

# Data Science Project Pipeline.

## EDA.

- EDA stands for Exploratory Data Analysis. It refers to the process of examining and analyzing data to summarize their main characteristics and gain insight into the data.
- It typically involves the use of statistical graphics and data visualization techniques to help identify patterns, relationships, and anomalies in the data.
- The main objective of EDA is to get better understanding of the data and to uncover any hidden insights or relationships that may not be immediately apparent, which can help in formulating hypotheses and developing predictive models as well as decision-making based on the data.

## Description of data :-

- It refers to summary of main characteristic of dataset. This typically include information such as mean, median, mode, std deviation range, min and max values and relevant statistics.
- The purpose of it is to provide overview of the dataset and to help identify patterns or trends that may be present which can help better understand data and to make informed decisions about how to analyze it further.
- There are number of techniques can be used to describe data. like tabular summarization, graphical display and statistical measures.

- Descriptive statistics such as measure of central tendency and variability are commonly used to summarize numerical data. while frequency distribution and histogram are oftenly used to summarize categorical data.
- it imp to note that it doesn't provide insight or explanation for the data it is simply summary of dataset and serves as starting point for further analysis and interpretation.

## what is Correlation , Co-variance and VIF ?

- correlation , covariance and vif (Variance Inflation Factor) are all statistical measures that are used to access the relationship between variable in dataset
- Correlation :-  
it is statistical measures that indicates the degree to which two variable are related to each other it ranges from -1 to 1 , where value of -1 indicates perfect negative correlation ( i.e as one variable increases other decreases ) ; a value zero indicates no correlation a value of 1 indicates perfect positive correlation . ( as one variable increases other also increases ) . it can be calculated by using various methods such as pearson Correlation coefficient , spearman's rank Correlation coefficient .

## Covariance :-

it is measure of joint variability between two variables a positive covariance indicates both vary in same direction whereas negative covariance tend to vary in opposite direction ; unlike correlation . covariance is not normalized so it can be difficult to interpret the magnitude of value.

## VIF (Variance inflation factor) :-

it is a measure used to detect multicollinearity in regression analysis ; multicollinearity occurs when two or more independent variables in regression model are highly correlated with each other making it difficult to interpret the effect of each variable separately . VIF measures the degree to which the variance of estimated coefficient of independent variable is increased due to multicollinearity with other independent variable in the model . A VIF value of 1 indicates value of no correlation while value greater than one suggest increasing levels of multicollinearity .

In summary Correlation and covariance are measure of relationship b/w two variables while VIF measure is used to detect multicollinearity in regression analysis .

$$VIF = \frac{1}{1 - R^2} \quad 1 \leq VIF \leq \infty$$

if  $R^2$  is high VIF will be also high .

## Univariate & Bivariate Analysis.

Univariate analysis and bivariate analysis are two types of statistical analysis used to study a set of data.

### Univariate Analysis:-

- it is analysis of single variable in isolation. In other words, it involves examining the distribution and properties of single variable w/o considering relationship between that variable and other variable in the same dataset.
- Examples of univariate analysis techniques include measure of central tendency (such as mean, median, mode), measure of dispersion (range, variance & std) and graphical representation like histogram and box plot.

#### for categorical data.

- countplot : <syntax>  
`sns.countplot(df['<name of column>'])`
- Pie chart  
`fig, ax = plt.subplots()  
ax.pie(sizes, labels = labels, autopct = '%.1f %%')`

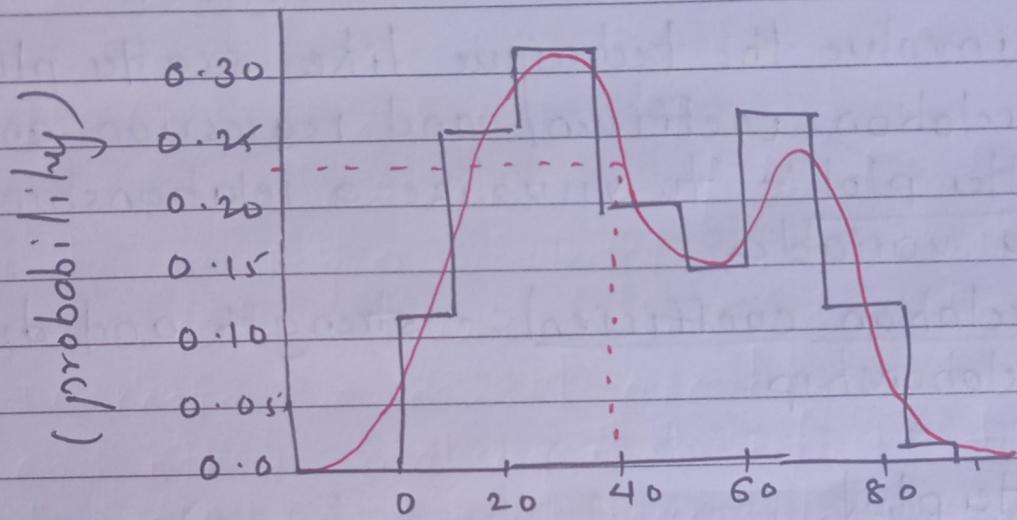
#### for Numerical data

- a. Histogram : it helps to understand the distribution of data. It partitions data into bins or range  
 for ex : if Age column  
`0-10, 10-20, 20-30, 30-40, 40-50, 50-60.`

`plt.hist(cdf[['<age column>']]).`

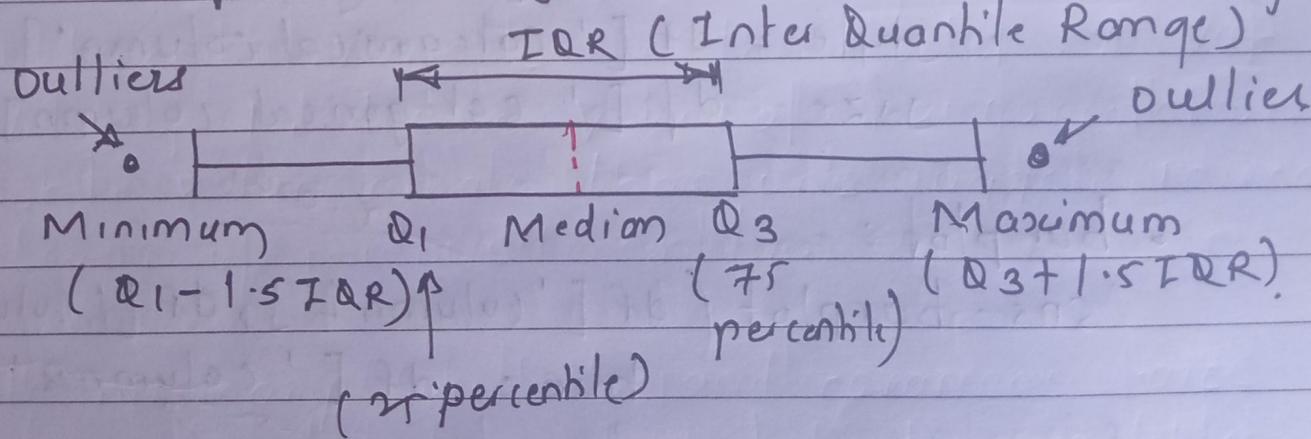
### b. Distplot.

it is almost similar to histogram only difference is curved line which called as kde (kernel density estimation) or pdf (probability density function)



probability of  $x=40$  is between 0.20 to 0.28

c. Box plot :- it gives five number summary



`sns.boxplot(cdf['<name of column>'])`

## Bivariate Analysis :-

- it is analysis of relationship between two variables. for examining the relationship between two variables to determine if there is any correlation or association between them.
  - it often used to test hypothesis and to identify the strength direction of a relationship between two variables.
  - it involve the technique like scatter plot, Correlation coefficient and regression analysis.
- Scatter plot :- to visualize a relationship between two variables.

Correlation coefficient :- strength and direction of relationship.

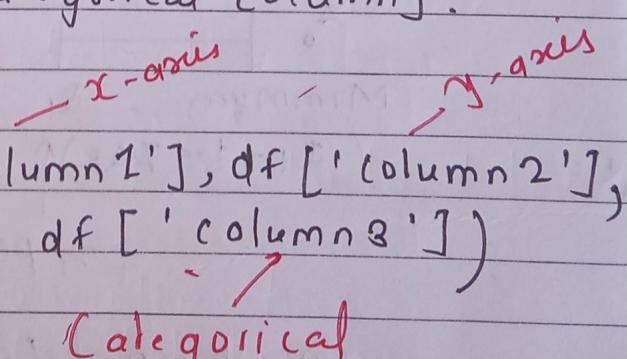
### Scatter plot

`sns.scatterplot(df['column1'], df['column2'])`

$\text{hue} = \text{df['categorical column']}$ ,  
 $\text{style} = \text{df['categorical column']}$ ,  
 $\text{size} = \text{df['categorical column']}$ .

### bar plot :-

`sns.barplot(df['column1'], df['column2'], hue = df['column3'])`

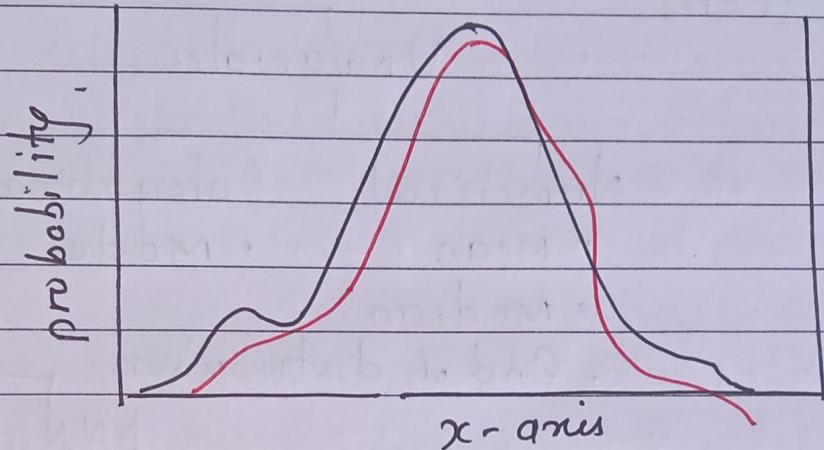


### box plot :-

`sns.boxplot(df['column1'], df['column2'], hue = df['column3'])`

#### 4. distplot (Numerical - categorical)

`sns.distplot(df[df['column1'] == 0]['column2'], hist=False)`  
`sns.distplot(df[df['column1'] == 1]['column2'], hist=False)`



#### 5. Heatmap (categorical - Categorical)

6. pairplot : it automatically identify all the numerical column in the dataset and plot their relation. (scatter)

`sns.pairplot('dataset') hue = 'categorical column'`

#### 7. Lineplot -( Numerical - Numerical)

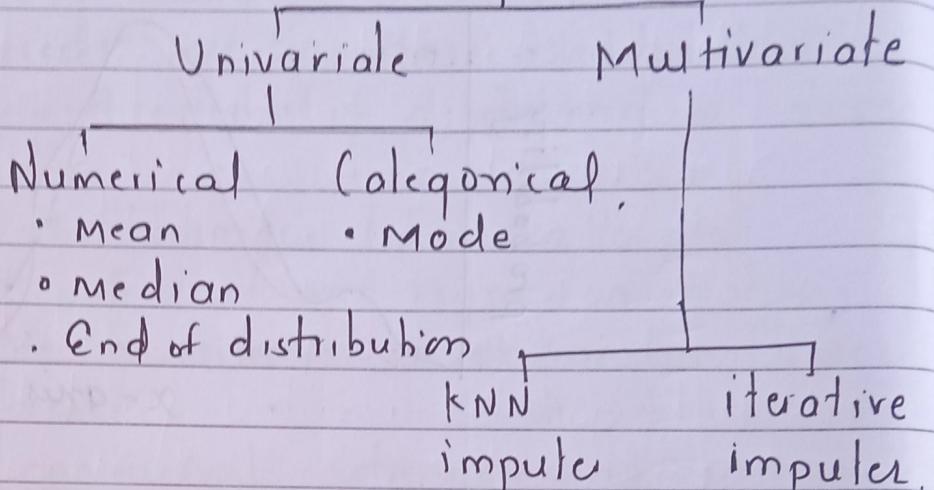
`sns.lineplot(df['column1'], df['Column2'])`

# Missing Value treatment

## Handling Missing Values technique

① Dropping  
(CCA)

② Impulation



### ① Dropping or Removing.

- Complete Case Analysis (CCA) also called as list wise deletion it consist of discarding (rows) Observation where value in any of the (columns) Variable are missing
  - it means literally allowing only those observation for which there is information in all of variable in the dataset

Assumption :- data should be randomly distributed, missing

Advantages :- Easy to implement

2) preserve variable distribution

Disadvantage :- if missing data is abundant it can exclude large fraction of original dataset.

2. Excluded info could be imp for analysis
3. when using the model in production  
model may not know how to handle missing data!

when to use?

it applicable when any column having missing values less than 5%

Note:- after deletion of those rows make sure your distribution of data is still same.

code: dropna(.)

