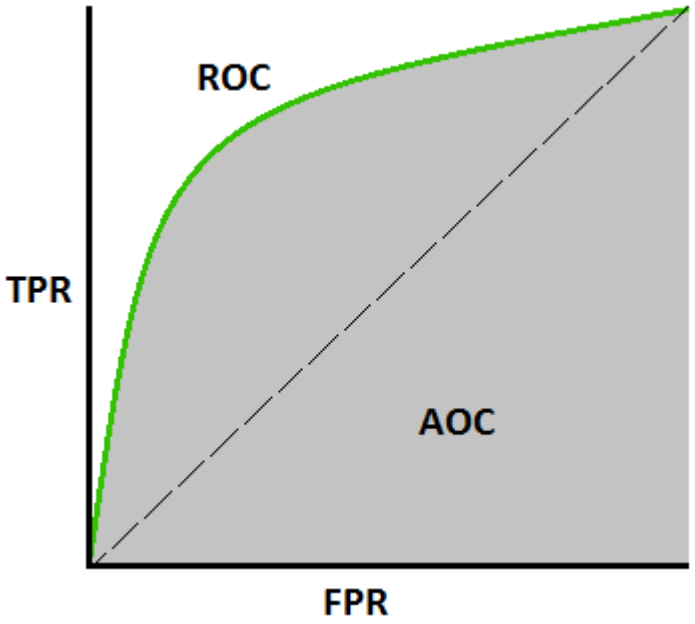
Class - 2

AUROC Curve

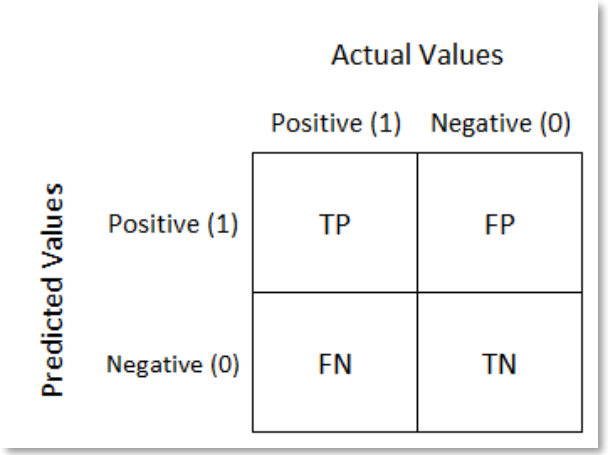
- When we need to check or visualize the performance of the multi-class classification problem,
  - We use the AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve.
- It is one of the most important evaluation metrics for checking any classification model's performance.
- It is also written as AUROC (**Area Under the Receiver Operating Characteristics**)

What is the AUC - ROC Curve?

- AUC - ROC curve is a performance measurement for the classification problems at various threshold settings.
- ROC is a probability curve
  - The ROC curve is plotted with TPR (True Positive Rate) against the FPR (False Positive Rate) where TPR is on the y-axis and FPR is on the x-axis.
- AUC represents the degree or measure of separability.
  - It tells how much the model is capable of distinguishing between classes.
  - Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.
  - The AUC value is essentially the probability that the classifier will rank a random positive test case higher than a random negative test instances.
- A random classifier corresponds to a diagonal line in the ROC plot.
  - We expect that half of the true positives and true negatives will be identify correctly.



Defining terms used in AUC and ROC Curve.



- TPR (True Positive Rate) / Recall/Coverage /Sensitivity
  - It is the percentage of persons with the disease who are correctly identified
  - proportion of truly diseased persons in a screened population who are identified as being diseased by the test.

$$\textbf{TRP/ Recall / Sensitivity} = \frac{\textbf{TP}}{TP+FN}$$

- Specificity-
  - It is the percentage of persons without the disease who are correctly excluded by the test.
  - It is the proportion of truly non-diseased persons who are so identified by the screening test.

$$\textbf{\textit{Specificity}} = \frac{\textbf{TN}}{TN+FP}$$

- FPR (False Positive Rate)

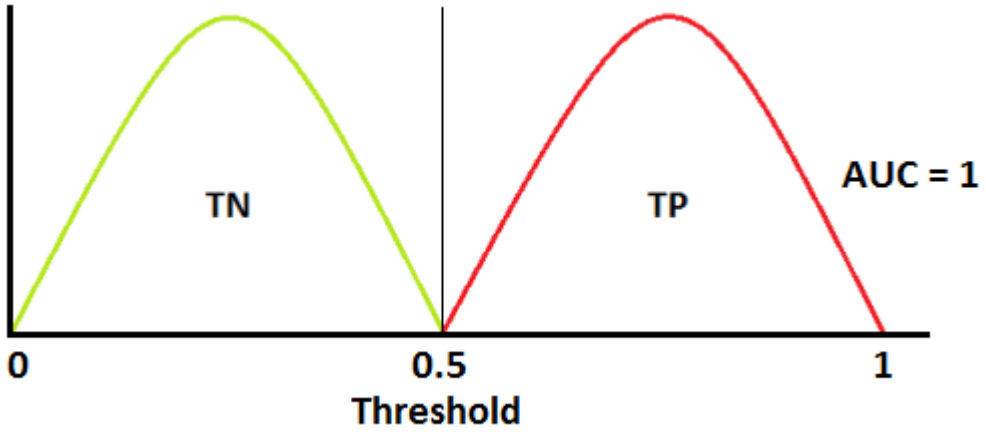
$$\textbf{FPR = 1 - Specificity} = \frac{\textbf{FP}}{\textbf{TN + FP}}$$

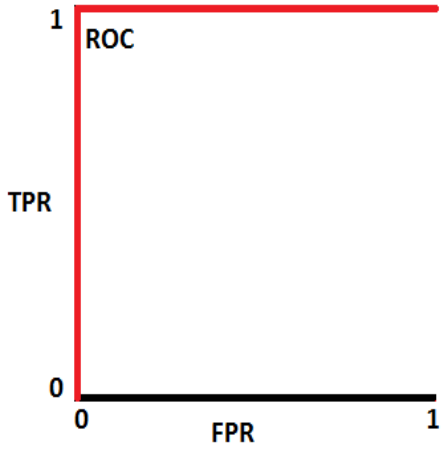
How to speculate about the performance of the model?

- An excellent model has AUC near to the 1 which means it has a good measure of separability.
- A poor model has an AUC near 0 which means it has the worst measure of separability (reciprocating the result).
  - It is predicting 0s as 1s and 1s as 0s.
- And when AUC is 0.5, it means the model has no class separation capacity whatsoever.
- “thresholds” in the context of ROC curves
  - Different thresholds represent the different possible classification boundaries of a model.

How to speculate about the performance of the model? …

- As we know, ROC is a curve of probability.
  - So let's plot the distributions of those probabilities:
- Note:
  - Red distribution curve is of the positive class (patients with disease) and
  - Green distribution curve is of the negative class(patients with no disease).