### Handling Missing Numerical Data.

• Univariate  $\rightarrow$  Numerical

1st Mean/Median Imputation.

mean:- it is used when data is normally distributed.

Median:- when data contain outlier due to which distribution is skewed to left or right

Benefits: simple. disadvantages:

- shape of distribution changes.
- outliers emerges.
- Effect on covariance/Correlation.

when to use:- when data completely

- ① missing at Random.
- ② missing value < 5%

Arbitrary value imputation.

Filling missing value with any arbitrary number

Numerical  $\rightarrow$  99, -1,

Categorical  $\rightarrow$  1, 0, 999.

It makes difference in distribution of dataset

When to use :- when missing data is not randomly distributed.

End of distribution Imputation:-

It like arbitrary value imputation but diff is instead of putting any value we put extreme value of dataset

Normally distributed for skewed.

$\text{mean} + 3\sigma$   $Q_1 - 1.5 \text{IQR}$

or  $\text{mean} - 3\sigma$   $Q_3 + 1.5 \text{IQR}$

where  $\text{IQR} = Q_3 - Q_1$

and  $Q_3 = 75^{\text{th}}$  percentile  
missing.  $Q_1 = 25^{\text{th}}$  percentile

When to use :- when data is not randomly distributed

Univariates  $\rightarrow$  categorical

Handling missing categorical data.

Most frequent category	or create new category "missing"
<ul style="list-style-type: none"> <li>• Mode.</li> <li>• simple to implement</li> </ul>	when to use :- when missing data is more than 10%

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='constant',
                          fill_value='missing')
```

### Random Imputation:

it is technique which can be used for both numerical as well as categorical data.

- it fills the missing value with any random number from same column.

Benefit :-

- ① easy to implement.
- ② it don't affect distribution of data.

### Missing indicator.

- for the column which has missing values we create new column and whatever it has missing value corresponding to that row in newly created column it will show True and rest it will show False
- this way model learn to difference between missing and non missing value and could improve accuracy.

Automatically select value for imputation  
Gridsearch cv - this is available in scikit-learn which itself try all the method and give best suitable method and parameters.

## Multivariate Imputation

In multivariate Imputation we take help from other columns as well unlike univariate analysis.

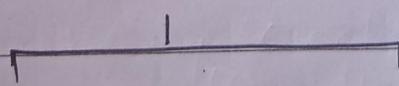
kNN

	$x_1$	$x_2$	$x_3$	$x_4$	$y$
R <sub>1</sub>			-		
R <sub>2</sub>	-				
R <sub>3</sub>	.		-		

- 4 features are four dimensions
- all the datapoints will be plotted in 4d
- for any missing value containing row or datapoint whichever other datapoint will be close its corresponding observation will used to fill the missing value for required missing datapoint.

## Oulier Treatment

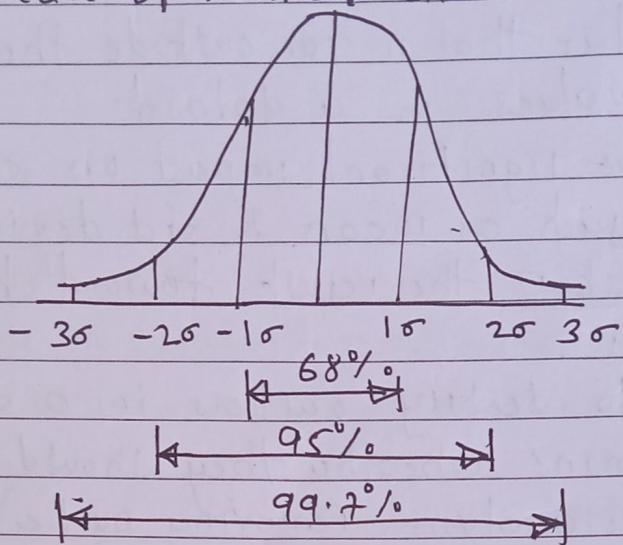
- what are the outliers?  
it is value that is far outside the normal range of values in a dataset
- they have significant impact on statistical analysis such as mean & std deviation as they can skew the result toward higher or lower values.
- it's imp to identify outliers in a dataset and determine whether they should removed or kept in the analysis. removing outlier can help improve in accuracy but it can also lead to loss of imp function. therefore it's crucial to carefully examine the data and the context in which it was collected before deciding what do with outliers.
- Effect of outlier in ML Algorithm  
outlier has very significant effect in weight based algorithm like
  - 1) Linear regression & logistic Regression
  - 3) Adaboost 5) Deep learning
 whereas it is not much harmful for tree based algorithm
  - 1) Decision Tree 2) Random forest
  - 3) XGBoost 5) Bagging & Boosting.
- How to treat outliers.



- Trimming
- Capping
- makes data thin
- fast

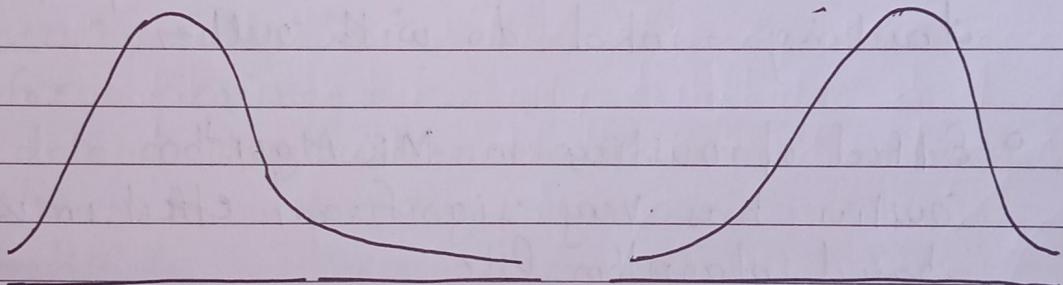
How to detect outliers?

1). In case of normal distribution:

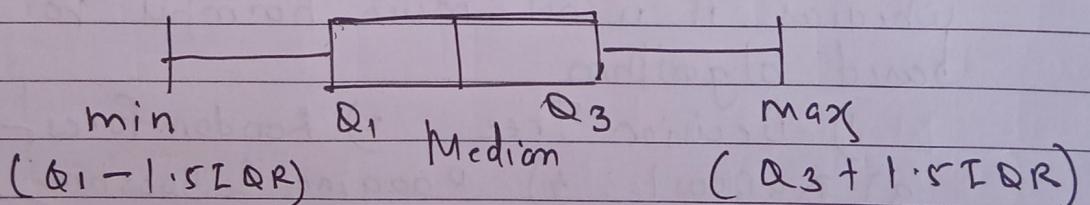


if  $(\mu - 3\sigma) < x < (\mu + 3\sigma)$  then it is outlier

2) if the data is skewed distribution.

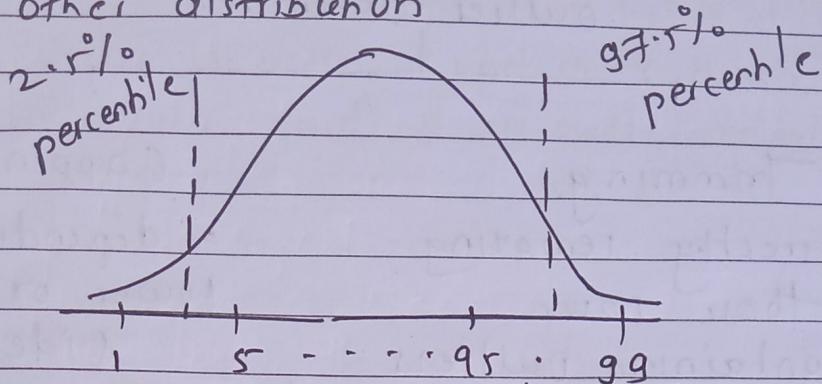


We can use box plot



anything less than min and greater than max is outlier.

In other distribution



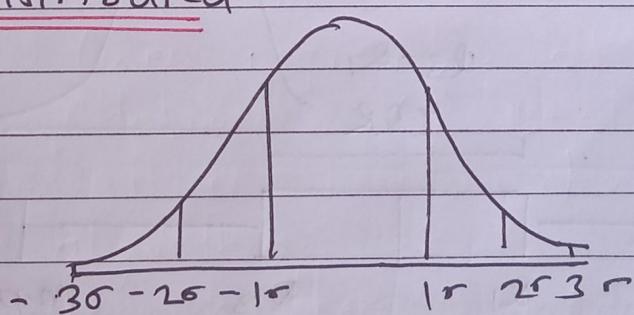
beyond 2.5<sup>th</sup> percentile and 97.5<sup>th</sup> percentile  
is outlier.

## Techniques for outlier detection and Removal.

- 1) Z-score treatment
- 2) IQR Based filtering
- 3) Percentile
- 4) winsorization.

outlier removal using Z-score.

\* Condition - data should be Normally distributed.



$$\mu \pm \sigma = 68\%$$

$$\mu \pm 2\sigma = 95\%$$

$$\mu \pm 3\sigma = 99.7\%$$

$$Z\text{-score} : z_i = \frac{x_i - \mu}{\sigma}$$

$$z_i = \frac{\text{data} - \text{mean}}{\text{std dev.}}$$

this is way of bringing any data distribution under standard normal distribution.

## Outlier treatment

### Trimming

- o directly removing those rows containing outliers

### Capping.

- o depending upon lower or higher side - we cap the value with lowest/higher restricted value

Ex. -30 30

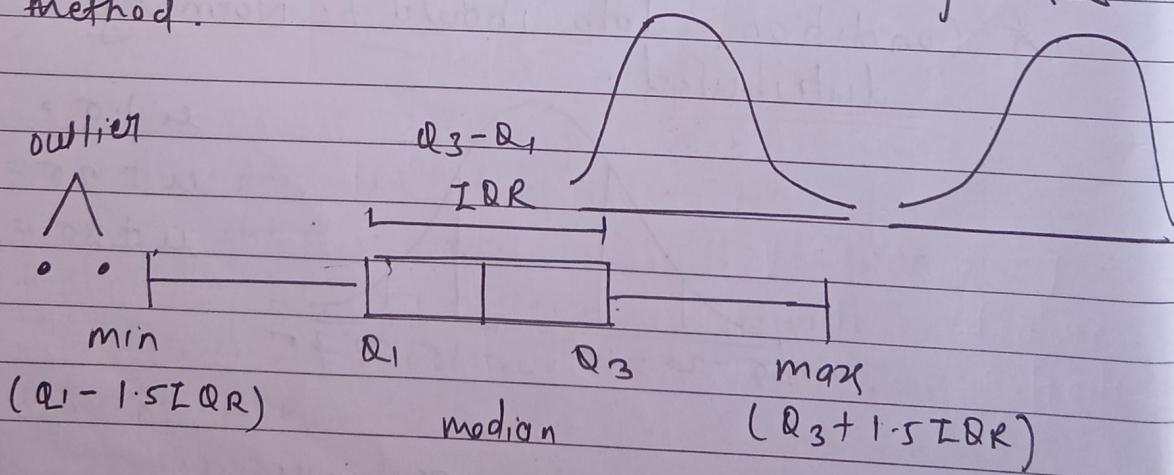
8

1 → 5

85 → 80.

80

Up. where (condition is true, — , if not — )  
**Outlier detection and Removal using IQR method.**  
 when data is not normally distributed or it is skewed then outliers treatment is done by ZQR method.



here is also same method of trimming or capping. if data is huge then we can do trimming or else we can do capping.

## Summary

If data is normally distributed then we can convert it into standard normal distribution by

$$\frac{\text{data} - \text{mean}}{\text{std. dev}} : z_i = \frac{x_i - \mu}{\sigma}$$

This is nothing but Z-score now if we consider upper limit and lower limit

$$\text{lower limit} = \mu - 3\sigma$$

$$\text{upper limit} = \mu + 3\sigma$$

} beyond this range anything is outlier,  
is based on number of data we can do trimming or capping.

If data is skewed then we can find upper limit and lower limit by IQR method.

$$\text{lower limit} = Q_1 - 1.5 \text{ IQR}$$

$$\text{upper limit} = Q_3 + 1.5 \text{ IQR}$$

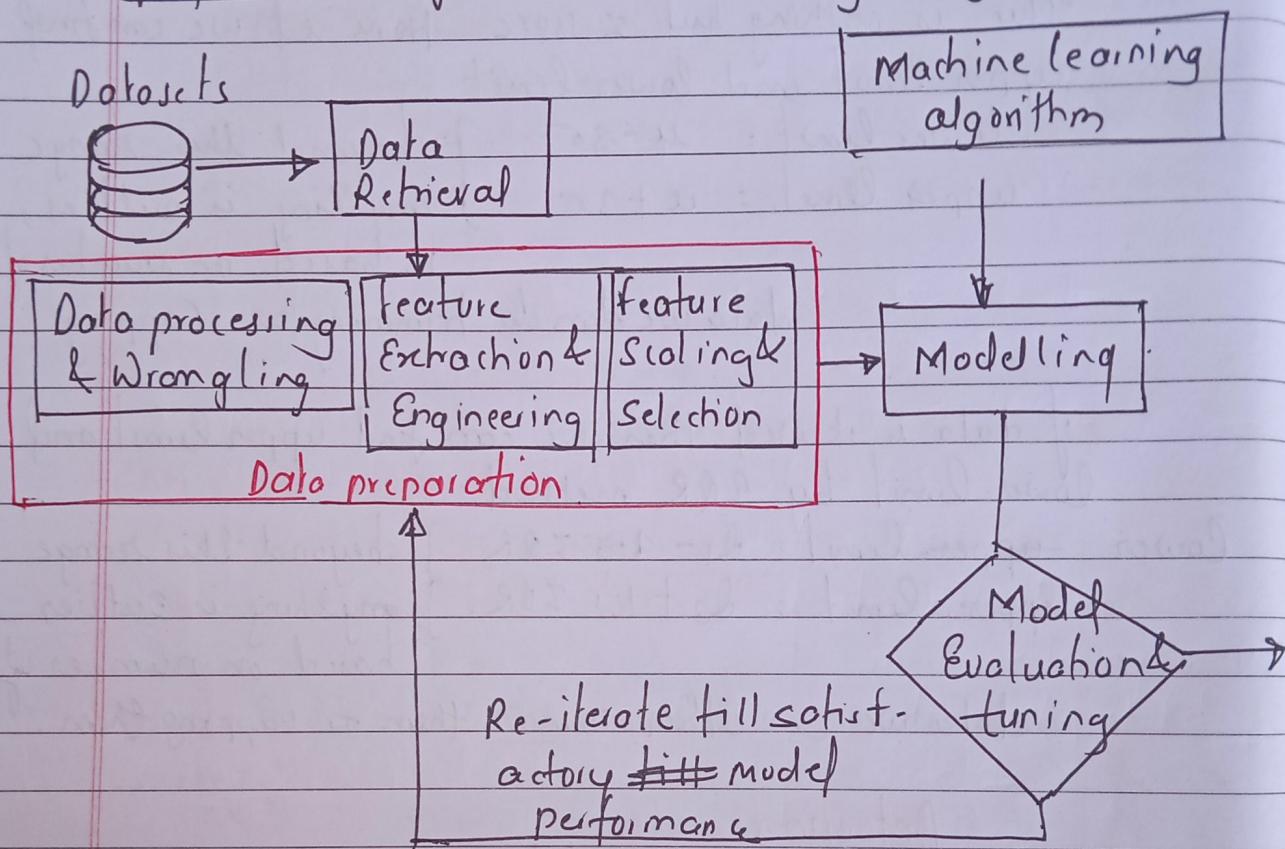
} beyond this range anything is outlier  
based on number of data we can either remove them or capping them.

## Detection Outlier Using Percentile Method (Winsorization Technique)

This is similar to rest of methods we only limit the distribution with percentile like 1st for lower limit and 99th percentile for upper limit any data point beyond it outlier and can trim or cap them based on user choice.

# Feature Engineering.

Feature Engineering is process of using domain knowledge to extract the features from raw data. These features can be used to improve the performance of machine Learning algorithm.



## Feature Engineering

- feature Transformation
- feature Construction
- feature Selection
- feature Extraction
- o missing value imputation
- o Handling categorical features
- o outlier detection
- o feature Scaling

## Feature Scaling

feature scaling is technique to standardize a data or independent features present in the data in a fixed range

for ex

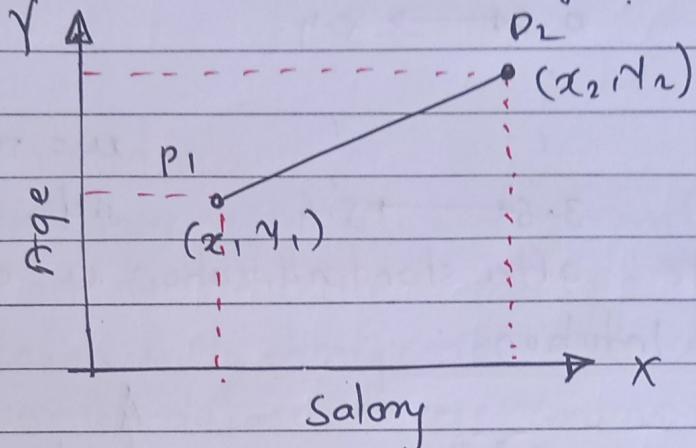
I.Q.

GPA

LPA

$$(10-100) \quad (1-10) \quad (5LPA - 20LPA)$$

why we do need feature scaling?



Deployment  
Monitoring.

Euclidean distance between  $P_1$  and  $P_2$

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$(x_2 - x_1)^2 = (83000 - 48000)^2 \\ = 1225000000 \quad \left. \right\} \text{this will dominate}$$

$$(y_2 - y_1)^2 = (50 - 27)^2 \\ = 529$$

In such scenario there is high chance your ML model will not give good performance

### Types of feature scaling

#### ① Standardization

#### ② Normalization

- Min-Max scalar

- Robust scalar

- mean normalization

- Max absolute scaling

standardization

Also called as Z-score Normalization.

$$x_i \rightarrow x'_i = \frac{x_i - \mu}{\sigma} = \frac{\text{data - mean}}{\text{std. deviation}}$$

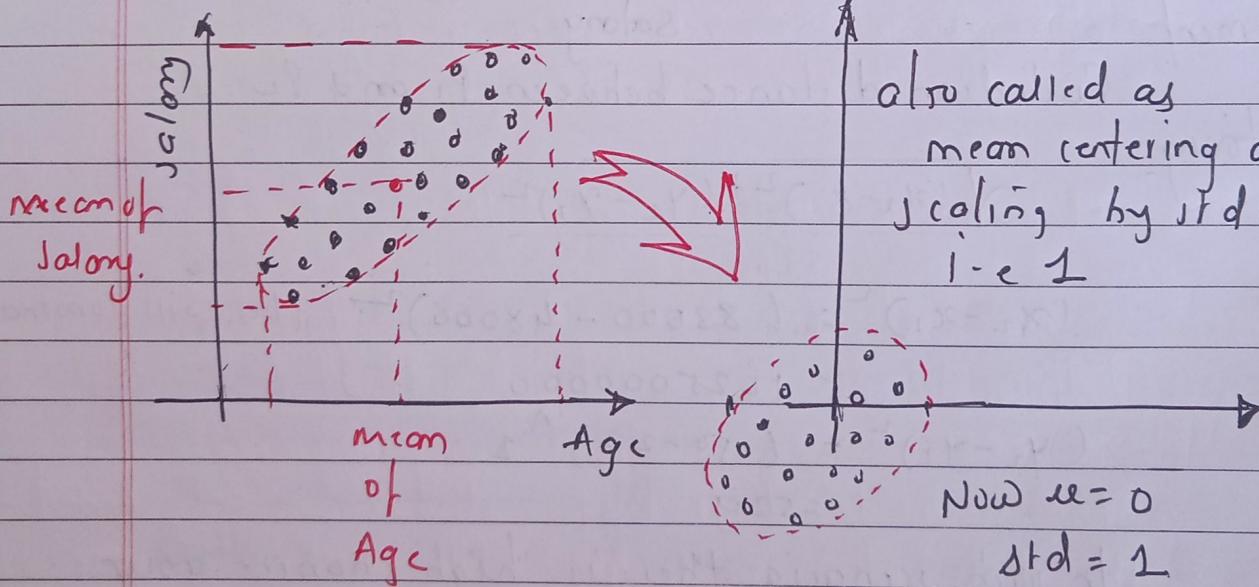
Salary	Age	Age'	salary'	
38000	27	-0.5	2.4	
40000	35	<del>0.3</del>	2	
23000	38	0.6	5.4	
78000	68	3.6	5.6	

if  $\mu = 32$   
std = 10 } for Age

$\mu = 50000$   
std = 5000

Note :- after standardization  $\mu = 0 \quad \sigma = 1$

Geometric Intuition :



When to use standardization?

Algorithm

① k-means

② k-Nearest Neighbor

Reason for applying feature scaling  
use Euclidean distance to measure

Measure the distance between pair of samples and these distance are influenced measurement units.

③ principle Component Analysis Try to get features with max variance  
(PCA)

④ Artificial Neural Network to apply gradient descent

⑤ Gradient descent improve theta calculation and learning rate.

## Normalization.

it is technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numerical column in the dataset to use common scale w/o distorting differences in the range of values or losing information.

there are four types

- ① Min max scaling      ③ Max abs scaling
- ② Mean Normalization      ④ Robust scaling

## Min-Max Scaling - Intuition

weight

130

67

81

92

:

108

$$\text{formula: } \frac{\text{data} - \text{min}}{\text{max} - \text{min}}$$

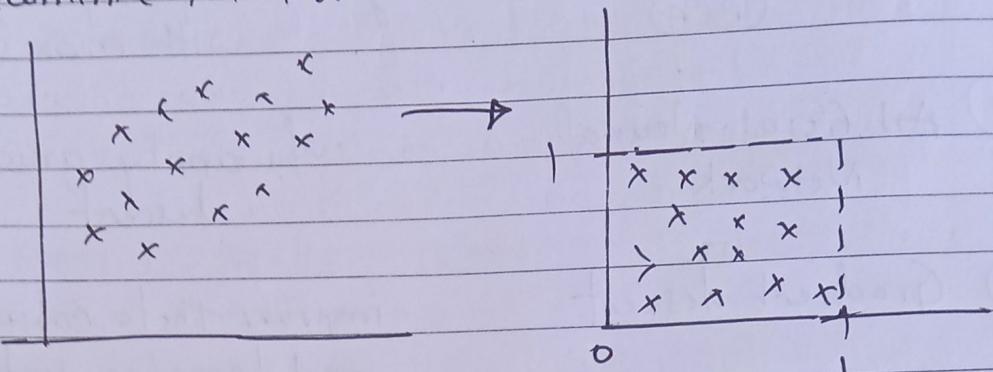
$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

$$x_{\min} = 67, x_{\max} = 130$$

$$\frac{130 - 67}{130 - 67} = 1$$

in this transformation your updated value must be in range of 0 cm [0, 1].

## Geometric Intuition



here we try to squeeze the data into unit square

Code