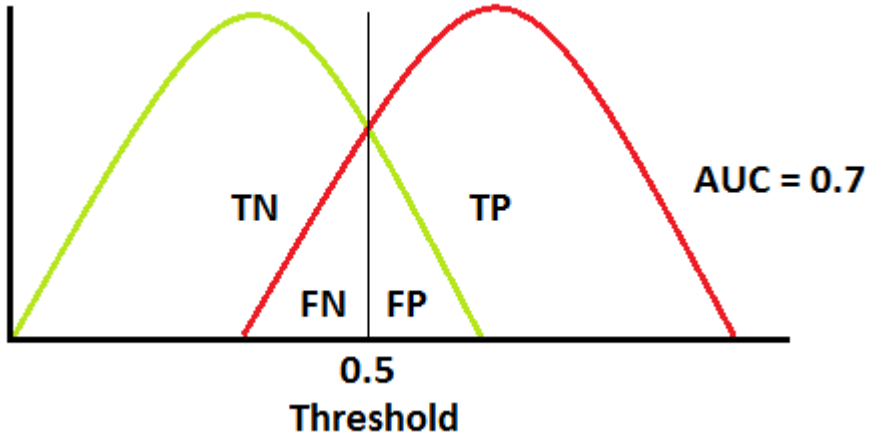


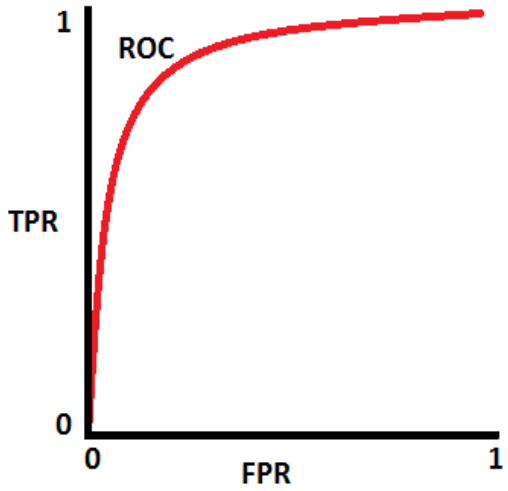


- This is an ideal situation.
  - It is perfectly able to distinguish between positive class and negative class.

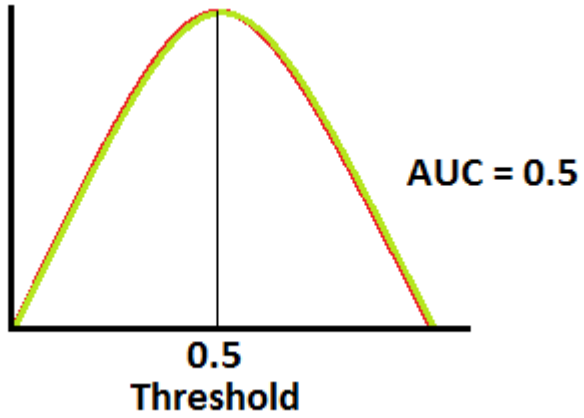
How to speculate about the performance of the model? …

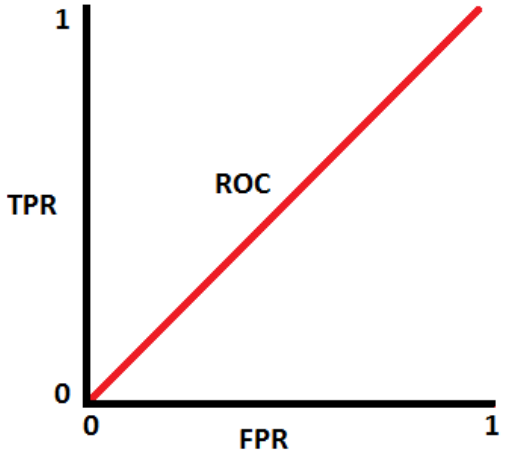
- When two distributions overlap,
  - we introduce type 1 and type 2 errors.
- Depending upon the threshold,
  - we can minimize or maximize them.
- When AUC is 0.7,
  - it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.





How to speculate about the performance of the model? …

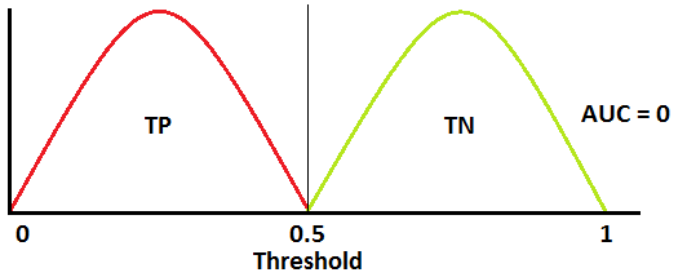


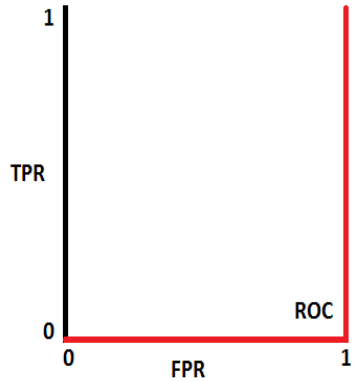


- This is the worst situation.
- When AUC is approximately 0.5,
  - the model has no discrimination capacity to distinguish between positive class and negative class.

How to speculate about the performance of the model? …

- When AUC is approximately 0,
  - the model is actually reciprocating the classes.
- It means the model is predicting a negative class as a positive class and vice versa.





Name	Formula	Explanation
True Positive Rate (TP rate)	TP / (TP + FP)	The closer to 1, the better. TP rate = 1 when FP = 0. (No false positives)
True Negative Rate (TN rate)	TN / (TN + FN)	The closer to 1, the better. TN rate = 1 when FN = 0. (No false negatives)
False Positive Rate (FP rate)	FP / (FP + TN)	The closer to 0, the better. FP rate = 0 when FP = 0. (No false positives)
False Negative Rate (FN rate)	FN / (FN + TP)	The closer to 0, the better. FN rate = 0 when FN = 0. (No false negatives)

The Formula of the ROC Curve

The relation between Sensitivity, Specificity, FPR, and Threshold

- Sensitivity and Specificity are inversely proportional to each other.
- So when we increase Sensitivity, Specificity decreases, and vice versa.

Sensitivity ↑, Specificity ↓ and Sensitivity ↓, Specificity ↑

- When we decrease the threshold, we get more positive values thus
  - it increases the sensitivity and decreasing the specificity.
- Similarly, when we increase the threshold, we get more negative values thus
  - We get higher specificity and lower sensitivity.
- As we know FPR is 1 – specificity
  - So when we increase TPR, FPR also increases and vice versa.