```
from sklearn.preprocessing import MinMaxScaler
```

### Mean Normalization

$$x'_i = \frac{x_i - \bar{x}_{\text{mean}}}{x_{\text{max}} - x_{\text{min}}} \quad \left\{ \text{mean centering} \right\}$$

### Max Abs scaling.

$$x'_i = \frac{x_i}{|x_{\text{max}}|}$$

```
from sklearn.preprocessing import MaxAbsScaler
```

used when data is sparse (i.e. when data has many zeros)

### Robust scaling

$$x'_i = \frac{x_i - \bar{x}_{\text{median}}}{ZQR} \quad \left\{ Q_3 - Q_1 \right\}$$

Robust scaling

- Note: it is Robust to outliers

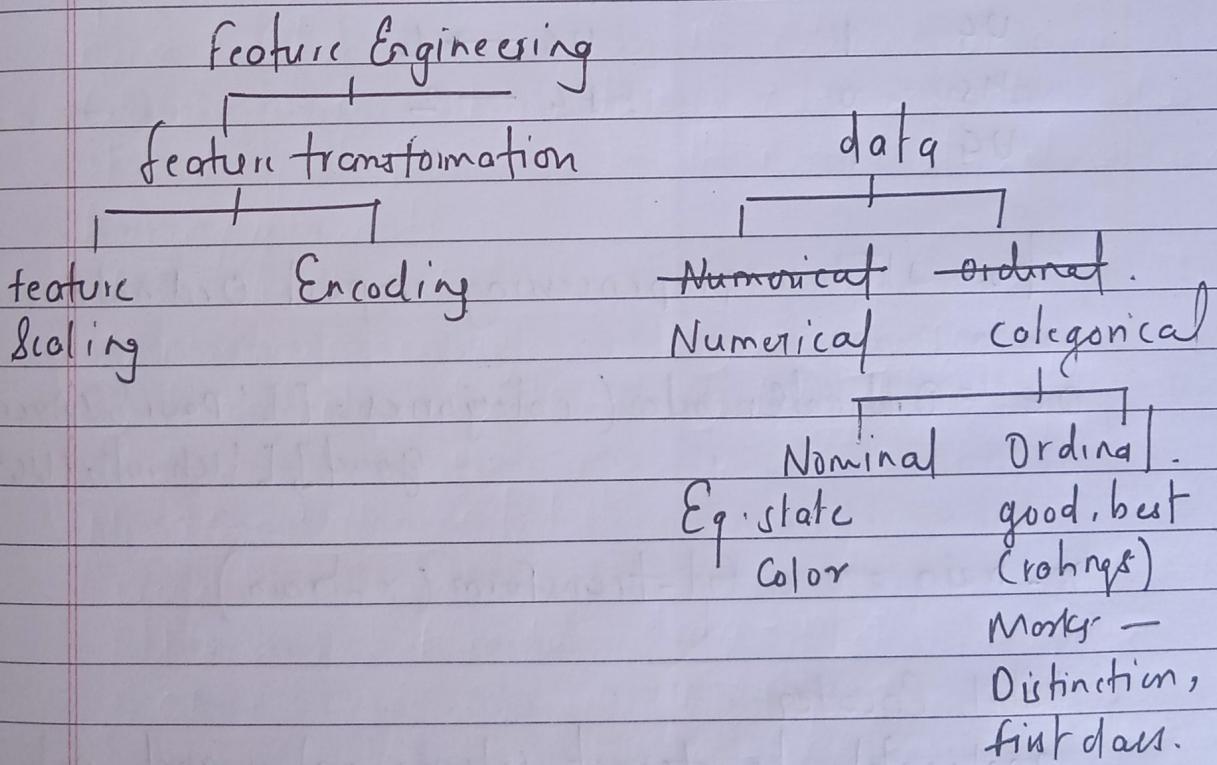
## How to choose Normalization or standardization

Q. 1: first confirm whether we actually need to do scaling or not required?

Ans: Actually most of the time we do standardization very less time we prefer for normalization when we know the max and min value. like image processing in CNN.

## Encoding Categorical Data

- ordinal Encoding | Label Encoding.



Let's see how to encode categorical data?

There are 2 techniques

- ① Ordinal Encoding → apply on ordinal data
- ② One-hot Encoding → apply on Nominal data

Label Encoding is like ordinal Encoding only difference is only applicable to target column y.

Code is -