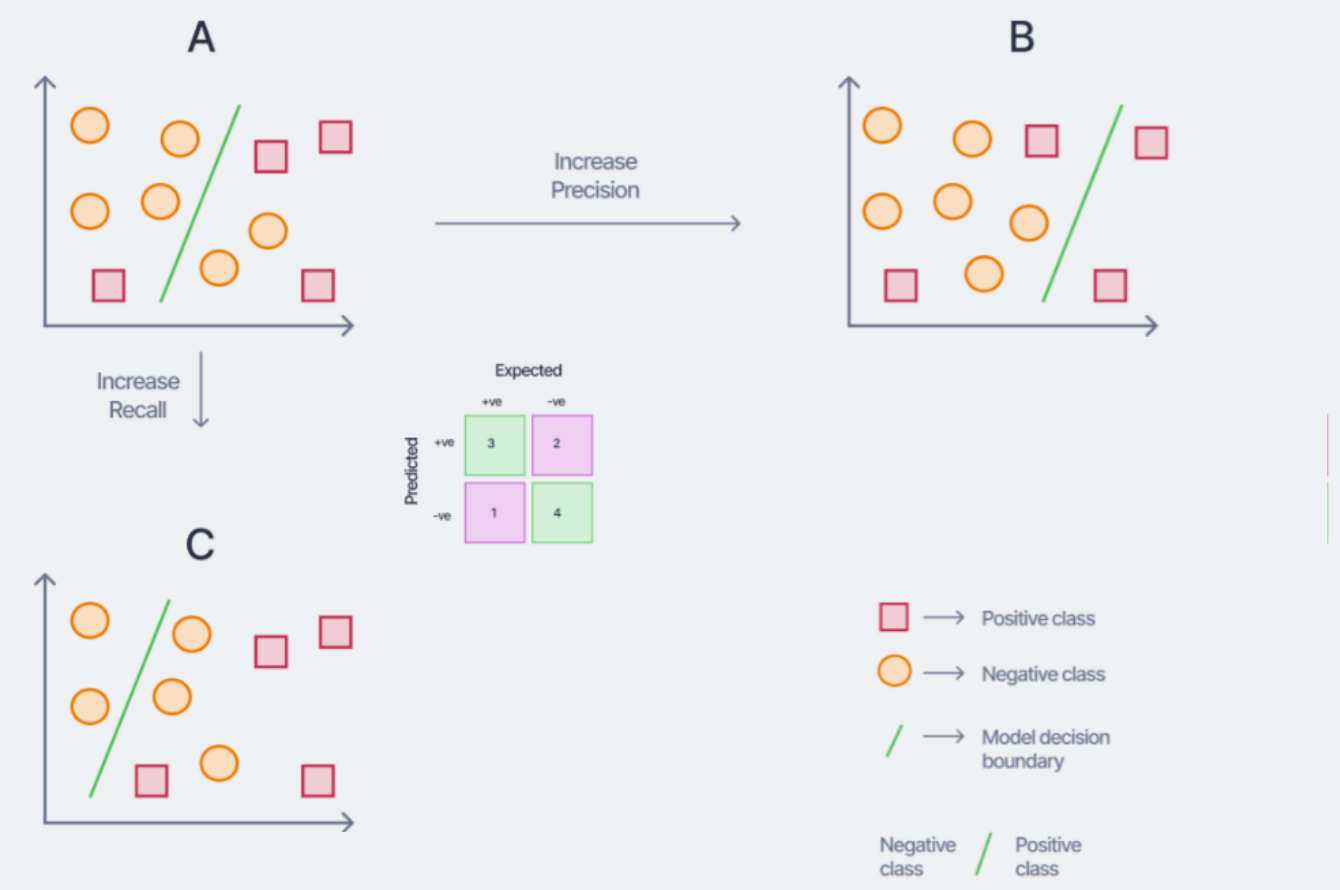
TPR ↑, FPR ↑ and TPR ↓, FPR ↓

The relation between Sensitivity, Specificity, FPR, and Threshold…

- Different thresholds represent the different possible classification boundaries of a model.
- Suppose we have a binary class dataset with 4 positive class samples and 6 negative class samples, and the model decision boundary is as shown by the green line in case (A).
- The RIGHT side of the decision boundary depicts the positive class, and the LEFT side depicts the negative class.

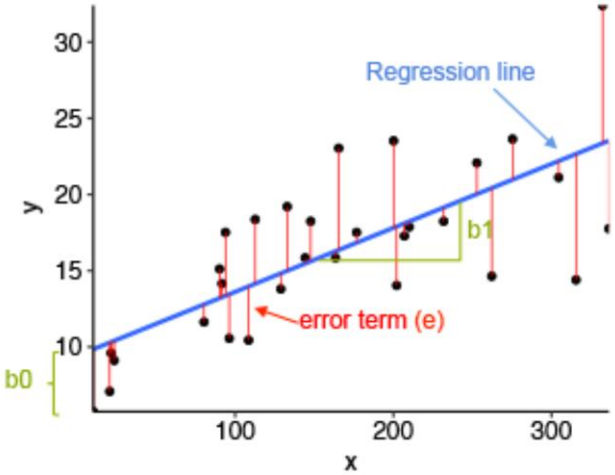
- Now, this decision boundary threshold can be changed to arrive
  - at case (B), where the Precision=? In (%) Recall=? In (%) or
  - to case (C) where the Recall=? In (%) (but Precision=? In (%)
- The corresponding confusion matrices are shown.
- The TPR and FPR values for these three scenarios with the different thresholds are thus as shown below.

	(A)	(B)	(C)
TPR	??? (in %)		
FPR			



Understanding Linear Regression

- This model finds the best fit linear line between the independent and dependent variable i.e
  - it finds the linear relationship between the dependent and independent variable.



- The vertical distance between the data point and the regression line is known as error or residual.
- Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

i.e $$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

- ✓ ***Residual plots expose a biased model than any other evaluation metric.***
- ✓ ***If own residual plots look normal, go ahead, and evaluate your model with various metrics.***

Evaluation metrics for a linear regression model

Mean Absolute Error (MAE)

- This is simply the average of the absolute difference between the target value and the value predicted by the model.
- Not preferred in cases where outliers are prominent.
- MAE does not penalize large errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Mean Squared Error (MSE)

- The most common metric for regression tasks is MSE.
- It is the average of the squared difference between the predicted and actual value.
- Since it is differentiable and has a convex shape, it is easier to optimize.
- MSE penalizes large errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Root Mean Squared Error (RMSE)

- This is the square root of the average of the squared difference of the predicted and actual value.
- RMSE penalizes large errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Solving a regression problem:

- Solving a regression problem involves obtaining a relation between
  - ✓ the response $Y$and the predictors
- i.e. determine a function $f(X_1, X_2, \ldots, X_p)$ such that $Y = f(X) + \in$,
  - ✓ based on the given set D of patterns.

where $\in$ is called a random error such that $E(\in) = 0$

	Advertising Budget on			Sales (Y)
	TV (X_1)	Radio (X_2)	News Paper (X_3)	
x_1	x_{11}	x_{12}	x_{13}	y_1
x_2	x_{21}	x_{22}	x_{23}	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	x_{i1}	x_{i2}	x_{i3}	y_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_N	x_{N1}	x_{N2}	x_{N3}	y_N

➢ In Multiple Linear Regression $f(X)$is assumed to be in the form:

(Model) $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ (1)

$$y \approx \hat{y} = f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

Goal: Determine β so as to satisfy the following N equations:

$$\left.\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} \\
&\vdots \\
y_N &= \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_j x_{Nj} + \cdots + \beta_p x_{Np}
\end{aligned}\right\} N \text{equations and } (p+1) \text{ unknowns}$$

Fact from Linear Algebra:

So, we seek a β for which the norm of r = y- **X**β (called the residual) is as small as possible.