```
from sklearn.preprocessing import LabelEncoder.
```

## Ordinal Encoding

Education

HS	0	- PG > UG > HS — ordinal data
UG	1	so here you give order to
PG	2	category
PG	2	HS → 0
UG	1	UG → 1
HS	0	PG → 2
UG	1	

```
from sklearn.preprocessing import OrdinalEncoder
```

```
oe = OrdinalEncoder(categories=[['poor', 'Average', 'good'], ['school', 'UG', 'PG']])
```

```
x-train = oe.fit_transform(x-train)
```

## Label Encoder

applicable to target column only and not feature column.

```
from sklearn.preprocessing import LabelEncoder
```

```
le = LabelEncoder()
```

```
le = fit(y-train)
```

One hot Encoding  
used for Nominal categorical data

- One hot Encoding.
- Dummy Variable trap.
- One using most frequent variable
- Example

One hot Encoding.

Color	Target	Color Y	Color B	Color R	Target
yellow	0	1	0	0	0
yellow	1	1	0	0	1
Blue	1	0	1	0	1
yellow	0	1	0	0	0
red	0	0	0	1	0
yellow	1	1	0	0	1
Red	1	0	0	1	1
red	1	0	0	1	1
Blue	0	0	1	0	0

Dummy Variable Trap

if we have n number category  
then keeps only n-1 column  
remove 1 column since it is not needed  
logically.

One hot Encoding using frequent variable category.

if there is frequent 50 category and 40 are  
most frequent then rest 10 category will merge to  
one category named 'others'

1) One hot Encoding using Pandas  
pd.get\_dummies(df, columns=['fuel', 'owner'])

2) dummy trap variable trap (k-1 hot Encoding)  
pd.get\_dummies(df, columns=['fuel', 'owner'], drop\_first=True)

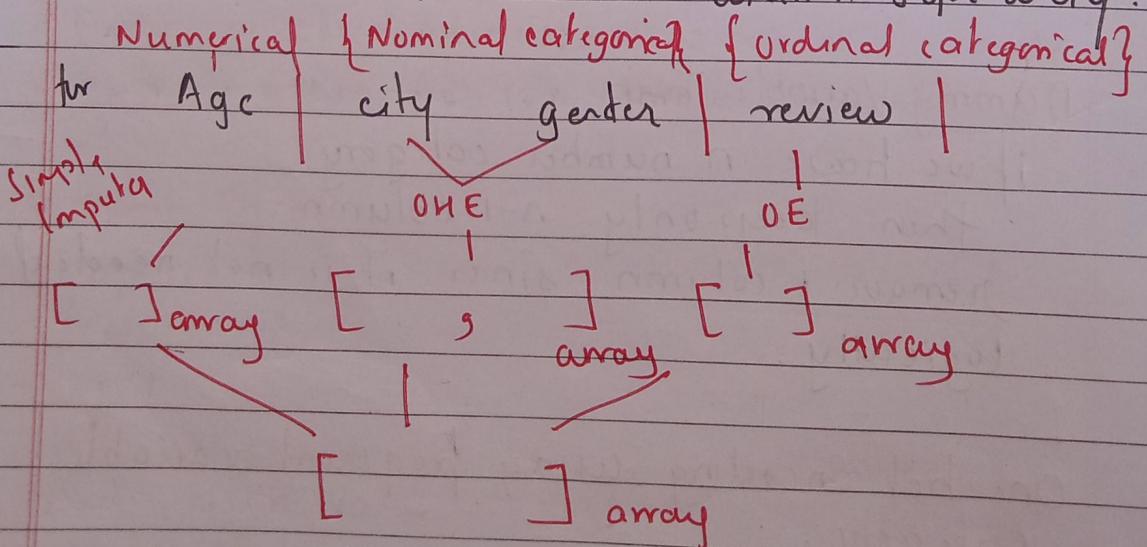
3. One hot Encoding using sklearn.

Code :-

- from sklearn.preprocessing import OneHotEncoder
- o ohe = OneHotEncoder(drop='first', sparse=False, dtype=np.int32)
- o x\_train\_new = ohe.fit\_transform(x\_train[['fuel', 'owner']])

### Column Transformer.

In a dataset we have different features column and need to perform different operation at it is difficult to handle for each column separately.



which is cumbersome task.

from sklearn.compose import ColumnTransformer

transformer = ColumnTransformer(transformer=[  
 ('tnf1', SimpleImputer(), ['Fever']),  
 ('tnf2', OrdinalEncoder(categories=[[['mild', 'strong']]],  
 ['cough']),  
 ('tnf3', OnehotEncoder(sparse=False, drop='first'),  
 ['gender', 'city'])], remainder=  
 ('passthrough')

transformer.fit\_transform(x\_train).shape

transformer.transform(x-test).shape

## Variable Transformation

mathematical transformation.

1) log transformation

Box-Cox

2) Reciprocal transformation

Yeo-Johnson

3) power ( $\sqrt{q}$  /  $\sqrt{r}$ )

The role of above transformation is that when we will not have normal distribution data thus mathematical transformation will make Normal distribution.

function Transformer

sklearn

function transformer

power  
transformer

Quantile  
transformer

- log transform

- Box-Cox

- Reciprocal

- Yeo-Johnson

-  $\sqrt{q}$  /  $\sqrt{r}$

- Custom

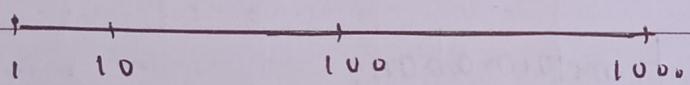
Since we apply these transformation when data is not normally distributed. How we will get to know whether data is normal dist or not?

- 1) sns.distplot
- 2) pd.skew()
- 3) Q-Q-plot

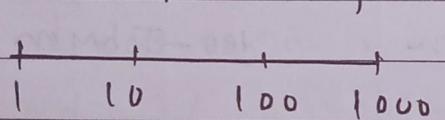
### ① Log transform.

- it is not applicable to the features containing negative values
- it is applicable to right skewed data

Note:- Working of log.



after  $\rightarrow \log$



bring it at equidistant point

### ② Reciprocal transformation. $1/x$

Small value  $\Rightarrow$  Large dot value

### ③ Jquare ( $x^2$ )

comes b/wed when data is left skewed.

### ④ sqrt $\cdot \sqrt{x}$

Code:-

```
from sklearn.preprocessing import FunctionTransformer  
trf = FunctionTransformer(func=np.log1p)
```

`trf x-train-transformed = trf.fit_transform(x-train)`  
`x-test-transformed = trf.transform(x-test)`

### Power transformations

Box-Cox

Yeo-Johnson

#### Box-Cox transformation

$$x_i' = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$$

- the exponent here is a variable called lambda ( $\lambda$ ) that varies over the range of -5 to 5 and in the process of searching we examine all values of  $\lambda$  finally we choose optimal value (resulting in best approximation to a normal distribution) for your variable.

- this transformation is only applicable to a number not greater than zero

$$\therefore n \neq n \geq 0 \quad \& \quad n \neq -ve$$

this restriction is overcome by Yeo-Johnson

#### Yeo-Johnson Transform

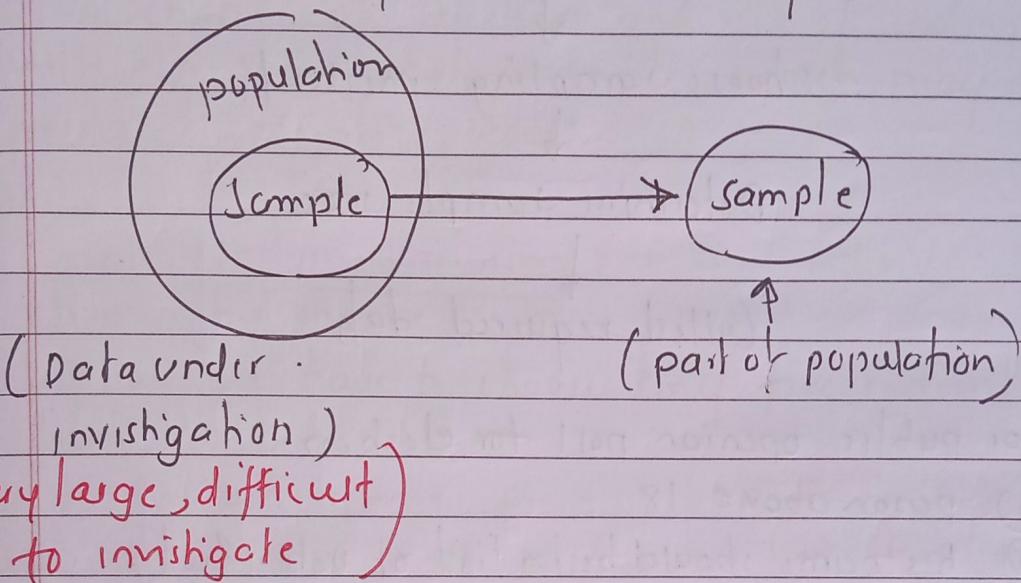
$$x_i^{(\lambda)} = \begin{cases} [(x_i + 1)^\lambda - 1] / \lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i) + 1 & \text{if } \lambda = 0, x_i > 0 \\ - [(-x_i + 1)^{2-\lambda} - 1] / (2-\lambda) & \text{if } \lambda \neq 2, x_i < 0 \\ - \ln(-x_i + 1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

This transformation is somewhat adjustment to Box-Cox Transformation by which we can apply it to negative numbers

```
from sklearn.preprocessing import PowerTransformer  
pt = PowerTransformer(method='box-cox')
```

feature selection "Sampling"

Sampling is a method that allows us to get information about the population based on statistics from subset of the population (sample) without having to investigate every individual.



for Ex : to calculate avg height of Men in Bombay  
it's not possible to reach every male so we can't realize entire population

- instead we can take multiple sample and calculate average height, but another question may arise how to take the samples.
  - like if we take the sample of only basketball players this would not consider a good sample because basketball players are taller than avg men
- why done need sampling*
- 1) Sampling is done to draw conclusion about population from samples
  - 2) It requires less time rather selecting every item in population
  - 3) Sample selection is cost efficient method.
  - 4) It is less cumbersome and more practical than analysis of entire population.

## Steps involved in sampling:-

Step 1:- identify and define target population

↓  
Select sampling frame

↓  
choose sampling method.

↓  
Determine sample size

↓  
Collect required data

for public opinion poll for election

- ① person above 18
- ② his name should be in list of voter list
- ③ different sample taken from different region
- ④ larger the sample size, more accurate our inference about the population
- ⑤ collect data from sample

## Different types of Sampling technique

### Sampling Methods

#### probability Sampling

- ✓ 1. Simple random
- ✓ 2. Systematic
- ✓ 3. stratified
- ✓ 4. cluster

#### Non probability Sampling

- 1. Convenience
- 2. Quota
- 3. Judgment
- 4. Snowball

Probability Sampling :- every element has equal chance of selecting it's true representation of the population

Non probability Sampling :- all elements do not have equal chance of selecting and risk of ending up with non-representative sample which may not produce generalize result.

Simple random sampling :- Every individual can be chosen by chance

Suppose we have to choose any five individuals from 20.

1 2 3 4 5  
6 7 8 9 10  
11 12 13 14 15  
16 17 18 19 20

Random sampling  
2 6 10  
13 18

Systematic Sampling :- in this first individual select randomly but others are selected using fixed sampling interval.

$$3, 3+4=7, 7+4=11, 11+4=15,$$

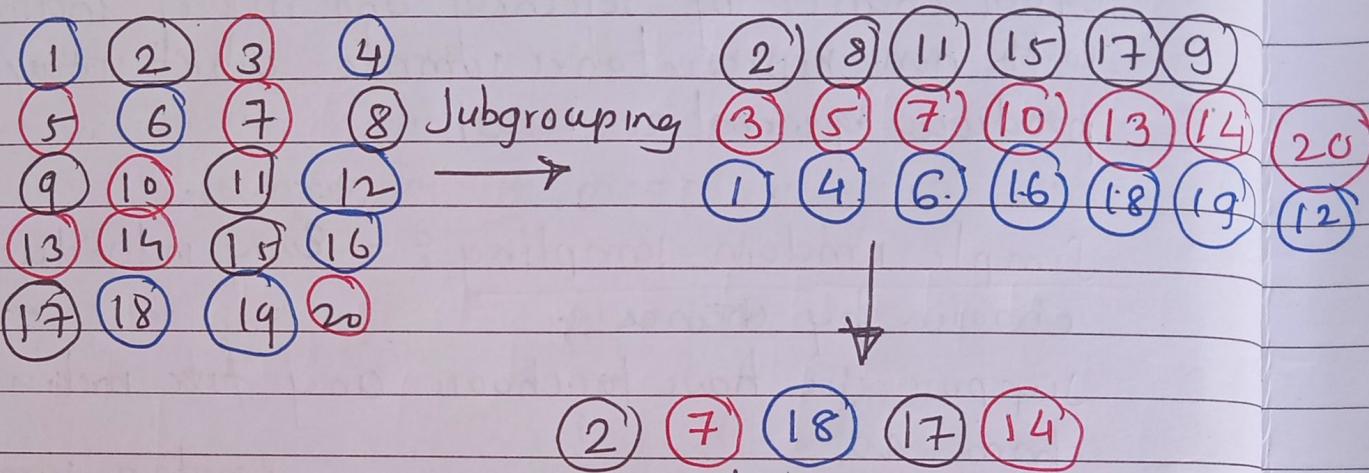
1 2 3 4  
5 6 7 8  
9 10 11 12  
13 14 15 16  
17 18 19 20

15+4=19  
3 7 11  
15 19

it is more convenient than simple random sampling

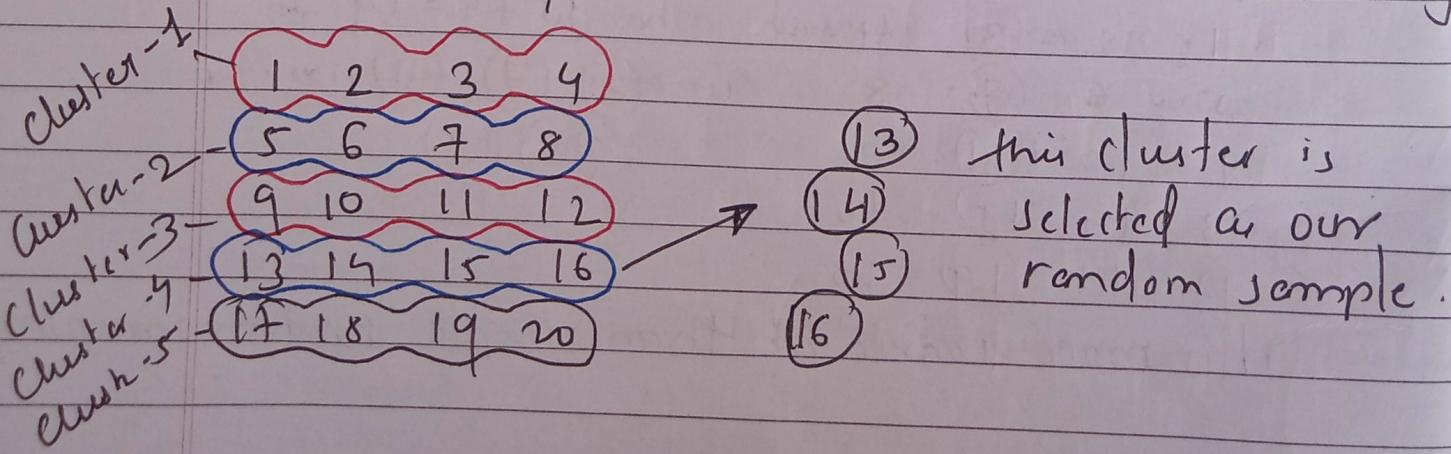
## stratified sampling :-

here we divide population into subgroups (called strata) based on different traits like gender, category, etc. and then select sample from these subgroups.



We used this type of sampling when we want representation from all subgroups of the population.

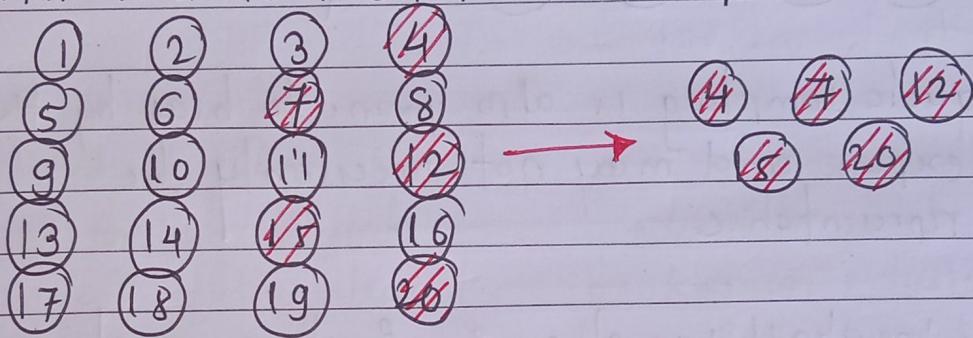
cluster sampling :- here we use the subgroups of population as sampling unit rather than individuals, the population divided into subgroups, known as cluster and whole cluster is randomly selected to be included in study.



## Non-Probability Sampling :-

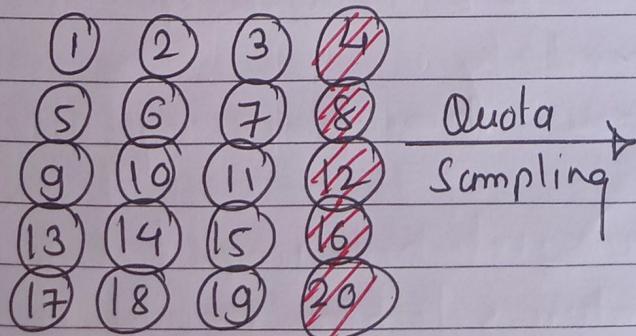
① Convenience Sampling :- this is perhaps easiest method of sampling because individual are selected based on their availability and willingness to take part

Here let say individual number 4, 7, 12, 15 & 20 want to be part of our sample and hence we will include them in the sample



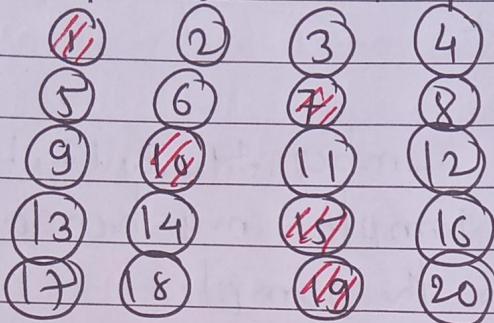
② Convenience sampling is prone to significant bias, bcoz the sample may not be the representation of the specific characteristic such as religion or say, the gender of the population

② Quota Sampling :- In this type of sampling, we choose item based on predetermined characteristic of the population; consider that we have to select individual having a number in multiples of four for our sample.



## Judgement sampling :-

it is selective sampling. it depends on the judgement of the expert when choosing whom to ask to participate.



} like random sampling  
but someone uses his  
experience to pick them  
up

- quota sampling is also prone to bias by the expert and may not necessarily be representative

Snowball sampling :- Existing people are asked to nominate further people known to them so that sample increases in size like a rolling snowball.

$$1 \rightarrow 6 \rightarrow 11 \rightarrow 14 - 19$$

there is significant risk of selection bias in snowball sampling as the referenced individual will share common traits with the person who recommended them

## feature Selection

In real life data science problem it's almost rare that all variables in dataset are useful for building model.

### what is feature selection ?

To find subset of features that allows one to build optimised model of studied phenomena.

#### Supervised Techniques

- used for labelled data & to identify the relevant features for increasing efficiency of the supervised model like classification

#### Unsupervised techniques

- used for unlabelled data Ex. k-means, clustering, PCA, Hierarchical clustering

#### & Regression

- Linear Regression
- decision Tree
- SVM

#### filter Method

- information Gain
- chi-square test
- correlation coefficient
- variance threshold

#### Wrapper Method

- forward feature selection
- Backward feature elimination
- recursive feature elimination

#### Embedded Method

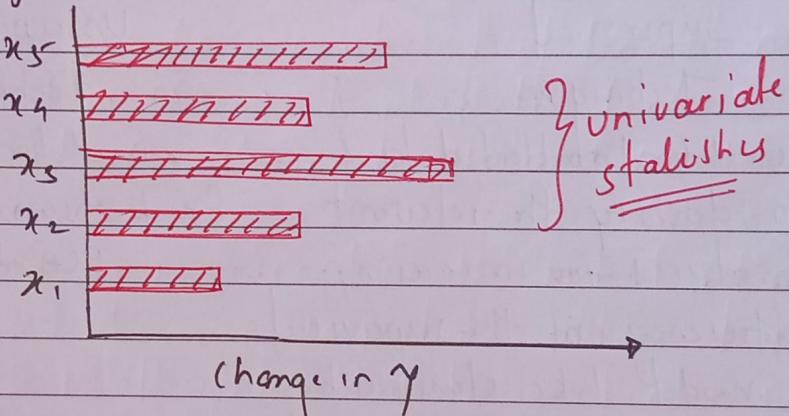
- LASSO Regularization (L1)
- Ridge (L2)
- Random forest

## filter Method :-

this method is faster and less computationally expensive than wrapper method.

## information Gain :-

- it calculate reduction in entropy from transformation of dataset
- it can be used for feature selection by evaluating the information gain of each variable with respect to target variable



## Chi-Square test :- it is used for categorical data of

- we calculate chisquare between each feature and target and then select desired number of feature with best chisquare score
- In order to correctly apply the chisquare test the following condition have to be met.
  - the variable have to be categorical
  - sample independently
  - and values should have frequency greater than 5.

from sklearn, feature\_selection import SelectKBest  
from sklearn.feature\_selection import chi2

## Correlation coefficient :-

- it is measure of linear relationship between two or more variables, through correlation we can predict one variable with the other
- th. logic is good variable highly correlate with target but they should uncorrelated among themselves
- if two feature are correlated we choose only one as second doesn't add additional information
- we use for it Pearson Correlation

Variance threshold:- it is simple baseline approach to feature selection. It remove all the feature who does not meet some threshold. By default it removes all zero variance features i.e. feature with same value in all samples or observation.

## Wrapper Method :-

- The wrapper method usually result in better predictive accuracy than filter method.
- here based on specific machine learning algorithm we are trying to fit a given dataset. It follows greedy search approach by evaluating all the possible combination of feature against the evaluation criterion.

### Type : forward feature selection

This is iterative method wherein we start with performing feature against target variable. Next we select another variable along with first selected feature variable against same target variable in order check improved in accuracy. This process continues until break condition is achieved.

## Backward feature Elimination :-

- it exactly works opposite to forward feature selection
- Here we start with all the features to build the model and iteratively remove the features and based on evaluation measure we repeat the process until preset condition is achieved.