We have $\hat{y}_i = f(x_i)$, where $\hat{y}_i$ denotes the computed value from the model for input x_i

- **The above system may be written as Xβ = y, where X is an N*(p+1) matrix and y is an N*1 vector.**
- **For an over determined system, these equations have a solution only if y is a linear combination of columns of X.**
- **For most choices of y, however, there is no vector** β for which **Xβ = y.**

So, we seek a β for which the norm of r = y- **X**β (called the residual) is as small as possible.

We have $\hat{y}_i = f(x_i)$, where $\hat{y}_i$ denotes the computed value from the model for input x_i

Residual (error) $= y_i - \hat{y}_i$

Residual Sum of Squares (RSS) is defined as

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

New Goal: Determine β so as to minimise $RSS(\beta)$ where,

$$RSS(\beta) = \sum_{i=1}^{N}\left[y_i - \left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)\right]^2$$

is a function of $f(\beta_0, \beta_1, \cdots, \beta_p)$

We need optimization!

Fact from Optimization:

- ✓ $RSS(\beta)$is minimized for some $\beta = \tilde{\beta}$
  - ✓ if and only if $\nabla f]_{\beta=\tilde{\beta}} = 0$ and Hessian of $f$at $\beta = \tilde{\beta}$is +ve definite.

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial \beta_0} \\ \frac{\partial f}{\partial \beta_1} \\ \frac{\partial f}{\partial \beta_2} \\ \vdots \\ \frac{\partial f}{\partial \beta_p} \end{pmatrix}, \text{ and Hessian of } f = \begin{bmatrix} \frac{\partial^2 f}{\partial \beta_0^2} & \frac{\partial^2 f}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_0 \partial \beta_2} & \cdots & \frac{\partial^2 f}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 f}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 f}{\partial \beta_1^2} & \frac{\partial^2 f}{\partial \beta_0 \partial \beta_2} & \cdots & \frac{\partial^2 f}{\partial \beta_1 \partial x_p} \\ \frac{\partial^2 f}{\partial \beta_2 \partial \beta_0} & \frac{\partial^2 f}{\partial \beta_2 \beta_1} & \frac{\partial^2 f}{\partial \beta_2^2} & \cdots & \frac{\partial^2 f}{\partial \beta_2 \partial x_p} \\ \vdots & & & & \\ \frac{\partial^2 f}{\partial \beta_n \partial \beta_0} & \frac{\partial^2 f}{\partial \beta_n \partial \beta_1} & \frac{\partial^2 f}{\partial \beta_n \partial \beta_2} & \cdots & \frac{\partial^2 f}{\partial \beta_p^2} \end{bmatrix}$$

- For a given function$f(x_1, x_2, \cdots, x_n)$ $f(x)$is minimized for some $x = \tilde{x}$ **if $\nabla f]_{x=\tilde{x}} = 0$** and **Hessian of $f$at $x = \tilde{x}$ is +ve definite**.

where,

$$\text{gradient of } f = \nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \text{ and hessian of } f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & & \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Matrix Calculus Formulae

Let $x \rightarrow$ vector (column vector),
$A \rightarrow$ Matrix

$\boldsymbol{y}$	$\boldsymbol{\frac{\partial y}{\partial x}}$
Ax	$\frac{\partial(Ax)}{\partial x} = A^T$
$x^T A$	$\frac{\partial(x^T A)}{\partial x} = A$
$x^T x$	$\frac{\partial(x^T x)}{\partial x} = 2x$
$x^T Ax$	$Ax + A^T x$ =2Ax if A is symmetric
$x^T \alpha = \alpha^T x$	α ($\alpha$is a column vector independent of x)

Let $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}_{N\times(p+1)}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}_{(p+1)\times 1}$

then $RSS(\beta) = (y - X\beta)^T (y - X\beta)$

hence, $\frac{\partial}{\partial\beta} RSS(\beta) = \frac{\partial}{\partial\beta}(z^T z)$ where, $z = y - X\beta$

$$= \frac{\partial z}{\partial\beta}\frac{\partial}{\partial z}(z^T z)$$
$$= \frac{\partial}{\partial\beta}(y - X\beta)2z$$
$$= \left[\frac{\partial}{\partial\beta}y - \frac{\partial}{\partial\beta}(X\beta)\right]2z$$
$$= \left[0 - \frac{\partial}{\partial\beta}(X\beta)\right]2z$$
$$= [-X^T]2z$$
$$= [-X^T]2(y - X\beta)$$
$$= -2X^T(y - X\beta)$$

since, $\frac{\partial}{\partial\beta} RSS(\beta) = 0$

$$\Rightarrow -2X^T(y - X\beta) = 0$$
$$\Rightarrow X^T(y - X\beta) = 0$$
$$\Rightarrow X^T y - X^T X\beta = 0$$
$$\Rightarrow X^T X\beta = X^T y$$
$$\Rightarrow (X^T X)^{-1}(X^T X)\beta = (X^T X)^{-1}(X^T y)$$
$$\Rightarrow \beta = (X^T X)^{-1}(X^T y) = \hat{\beta} \text{ (say)}$$

Note $(AB)^T = B^T A^T$

Hessian Matrix of $RSS(\beta)$

$$= \begin{bmatrix} \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_0} & \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_1} & \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_2} & \cdots & \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_P} \\ \frac{\partial}{\partial\beta_1}\frac{\partial RSS}{\partial\beta_0} & \frac{\partial}{\partial\beta_1}\frac{\partial RSS}{\partial\beta_1} & \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_2} & \cdots & \frac{\partial}{\partial\beta_1}\frac{\partial RSS}{\partial\beta_P} \\ \frac{\partial}{\partial\beta_2}\frac{\partial RSS}{\partial\beta_0} & \frac{\partial}{\partial\beta_2}\frac{\partial RSS}{\partial\beta_1} & \frac{\partial}{\partial\beta_0}\frac{\partial RSS}{\partial\beta_2} & \cdots & \frac{\partial}{\partial\beta_2}\frac{\partial RSS}{\partial\beta_P} \\ \vdots & & & & \\ \frac{\partial}{\partial\beta_N}\frac{\partial RSS}{\partial\beta_0} & \frac{\partial}{\partial\beta_N}\frac{\partial RSS}{\partial\beta_1} & \frac{\partial}{\partial\beta_N}\frac{\partial RSS}{\partial\beta_2} & \cdots & \frac{\partial}{\partial\beta_N}\frac{\partial RSS}{\partial\beta_P} \end{bmatrix}$$

$$= \frac{\partial}{\partial\beta}\left(\frac{\partial RSS(\beta)}{\partial\beta}\right)^T$$

$$= \frac{\partial}{\partial\beta}\left(\frac{\partial RSS(\beta)}{\partial\beta^T}\right)$$

$$= \frac{\partial^2}{\partial\beta}\frac{\partial RSS(\beta)}{\partial\beta^T}$$

$$= \frac{\partial}{\partial\beta}\left[\left(-2X^T(y - X\beta)\right)^T\right]$$

$$= (-2)\frac{\partial}{\partial\beta}\left[\left(X^T(y - X\beta)\right)^T\right]$$

$$= (-2)\frac{\partial}{\partial\beta}[(X^Ty - X^TX\beta)^T]$$

$$= (-2)\left[0 - \frac{\partial}{\partial\beta}(X^TX\beta)^T\right]$$

$$= 2\frac{\partial}{\partial\beta}[\beta^T(X^TX)^T]$$

$$= 2\frac{\partial}{\partial\beta}(\beta^TX^TX)$$

$$= 2X^TX$$

Note

1. For any matrix X we can prove X^TX is a square, and symmetric matrix.
2. A symmetric matrix S is positive definite **if and only if** there is a matrix A with independent columns such that $S=A^TA$ ($x^T(A^TA)x = x^TA^TAx = (Ax)^TAx > 0, x \neq 0$, since A has independent columns.)
3. X^TX is invertible if X has full column rank (Columns of X are independent)

For a new $x_0 \in \mathbb{R}^p$,

the response, $y_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_{0p} x_{0p}$

$$y_0 = \hat{\beta}^T \tilde{x}_0,$$

where,

$$\tilde{x}_0 = \begin{pmatrix} 1 \\ x_0 \end{pmatrix} = \begin{pmatrix} 1 \\ x_{01} \\ x_{02} \\ x_{03} \\ \vdots \\ x_{0p} \end{pmatrix}$$

Algorithm

Input: $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ and a new predictor vector $x_0 \in \mathbb{R}^p$

Output: y_0

1. Create the matrix X and the vector y.
2. Compute $\hat{\beta} = (X^T X)^{-1}(X^T y)$
3. Create the $\hat{x}_0$
4. Return $\hat{\beta}^T \hat{x}_0$

Important

$\hat{\beta} = (X^T X)^{-1}(X^T y)$ if the columns of X are linearly independent.

(Response Y is linearly related with the predictors)

Limitation of Multiple Linear Regression

1. If the data are not linearly related to the response then the prediction accuracy of the Multiple Linear Regression is low.
2. If some features are highly correlated then the matrix X may become computationally singular and hence $\hat{\beta}$ cannot be computed. If two features are highly correlated then remove one of them but not both.
3. When p is large linear regression fit does not predict well even when the data is not non-linear.

To overcome these limitations one of the approaches is shrinking the coefficient $(\beta's)$ of the model or setting some of the $\beta's$ equal to zero.

Logistic Regression

- It is a special case of linear regression where the target variable is categorical in nature.
- It predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is a dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

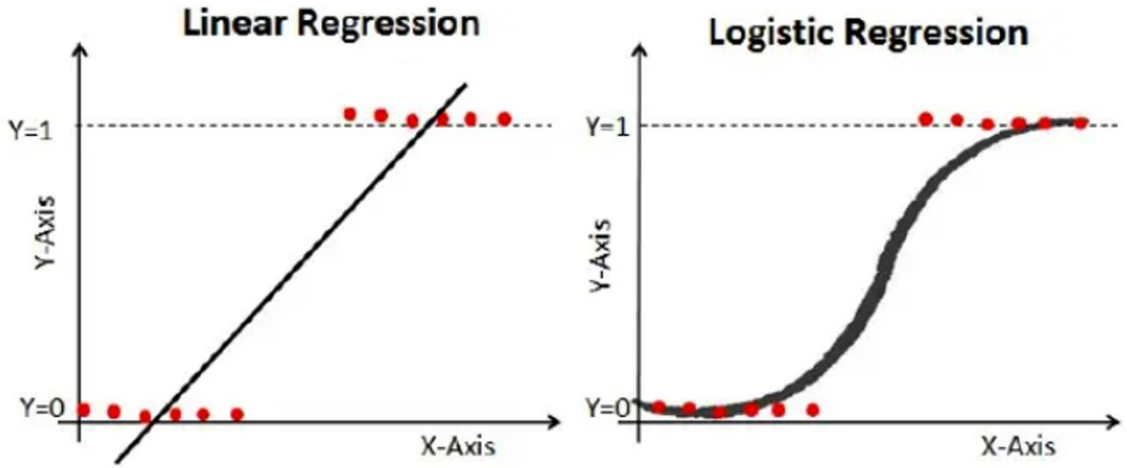
$$p = 1/1 + e^{-y}$$

Apply Sigmoid function on linear regression:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$

Difference between Linear and Logistic Regression

- Linear regression gives us a continuous output, but logistic regression provides a constant output.
  - An example of the continuous output is house price and stock price.
  - Example's of the discrete output is predicting whether a patient has cancer or not, predicting whether the customer will churn.
- Linear regression is estimated using Ordinary Least Square (OLS) while logistic regression is estimated using Maximum Likelihood Estimation (MLE) approach.



Sigmoid Function: $f(x) = \frac{1}{1+e^{-(x)}}$

Types of Logistic Regression

Binary Logistic Regression: The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.

Multinomial Logistic Regression: The target variable has three or more nominal categories such as predicting the type of Wine.

Ordinal Logistic Regression: the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.

K-NN (K - Nearest Neighbour)

- KNN is used to solve both Classification and Regression Problems.

KNN (K - Nearest Neighbour) for Classification Problem:

Given: $D = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \{1, 2, 3, \cdots, M\}$, K,

and an $x \in \mathbb{R}^p, x \neq x_i$

Goal: Determine the response $y \in \{1, 2, 3, \cdots, M\}$ associated with $x \in \mathbb{R}^p$.