## Recursive feature Elimination:-

- Given an external estimator that assign weights to features. (e.g coefficient of linear model) the goal of RFE is to select the features by recursively considering smaller and smaller sets of features
- first estimate is trained on the initial set of features and each feature importance obtained either through coefficient attribute or feature importances attribute

## Embedded Method :-

this method combines benefits of both filter and wrapper method by including interaction of features but also maintaining reasonable computational cost.

## LASSO Regression (L1) :-

Regularization consists of adding penalty to the different parameters of the machine learning model to reduce freedom of the model i.e. to avoid overfitting.

- In Lasso model regularization the penalty is applied over the coefficient that multiply each predictor
- from different types of regularization Lasso or L1 has property that can shrink some coefficient to zero therefore that feature can be removed from the model.

### Randomforest Importance :-

- it is kind of Bagging algorithm that aggregate specified number of decision tree
- the tree based strategies used by random forest naturally rank by how well they improve the purity of Node or in other words decrease the impurity (Gini Impurity) over all the trees
- Nodes with greatest ~~decrease~~ in impurity happen at the start of tree, while nodes with least impurity occurs at the end of the tree, thus by pruning trees below a particular node, we can create subset of most imp features

Model selection & training :- After preparing data and identify relevant features the next step is to select suitable model. This model can be regression model or classification or clustering model depend on the nature of problem

Model training :- Once the model is selected it needs to be trained using training data set. During the training process, the model is adjusted to minimize difference between the predicted output and actual output.

# Different Model Evaluation technique

## Model classification technique for classification problem

- accuracy
- precision
- f1 score
- AUC
- confusion Matrix
- Recall
- ROC

1) Accuracy :- In binary classification problem where output either 0 or 1

$$\text{accuracy} = \frac{\text{no. correct prediction}}{\text{total prediction}}$$

$$\text{for } 80 = \frac{80}{100} = 0.80 = \underline{\underline{80\%}}$$

2) for multiclassification problem :- it is similar to binary classification , No change

$$\text{accuracy} = \frac{\text{no. of correct prediction}}{\text{total prediction}}$$

Ques: How much accuracy is good ?  
it depend upon the problem we are solving

### The problem with Accuracy ?

from accuracy score we can only get to know that model is making som. mistakes but what type of mistakes that we might don't know.

Type 1:- class is actual belong to 1 but predict as zero .

Type 2 : class 1, actually belong to zero but predict 1

To overcome this we have another solution called as confusion Matrix.

		Prediction		
		1	0	0
Actual	1	True Positive (TP)	False Negative (FN)	0
	0	False Positive (FP)	True Negative (TN)	0

from confusion Matrix we can calculate accuracy but from accuracy we can't calculate confusion Matrix.

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Type 1 Error :- False positive is called as Type 1 means model is predicting 1 but actually it belongs to zero.

Type 2 Error : False negative is called as Type 2 means model is predicting 0 but actually it belongs to one.

for multi classification problem

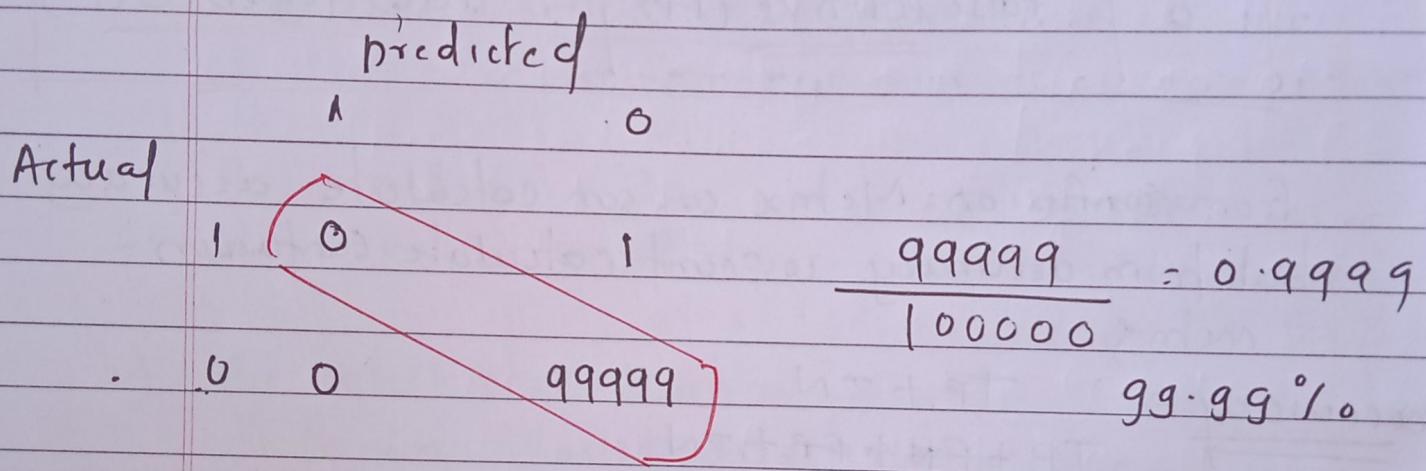
		0	1	2	
		0	7	0	5
Actual	1	3	2	3	0
	2	1	5	7	0

$$\text{accuracy} = \frac{\text{diagonal element (7+2+7)}}{\text{sum of all element}}$$

when accuracy is misleading ?

Ans:- Imbalanced dataset.

- for Ex. we have created model for terrorist detection. for which we have dataset in which ~~too~~ out of 100000 only 1 is terrorist.
- In such cases model will highly trained on normal people and will predict everyone as not terrorist with accuracy almost close to zero.



To overcome this problem when accuracy is misleading due to imbalance dataset we have other method i) Precision and Recall.

If type 1 Error are more dangerous then we use precision

If type 2 Error are more dangerous then we use recall.

Precision :- for type 1 Error  $\rightarrow$  FP  
which mean it is  $\uparrow$  belong to 0 but predicted 1

model A

for Ex

model B

Page

		predicted		predicted	
		spam	Notspam	spam	Notspam
Actual	spam	100 TP	120 FN	100 TP	190 FN
	Notspam	30 FP	700 TN	10 FP	700 TN

if we calculate accuracy for both model it will be same but which will finalize to deploy

in this case FP is more dangerous mean the mail was not spam but put in spam due to which we can miss imp information

$$FP_A > FP_B$$

that's why model B is good to finalize for deploy

precision:  $\frac{TP}{\text{predicted positive}}$

model A

$$\frac{100}{130} = 0.76$$

B

$$\frac{100}{110} = 0.90$$

{ good }

My when type 2 is dangerous mean FN is dangerous we take Recall.

		model A		model B	
		Cancer	Not Cancer	Cancer	Not Cancer
Cancer	Cancer	1000	200	1000	500
	Not Cancer	800	800	500	8000

in this case FN is more dangerous person has cancer but detect No cancer he may miss the treatment

$$\text{Recall} = \frac{\text{TP}}{\text{Actual positive}}$$

Model A

$$\frac{1000}{1200} = 0.83$$

Model B

$$\frac{1000}{1500} = 0.66$$

Model A is good to deploy since less no of mistakes or FN (false negative)

but sometimes we don't know which type Error is more dangerous type 1 or type 2 at that time we have another technique F1 score which harmonic mean of Precision and Recall.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Harmonic mean why?

- it is always on the lower side unlike arithmetic mean which gives exact centre
- Harmonic mean always penalize towards lower value.

### Multiclass Precision and Recall.

Predicted

	Dog	Cat	Rabbit		
Actual	Dog	25	5	10	40
Dog	0	30	4	34	
Cat	4	10	20	34	
Rabbit	29	45	34	108	

To calculate Precision

First we will calculate individual precision

$$\frac{TP}{Predicted} = \frac{25}{29}$$

$$= 0.86$$

$$P_C = \frac{30}{45}$$

$$= 0.66$$

$$P_R = \frac{20}{34}$$

$$= 0.58$$

There two method of combining to mean

1) Macro precision

$$\frac{0.86 + 0.66 + 0.58}{3} = 0.70$$

2) Weighted precision

$$\frac{0.86 \times \frac{40}{108} + 0.66 \times \frac{34}{108} + 0.58 \times \frac{34}{108}}{3} = 0.71$$

To calculate Recall

$$\frac{TP}{Actual Position} = R_D = \frac{25}{40}, R_C = \frac{30}{34}, R_R = \frac{20}{34}$$

$$= 0.62$$

$$0.88$$

$$= 0.58$$

again there two method

1) Macro Recall

$$\frac{0.62 + 0.88 + 0.58}{3} = 0.68$$

2) weighted Recall

$$0.62 \times \frac{40}{108} + 0.66 \times \frac{34}{108} + 0.58 \times \frac{34}{108}$$

$$(0.62 \times 0.37) + (0.66 \times 0.31) + (0.58 \times 0.31)$$

$$\frac{0.22 + 0.46 + 0.17}{0.20 + 0.17} = 0.59$$

similar we do f1 score

$$F1_{Dog} = \frac{2PR}{P+R}_{Dog}; F1_{Cat} = \frac{2PR}{P+R}_{Cat}; F1_{Robbi} = \frac{2PR}{P+R}_{Robbi}$$

Macro f1 score

Weighted f1 score

### Softmax Regression.

- accuracy, confusion matrix, Precision and Recall and f1score also work very fine with Binary class, but what if multiclass problem is there, then may we don't get good accuracy or result for that we have softmax Regression or Multinomial Regression
- logistic Regression is specialization of softmax

what is softmax function?

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}} \quad k = \text{no of class } \{3\}$$

placement

if yes  $\rightarrow 1$ , No  $\rightarrow 2$ , opt - 3

$$\sigma(z)_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

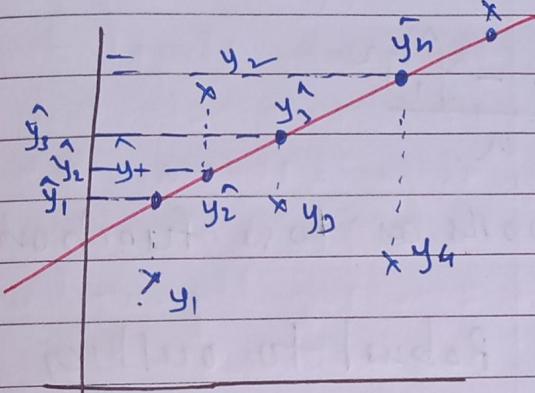
$$\sigma(z)_2 = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

$$\sigma(z)_3 = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

## Evaluation Matrix for regression Model.

- 1) MAE (Mean absolute Error)
- 2) MSE (Mean Square Error)
- 3) RMSE (Root Mean Square Error)
- 4) R<sup>2</sup>-score (coefficient of determination)
- 5) Adjusted R<sup>2</sup> score

1) MAE = Mean absolute Error