KNN for Regression Problem

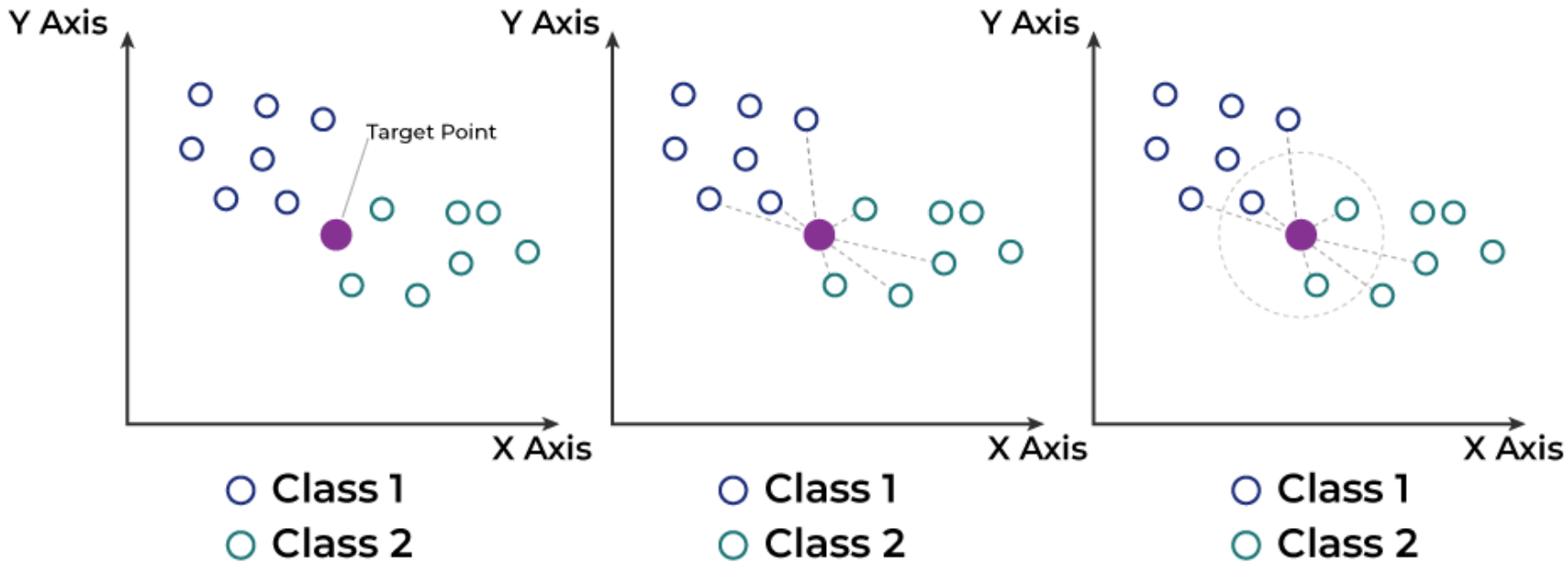
Given: $D = \{(x_i, y_i)\}_{i=1}^{N}$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, K

and an $x \in \mathbb{R}^p, x \neq x_i$

Goal: Determine the response $y \in \mathbb{R}$ associated with $x \in \mathbb{R}^p$.

Workings of KNN algorithm

- The K-Nearest Neighbors (KNN) algorithm operates on the principle of similarity, where
  - it predicts the **label** or **value** of a new data point by considering the labels or values of its K nearest neighbors in the training dataset.



Step-by-Step explanation of how KNN works:

Step 1: Selecting the optimal value of K

- K represents the number of nearest neighbors that needs to be considered while making prediction.

Step 2: Calculating distance

- To measure the similarity between target and training data points, Euclidean distance is used. Distance is calculated between each of the data points in the dataset and target point.

$$\text{distance}(x, X_i) = \sqrt{\sum_{j=1}^{d} (x_j - X_{i_j})^2]}$$

Step 3: Finding Nearest Neighbors

- The k data points with the smallest distances to the target point are the nearest neighbors.

Step 4: Voting for Classification or Taking Average for Regression

- In the classification problem, the class labels of K-nearest neighbors are determined by performing majority voting. The class with the most occurrences among the neighbors becomes the predicted class for the target data point.
- In the regression problem, the class label is calculated by taking average of the target values of K nearest neighbors. The calculated average value becomes the predicted output for the target data point.

Example

We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is?

Solution:

1. Determine parameter K = number of nearest neighbors

Suppose use K = 3

2. Calculate the distance between the query-instance and all the training samples

Coordinate of query instance is (3, 7), instead of calculating the distance we compute square distance which is faster to calculate (without square root)

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	**Square Distance to query instance (3, 7)**
7	7	$(7-3)^2+(7-7)^2=16$
7	4	$(7-3)^2+(4-7)^2=25$
3	4	$(3-3)^2+(4-7)^2=9$
1	4	$(1-3)^2+(4-7)^2=13$

Solution Cont…:

3. *Sort the distance and determine nearest neighbors based on the K-th minimum distance*

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	**Rank minimum distance**	**Is it included in 3-Nearest neighbors?**
7	7	$(7-3)^2+(7-7)^2=16$	**3**	**Yes**
7	4	$(7-3)^2+(4-7)^2=25$	**4**	**No**
3	4	$(3-3)^2+(4-7)^2=9$	**1**	**Yes**
1	4	$(1-3)^2+(4-7)^2=13$	**2**	**Yes**

Solution Cont…:

4. Gather the category **Y** *of the nearest neighbors.* Notice in the second row last column that the category of nearest neighbor (Y) is not included because the rank of this data is more than 3 (=K).

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Square Distance to query instance (3, 7)	Rank minimum distance	Is it included in 3-Nearest neighbors?	**Y = Category of nearest Neighbor**
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	**Bad**
7	4	$(7-3)^2+(4-7)^2=25$	4	No	-
3	4	$(3-3)^2+(4-7)^2=9$	1	Yes	**Good**
1	4	$(1-3)^2+(4-7)^2=13$	2	Yes	**Good**

5. *Use simple majority of the category of nearest neighbors as the prediction value of the query instance*

We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7 is included in **Good** category.

Example 2: The table below provides a training data set containing six observations, three predictors X1, X2, X3 and one qualitative response variable Y.

Observation	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

- Predict the value of Y when X1 = 1, X2=5, X3 = 0 using KNN with k=3.

Sol:

Observation	X1	X2	X3	Square distance to query instance (X1 = 1, X2=5, X3 = 0)	Rank Minimum Distance	Is it included in 3-NN	Y
1	0	3	0	$(0-1)^2+(3-5)^2+(0-0)^2 = 1+4+0=5$	1	Y	Red
2	2	0	0	$(2-1)^2+(0-5)^2+(0-0)^2 = 1+25+0=26$	5	N	-
3	0	1	3	$(0-1)^2+(1-5)^2+(3-0)^2 = 1+16+9=26$	4	N	-
4	0	1	2	$(0-1)^2+(1-5)^2+(2-0)^2 = 1+16+4=21$	3	Y	Green
5	-1	0	1	$(-1-1)^2+(0-5)^2+(1-0)^2 = 4+25+1=30$	6	N	-
6	1	1	1	$(1-1)^2+(1-5)^2+(1-0)^2 = 0+16+1=17$	2	Y	Red

- The predict value of Y when X1 = 1, X2=5, X3 = 0 using KNN with k=3 is **Red**.

KNN (K - Nearest Neighbour) for Classification Problem:

Algorithm

Input: Data Set $D, x \in \mathbb{R}^P$,K|

Output: Class Label of x

1. for i = 1 to N do
   compute the distance of x_i from x:
   $$d_i = \sqrt{\sum_{j=1}^{P}(X_j - x_{ij})^2}$$
2. sort $\{d_1, d_2, \cdots, d_N\}$ in non-decreasing order.
   choose the K-smallest distances in the sorted list and
   denote the set of points x_i which corresponds to K-smallest
   distances as N_x
3. for j = 1 to M do
4. count the number n_j of points of class j in N_x.
5. compute the maximum of the n_j's.
   Let this maximum be m.
   Determine the class label l of the class which has m points in N_x.
   or
   compute $l = \underset{1 \le j \le M}{argmax} n_j$
6. return l

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{pmatrix} \qquad x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$\|X - x_j\| = \sqrt{(X_1 - x_{i1})^2 + (X_2 - x_{i2})^2 + \cdots + (X_P - x_{iP})^2}$$

KNN (K - Nearest Neighbour) for Regression Problem:

Algorithm

1. for i = 1 to N do
   compute the distance of x_i from x.
   $$d_i = \sqrt{\sum_{j=1}^{P} (X_j - x_{ij})^2}$$
2. sort $\{d_1, d_2, \cdots, d_N\}$ in non-decreasing order.
   choose the K-smallest distances in the sorted list and denote the points x_i's which corresponds to smallest distance as N_x that is $|N_x| = k$.
3. $y = \frac{1}{k} \sum_{x_j \in N_x} y_j$ where y_j is the response of x_j.
4. return y

Advantages of KNN

- **It does not make any assumption concerning the distribution of data.**
- **So, the method is applicable to wide class of data.**

Disadvantages of KNN

1. It provides limited insight into relationship between the predictors and the classes.
2. Here one needs the entire dataset whenever one wants to classify a new point x in R^p.
   So, if N or p is large the computation is prohibitive.
3. The prediction is slow, because it requires comparing **distances to every point.**

Remarks:

1. The accuracy of the method depends upon the choice of K. So, one has to tune the value of K suitably.

   Theorem: The expected error of k-NN rule converges to 1+sqrt(8/k) times the error of the Bayes classifier.

2. When the number of features p is large, there tends to be a deterioration in the performance of KNN.

5. How to choose the Right Statistical Test

- **Statistical tests** are used in hypothesis testing, which can be used to:
  - determine whether a predictor variable has a statistically significant relationship with an outcome variable.
  - estimate the difference between two or more groups.
- Statistical tests assume a null hypothesis of no relationship or no difference between groups.
  - Then, they determine whether the observed data fall outside of the range of values predicted by the null hypothesis.
- If we already know what types of variables we are dealing with, then
  - we can choose the right statistical test for our data. (See the flowchart)

Finally, what does a statistical test do?

- Statistical tests work by calculating a **test statistic**
  - A number that describes how much the relationship between variables in our test differs from the null hypothesis of no relationship.
- It then calculates a ***p* value** (probability value)
  - If the null hypothesis of no relationship were true then
    - The *p*-value estimates how likely it is that we would see the difference described by the test statistic.
    - Therefore, the p-value gives us an estimate of how "strange" our sample is.
- If the value of the test statistic is more extreme than the statistic calculated from the null hypothesis, then
  - We can infer a **statistically significant relationship** between the predictor and outcome variables.
- If the value of the test statistic is less extreme than the one calculated from the null hypothesis, then
  - We can infer **no statistically significant relationship** between the predictor and outcome variables.

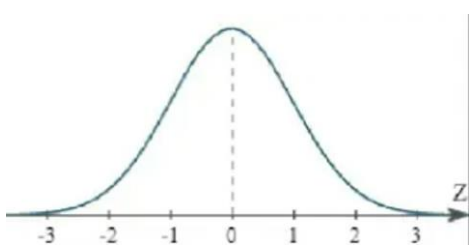
When to perform a statistical test

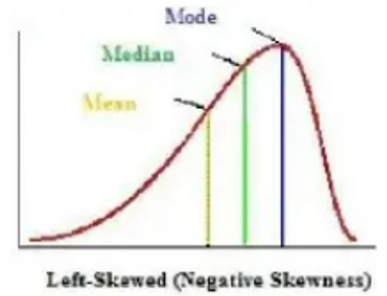
- We can perform statistical tests on data that have been collected in a statistically valid manner
  - – either through an experiment, or
  - – through observations made using probability sampling methods.
- For a statistical test to be valid,
  - our sample size needs to be large enough to approximate the true distribution of the population being studied.
- To determine which statistical test to use, we need to know:

  (a) whether our data meets certain assumptions. (in the next slide)

  (b) the types of variables that you're dealing with.

(a) Statistical Test Assumptions





- Statistical tests make some common assumptions about the data we are testing:
  1. **Independence of observations** (a.k.a. no autocorrelation): The observations/variables we include in our test are not related
     - ✓ (for example, multiple measurements of a single test subject are not independent, while measurements of multiple different test subjects are independent).
  2. **Homogeneity of variance**: the variance within each group being compared is similar among all groups.
     - ✓ If one group has much more variation than others, it will limit the test's effectiveness.
  3. **Normality of data**: the data follows a normal distribution (a.k.a. a bell curve).
     - ✓ This assumption applies only to quantitative data.
- If our data do not meet the assumption of independence of observations,
  - we may be able to use an experiment/test that accounts for structure in our data (repeated-measures tests or tests that include blocking variables).
- If our data do not meet the assumptions of normality or homogeneity of variance, we may be able to perform a **nonparametric statistical test**,
  - which allows us to make comparisons without any assumptions about the data distribution.

(b) Types of variables

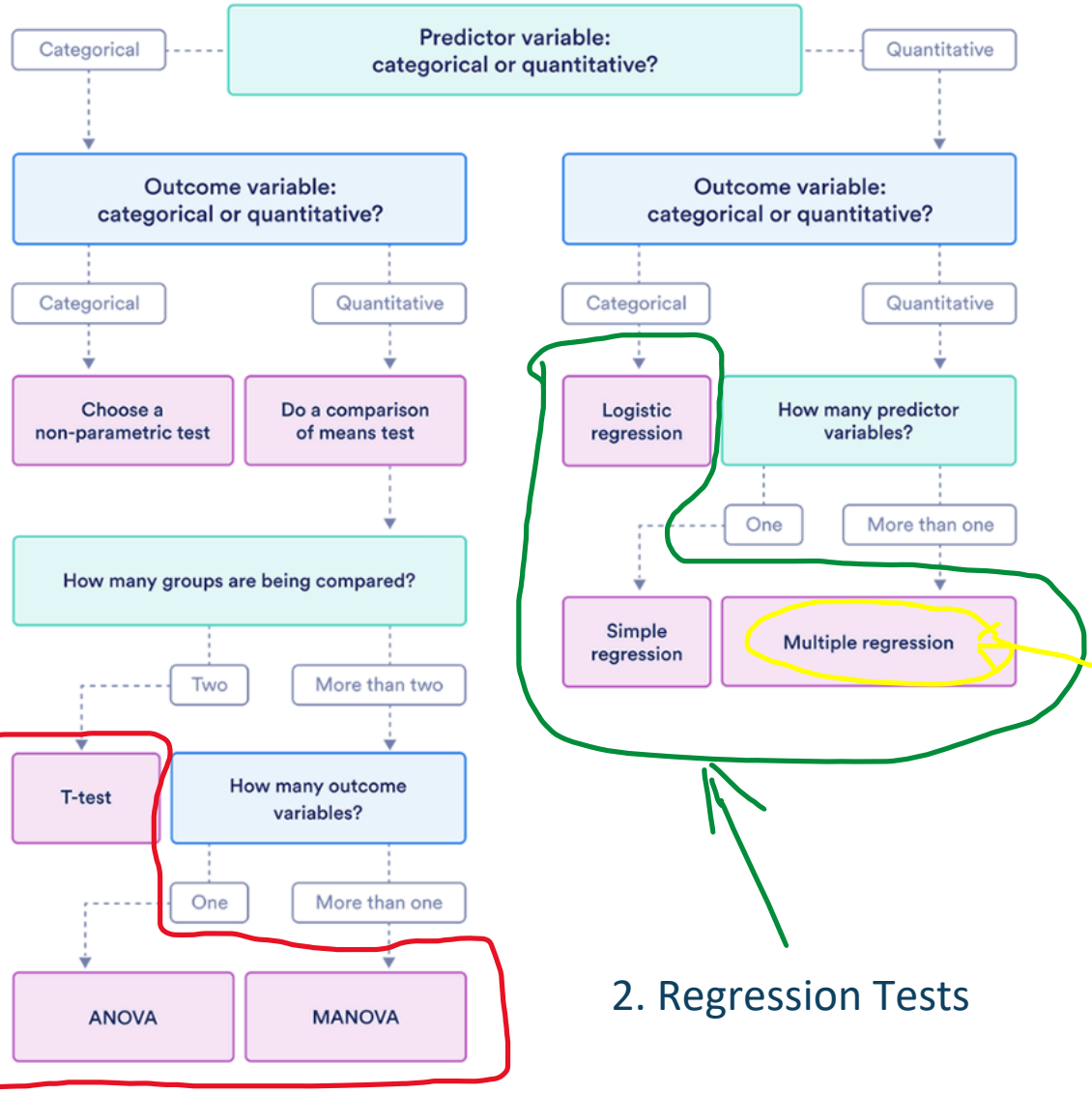
The types of variables have usually determine what type of statistical test we can use.