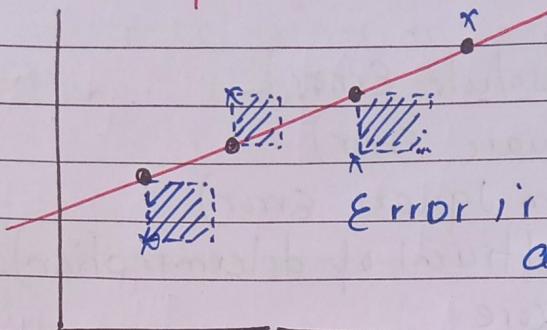
$$\text{MAE} = \frac{|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + |y_3 - \hat{y}_3| + \dots + |y_n - \hat{y}_n|}{n}$$

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

Advantages : Error is given in same unit as of y the target column  
is Robust to outliers.

Disadvantages: graph of modulus function is not differentiable at zero.

## Mean Squared Error (MSE)



Error is nothing but ~~area~~ area of square.

$$\text{MSE} = \frac{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2}{n}$$

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

Advantages :- it works as loss function

disadvantages :- Not Robust to outliers

Here since Error is in terms of square it's difficult while Explaining you have to explain with unduroot-

## RMS E (Root Mean square Error)

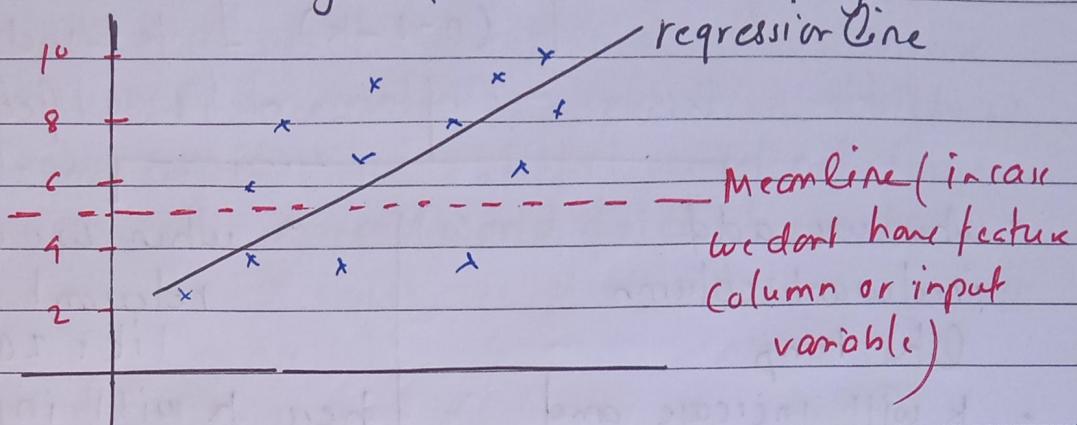
$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

benefit :- Error you get in terms of y-unit

disadvantages :- it is not robust to outliers

## $R^2$ - score

it is nothing but comparison of performance of regression line against mean line of  $y$ .



$$R^2 = 1 - \frac{SSR \text{ (Sum of Squared Error for Regression)}}{SSM \text{ (Sum of Squared Error for mean line)}}$$

$$= 1 - \frac{\left[ \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right] \text{Regression Line}}{\left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right] \text{mean line}}$$

it should be in range of [0-1] zero to one

if  $R^2$ -score = 0.80

- it interpretation like this.

Variance in output column is coming 80% ~~from~~ from Features we selected for model training.

- $R^2$ -score can be improved as no of feature are added
- but problem arise when we add irrelevant feature it supposed to decrease but it increase or remain same this flaw is overcome with Adjusted  $R^2$

Adjusted  $R^2$ 

$$R^2_{adj} = 1 - \left[ \frac{(1-R^2)(n-1)}{(n-1-k)} \right]$$

$R^2 = R^2 - \text{score}$   
 $n = \text{no of rows}$   
 $k = \text{independent features}$

- when we add irrelevant column like temp

- k will increase and slight decrease in denominator

- and since there is no effect of  $R^2$  nominator will be constant and fraction will increase

- so overall adj  $R^2$  will decrease

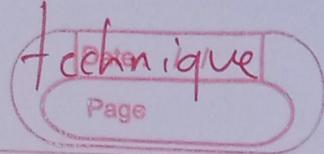
- when we add relevant column like IQ

here k will increase and that's why slight decrease in denominator but since it will increase  $R^2$  and  $(1-R^2)$  will become closer to zero rapidly and overall nominator will decrease rapidly and overall fraction will be less.

- so overall adj  $R^2$  will increase

# Classification Model Evaluation

(Free of cutoff value)



AUC-ROC curve.

(Receiving Operating characteristic curve)

- For classification model Evaluation we have metric like Recall, F1, accuracy and precision which is based to a cutoff value, but it could not same for all the case in real world dataset sometimes we may need to check for different range cutoff value [1% - 99%]
- Here we plot the graph between TPR and FPR rate for good model TPR supposed to be higher than FPR

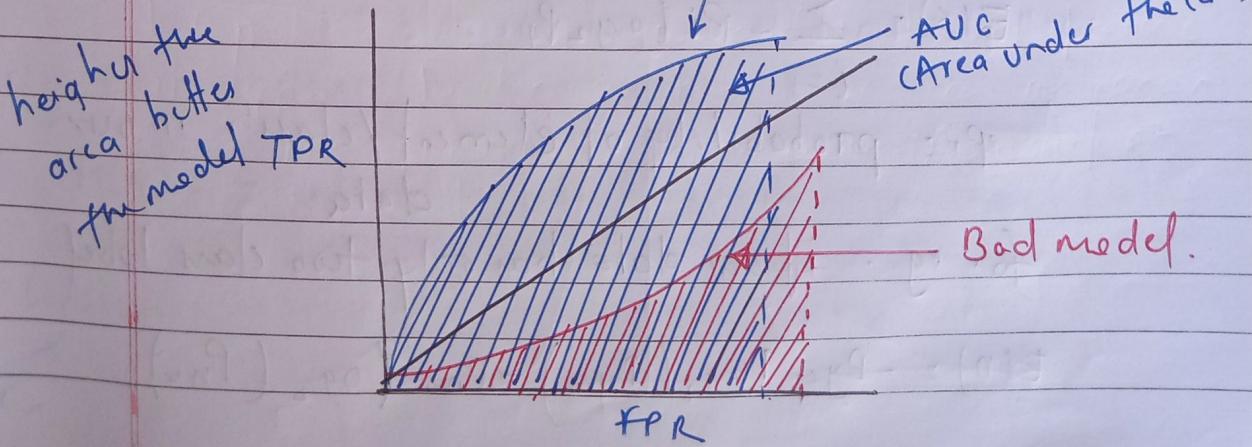
$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

		predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Inter. Question: Can you tell me metric which is free of cutoff

Ans AUC-ROC curve



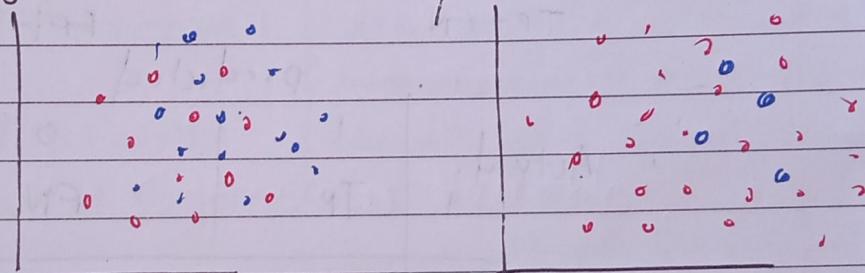
- Entropy, Information Gain and Gini impurity are validation measure for decision tree or tree based model.

what is Entropy :- In layman terms it is nothing but measure of disorder - or you can also call as measure of purity / impurity

In physics term

Entropy of vapour  $\gg$  liquid water entropy  $\gg$  Entropy of ice

In data science terminology :- More the knowledge lesser the entropy



to predict is difficult.

high Entropy

to predict is quite easy

less Entropy

How to calculate Entropy :

$$E(S) = \sum_{i=1}^C -P_i \log_2 P_i$$

$P_i$  : probability of element / class 'i' in our data

for eg . if our data has only two class label .

$$E(D) = -P_{ys} \log_2(P_{ys}) - P_{no} \log_2(P_{no})$$

for Example

Salary	Age	Purchase	Salary	Age	Purchase
20k	21	Yes	34k	31	No
10k	45	No	15	37	No
60k	22	Yes	69	21	Yes
15k	31	No	25	28	No
12k	18	No	32	27	No

$$H(D) = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$$

$$= -\frac{2}{3} \log_2\left(\frac{2}{3}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right)$$

$$= -\frac{1}{5} \log_2\left(\frac{1}{5}\right) - \frac{4}{5} \log_2\left(\frac{4}{5}\right)$$

~~$$= -\frac{2}{3} \frac{\log\left(\frac{2}{3}\right)}{\log 2} - \frac{3}{5} \frac{\log\left(\frac{3}{5}\right)}{\log 2}$$

$$H(D) = 0.72$$~~

$$= (-0.66 \times -0.58) - (0.6 \times -0.73) \quad \left\{ \text{Entropy is low} \right\}$$

$$= 0.38 + 0.43$$

$$= \underline{0.81} \quad \text{www}$$

$$= \underline{\underline{0.97}}$$

{ Entropy is high }

if we would have all the y label from same class either Yes or No then entropy would be zero.

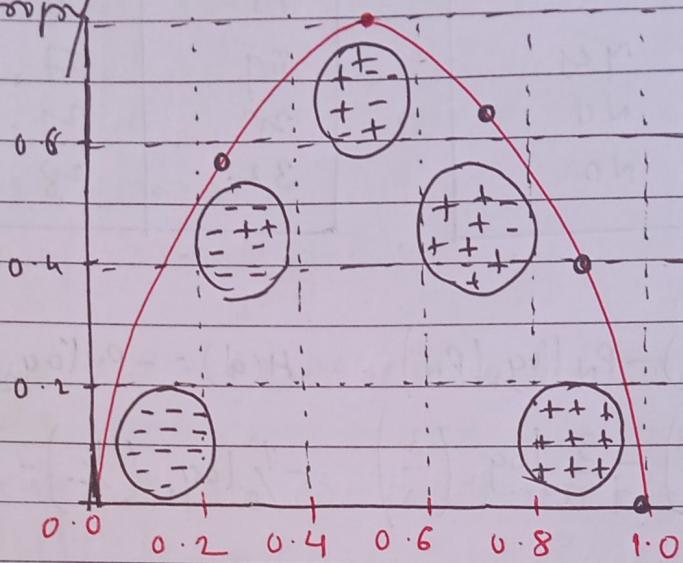


use for multiclass / 3 class problem

$$H(D) = -P_Y \log_2(P_Y) - P_N \log_2(P_N) - P_M \log_2(P_M)$$

Conclusion: More the uncertainty more the entropy.

- 2) For 2 class problem the min entropy is zero  
and max would be 1
- 3) More than 2 classes the min entropy is 0 but max can be greater than 1
- 3)  $\log_2$  or  $\log_e$  both can be used to calculate entropy

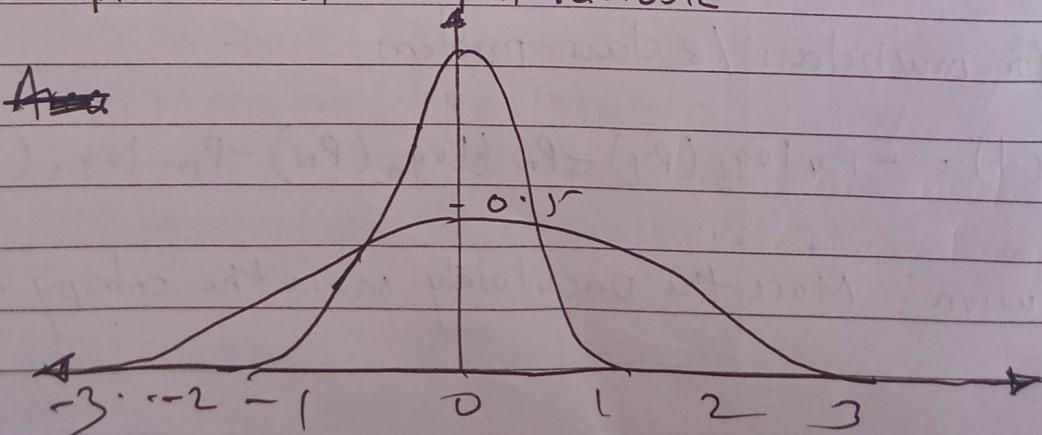


### Entropy for Continuous Variable

$x_1$	$x_2$	$y$
-	-	3.1
-	-	5.6
1	-	-
-	-	7.3

$x_1$	$x_2$	$y$
-	-	4.6
-	-	6.5
-	-	12.8

will plot for both output variable



whatever has less peak has higher entropy.

**Information Gain :-** is a metric used to train decision trees. Specifically, this metric measures the quality of split.

The information gain is based on entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns highest information gain.

$$\text{Information Gain} = E(\text{parent}) - \{\text{weight Avg}\} \times E(\text{child})$$

Consider a dataset

outlook	temp	Humidity	windy	play tennis
Sunny	Hot	High	false	No
Sunny	not	High	True	No
overcast	mild	High	F	Y
Rainy	cool	Normal	F	Y
Rainy	cool	Normal	T	N

shape. (14 x 5)

Step 1: Entropy of parent

$$= -P_Y \log(P_Y) - P_N \log(P_N)$$

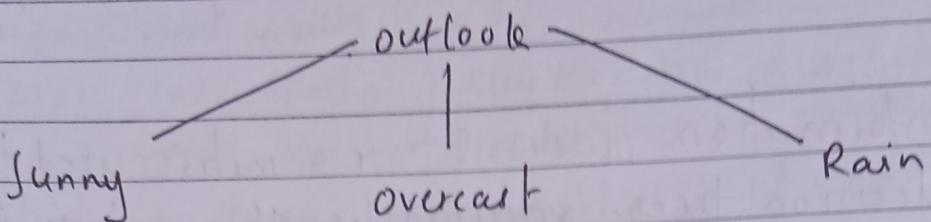
$$= -\frac{9}{14} \log\left(\frac{9}{14}\right) - \left(\frac{5}{14} \log\left(\frac{5}{14}\right)\right)$$

$$= 0.910.12 - (-0.79)$$

$$= 0.12 + 0.79$$

$$\boxed{\therefore = 0.91}$$

Now we will calculate entropy of children based on outlook



outlook	Humid	Wind	Play
Sunny		W	
overcast		N	
Rain		Y	
		Y	

$$E(S) = 0.97$$

$$E(OV) = 0$$

$$E(R) = 0.97$$

Step 3 :- calculate weighted entropy of children

$$\frac{9}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$$

$$WE(\text{children})_{\text{outlook}} = 0.69$$

Note: wherever entropy is 0 it is leaf node }

Step 4 :-

$$\text{Information Gain} := E(\text{Parent}) - [\text{wt Avg.}] \times E(\text{children})$$

$$0.97 - 0.69 \\ = 0.28$$

Interpretation goes like this

when you split the data based on outlook information gain or decrease in Entropy / Impurity is 0.28.

step 5: whichever column has higher information gain algorithm will select that column to split the data

step 6: find information gain recursively.

## Gini Impurity :-

formula for Entropy

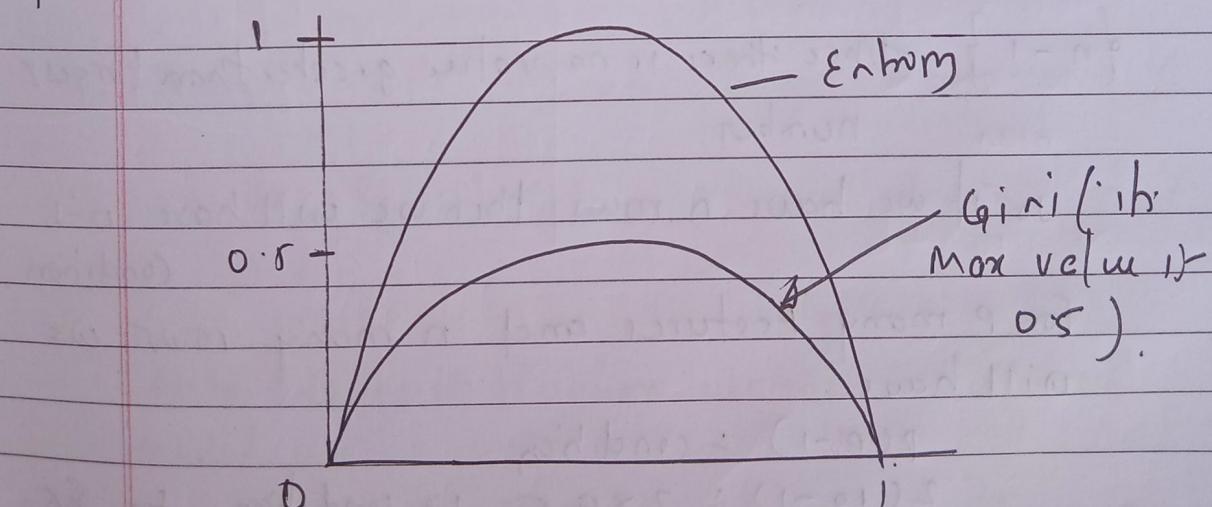
$$E = -P_Y \log(P_Y) - P_N \log(P_N)$$

formula for gini

$$G_2 = 1 - (P_Y^2 + P_N)^2$$

both are used for same purpose

	Gini	Entropy
$P_Y = 1; P_N = 0$	0	0
$P_Y = 0.5; P_N = 0.5$	0.5	1



till the time we were only considering categorical column what if input variable are numerical?

for let's see any continuous data.

$x_1$	$x_2$	$y$
1	8	0
2	9	0
3	7	0
:	4	0
:	5	0
10	10	0

$$\text{calculate gini impurity} \therefore 1 - (p_1^2 + p_2^2)^2$$

$$= 1 - p_1^2 - p_2^2$$

in such condition we have to test the model  
for different conditions for each column  $x_1$  &  $x_2$

$x_1 > 1$      $x_2 > 8$

$x_1 > 2$      $x_2 > 9$

⋮    ⋮

$x_1 > 9$      $x_2 > 9$