- **Quantitative (or Numerical) variables** represent amounts of things (e.g. the number of trees in a forest).
  - ✓Types of quantitative variables include:
    - **Continuous** (aka ratio variables): represent measures and can usually be divided into units smaller than one (e.g. 0.75 grams).
    - **Discrete** (aka integer variables): represent counts and usually can't be divided into units smaller than one (e.g. 1 tree).
- **Categorical variables** represent groupings of things (e.g. the different tree species in a forest).
  - ✓Types of categorical variables include:
    - **Ordinal**: represent data with an order (e.g. rankings).
    - **Nominal**: represent group names (e.g. brands or species names).
    - **Binary**: represent data with a yes/no or 1/0 outcome (e.g. win or lose).
- Choose the test that fits the types of predictor and outcome variables we have collected (if we are doing an experiment, these are the independent and dependent variables).
  - The **independent variable** is the cause. Its value is **independent** of other variables in your study.
  - The **dependent variable** is the effect. Its value **depends** on changes in the independent variable.

Example: Independent and dependent variables

- We design a study to test whether changes in room temperature have an effect on math test scores.
  - Our **independent variable** is the temperature of the room.
    - We vary the room temperature by making it cooler for half the participants, and warmer for the other half.
  - Our **dependent variable** is math test scores.
    - We measure the math skills of all participants using a standardized test and check whether they differ based on room temperature.

Flowchart: **choosing a statistical test**



1. Comparison Tests

2. Regression Tests

3. Correlation tests:
whether two variables we want to use in a multiple regression test are autocorrelated.
(Pearson's r test)

Choosing a Parametric Test:
Regression, Comparison, or Correlation

- Parametric tests usually have stricter requirements than nonparametric tests,
  - and are able to make stronger inferences from the data.
- They can only be conducted with data that adheres to the common assumptions of statistical tests.
- The most common types of parametric test include

  (a) Regression Tests,

  (b) Comparison Tests, and

  (c) Correlation Tests.

(a) Regression Tests (a Parametric Test)

- Regression tests look for **cause-and-effect relationships**.
- They can be used to estimate the effect of one or more continuous variables on another variable.

	Predictor variable	Outcome variable	Research question example
Simple linear regression	• Continuous • 1 predictor	• Continuous • 1 outcome	What is the effect of income on longevity?
Multiple linear regression	• Continuous • 2 or more predictors	• Continuous • 1 outcome	What is the effect of income and minutes of exercise per day on longevity?
Logistic regression	• Continuous	• Binary	What is the effect of drug dosage on the survival of a test subject?

(b) Comparison Tests (a Parametric Test)

- Comparison tests look for **differences among group means**.
  - They can be used to test the effect of a categorical variable on the mean value of some other characteristic.
- T-tests are used when comparing the means of precisely two groups (e.g., the average heights of men and women).
- ANOVA and MANOVA tests are used when comparing the means of more than two groups (e.g., the average heights of children, teenagers, and adults).

Example of Comparison tests…

	Predictor variable	Outcome variable	Research question example
Paired t-test	• Categorical • 1 predictor	• Quantitative • groups come from the same population	What is the effect of two different test prep programs on the average exam scores for students from the same class?
Independent t-test	• Categorical • 1 predictor	• Quantitative • groups come from different populations	What is the difference in average exam scores for students from two different schools?
ANOVA	• Categorical • 1 or more predictor	• Quantitative • 1 outcome	What is the difference in average pain levels among post-surgical patients given three different painkillers?
MANOVA	• Categorical • 1 or more predictor	• Quantitative • 2 or more outcome	What is the effect of flower species on petal length, petal width, and stem length?

Correlation tests (a Parametric Test)

- Correlation tests **check whether variables are related**
  - without hypothesizing a cause-and-effect relationship.
- These can be used to test whether two variables we want to use in a multiple regression test are autocorrelated.

✓ Suppose there is a strong correlation between two variables or metrics, and one of them is being observed acting in a particular way.
  ✓ In that case, we can conclude that the other one is also being affected similarly.
  ✓ This helps group related metrics together to reduce the need for individual data processing.

	Variables	Research question example
Pearson's *r*	• 2 continuous variables	How are latitude and temperature related?

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

Choosing a nonparametric test

- Non-parametric tests don’t make as many assumptions about the data, and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren’t as strong as with parametric tests.

12. Choosing a nonparametric test…

	Predictor variable	Outcome variable	Use in place of…
Spearman's *r*	• Quantitative	• Quantitative	Pearson's *r*
Chi square test of independence	• Categorical	• Categorical	Pearson's *r*
Sign test	• Categorical	• Quantitative	One-sample *t*-test
Kruskal–Wallis *H*	• Categorical • 3 or more groups	• Quantitative	ANOVA
ANOSIM	• Categorical • 3 or more groups	• Quantitative • 2 or more outcome variables	MANOVA
Wilcoxon Rank-Sum test	• Categorical • 2 groups	• Quantitative • groups come from different populations	Independent t-test
Wilcoxon Signed-rank test	• Categorical • 2 groups	• Quantitative • groups come from the same population	Paired t-test