{ $n-1$ } since there is no value greater than largest number.

so if we have  $n$  rows then we will have  $n-1$  condition.

for  $p$  many Features and  $n$  many rows we will have

$p(n-1)$  = condition

$2(10-1) = 2 \times 9 = 18$  condition for Ex

Among all possible splitting condition we will select that split where test impurity information gain is highest, or less entropy.

## Cross Validation . (Kfold)

In k-fold cross validation, we split the dataset into k- Number of folds (subsets) One chunk of the data is used as test data for evaluation and remaining part of data is used for training of the model each time a different chunk will be used as test data.

Dataset	for k=5	Accuracy
Iteration 1	Train Train Train Train Test	88%
2	Train Train Train Test Train	83%
3	Train Train Test Train Train	86%
4	Train Test Train Train Train	81%
5	Test Train Train Train Train	84%

Note : for each iteration we will select new model .

$$\frac{88 + 83 + 86 + 81 + 84}{5} = 84.4\%$$

## Advantages :-

- 1) Best suitable for small dataset
- 2) Better for multiclass classification
- 3) more reliable
- 4) useful for model selection

## Hyperparameter Tuning.

- Gridsearch cv

- Randomized search cv.

### Parameter

Model

parameters

Hyperparameters

These are the parameters of the model that can be determined by training with training data. These can be considered as internal parameters.

weights  
bias

$$y = wX + b$$

Hyperparameters are the parameters whose value control the learning process. These are adjustable parameters used to obtain optimal model also called as external parameters.

- Learning rate in gradient descent

- Number of Epochs
- n-estimator -

optimum values mean, for that particular set of hyperparameters we will get highest result

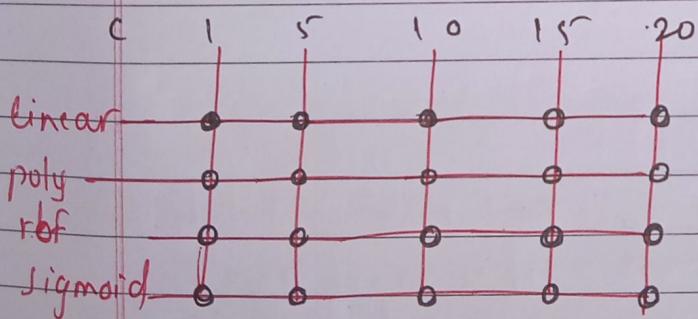
- we do model training → Best model parameter
- hyperparameter tuning → Best hyper parameters  
↓  
so that we can get highest accuracy or efficiency

### Hyperparameter tuning :-

hyperparameter tuning refers to the process of choosing the optimum set of hyperparameters for machine learning model this process is also called hyperparameter optimization.

### Hyperparameter tuning :-

- 1) GridSearch CV
- 2) Randomized search CV



for ex.

SVM

$c = [1, 5, 10, 15, 20]$

kernel = ['linear', 'rbf',  
'poly', 'sigmoid']

5 'c' values and 4 kernel

so in grid search cv we will try all 20 combination of c and kernel and among that whichever combination will give highest result will be finalized for best hyperparameter but in Randomized instead of trying all 20 we randomly choose few and among them whichever give highest result will

## finalize our best hyperparameters

when to use gridsearch cv :- Based on its way

- if working we can easily get to know it will take lot of time for large dataset so if we are time constraint then we should not use for large data for the small dataset it gives best is best choice
- for large dataset since we can't use gridsearch cv hence we would prefer to go with randomized search which can't low time although you might miss best hyperparameter since we have not tried all the combination.

## \* Descriptive statistics \*

### Descriptive statistics

#### A. Central tendency of data.

a) Mean

b) Median

c) Mode

#### B. Dispersion of data

a) Inter Quartile Range (IQR)

b) Range

c) Std deviation

d) Variance

#### C. Shape of data

a) Symmetric

b) Skewness

c) Kurtosis

There are two types of methods widely used for analyzing data

#### 1) Descriptive statistics

#### 2) Inferential statistics

It is used to summarize and describe the main features of the dataset such as central tendency, variability and distribution.

**A) Central tendency of data** :- This is nothing but center of distribution of data and concentrate where data is located. and three widely used measure are

a) mean    b) median    c) mode

a) Mean :- It is nothing but avg.

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Population mean -  $\mu$

Sample mean -  $\bar{x}$

{ outlier influence the central tendency of data }

↑  
Extreme behaviour:

Median :-

- Median is 50<sup>th</sup> percentile of data and it is exact centre point of data.
- arrange the data in ascending or descending order and split it into two halves and that's median.

Median is not affected by outliers  
for even number of data take two middle most numbers and take their mean.

Mode: most frequently occurred element

- if no number of element repeated in data then it don't have a mode

mode can be found out for both quantitative and Qualitative data

Dispersion :- it is nothing but spread of data

this dispersion can be measured by

a) standard deviation Var.

b) Variance

c) Range

d) IQR. (Boxplot)

a) standard deviation:- it is most common method to measure the spread.

• in simple terms - it measure of for the data deviates from mean value

• std dev formula varics for population & sample

sample std deviation = -

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \quad \text{this is deviation}$$

population std deviation =

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

std deviation is always positive or zero.  
it will be large when data values are spread out from the mean

Variance : it is measure of variability. it is average squared deviation from mean

$$\sigma^2 = \text{population Variance}$$

$$s^2 = \text{sample Variance}$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} \quad \text{population variance}$$

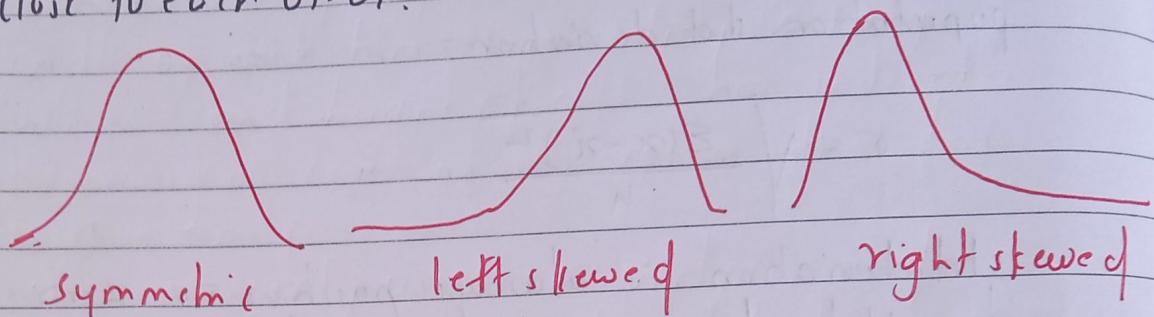
$$= \frac{\sum (x - \mu)^2}{n-1} \quad \text{sample variance}$$

3) shape of data :- shape describe the type of the graph. it is help to make decision of data making decision about the probability of data based on its shape

it is measured by three methodologies

- a) Symmetric
- b) Skewness
- c) Kurtosis

- In symmetric shape of the graph the data is distributed similar on both side.
- In symmetric both mean and median are very close to each other.



skewness :- it is measure of asymmetry in the distribution of data

- positively skewed
- negatively skewed

- Positively skewed :- here data values are clustered around the left side of distribution and right side is longer
- and mean and median will be greater than the mode

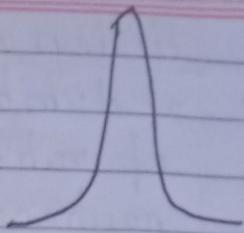
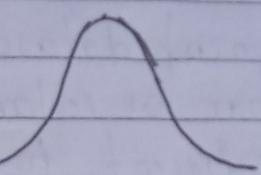
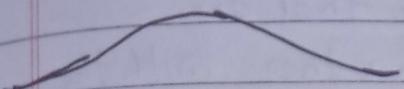
Negatively skewed :- here data values are clustered around right side of distribution and left side is longer  
mean and median will be less than mode

kurtosis :- it is also measure of describing the distribution of data.

- platykurtic
- mesokurtic
- leptokurtic

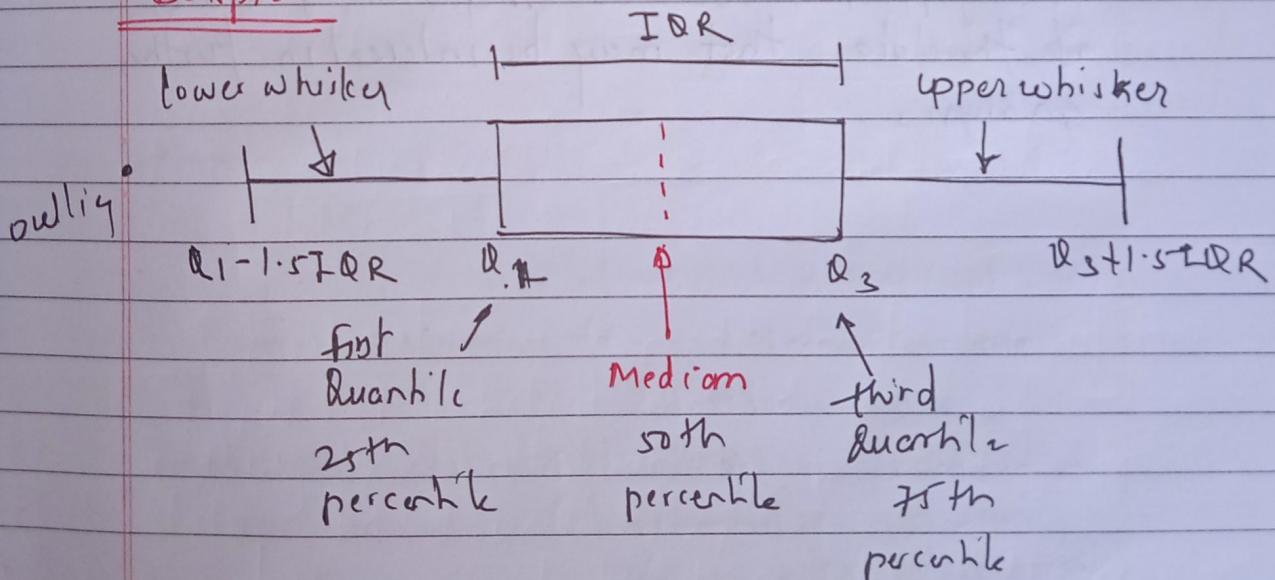
1. platykurtic

Mesokurtic



1. platykurtic : show the distribution with flat tails. Here data is distributed flatly, it indicates small outliers in the distribution.
2. Mesokurtic : it is normal distribution where data spread widely.
3. Leptokurtic : the data very closely distributed. The height of peak is greater than width of peak.

### Boxplot :-



Box plots are useful tool for summarizing and visualizing the distribution of the data and identifying any outliers. They are commonly used in EDA and in statistical inference to compare the distribution of data between groups.

Histogram :- it is graphical representation of distribution of dataset. It consists of series of vertical bar or rectangle that are positioned adjacent to each other with the width of each bar representing a range of values and height of each bar indicates frequency or count of observation that fall within range.

- it is commonly used in analysis and statistics to visualize the distribution of continuous variable such as age, weight, height by dividing range into equal interval or bins.
- Basically it is useful tool for visualizing the distribution of data and gaining insight into characteristics of dataset they can help identify pattern, outliers and other features of the data that may be involved in further analysis.

# Inferential statistics

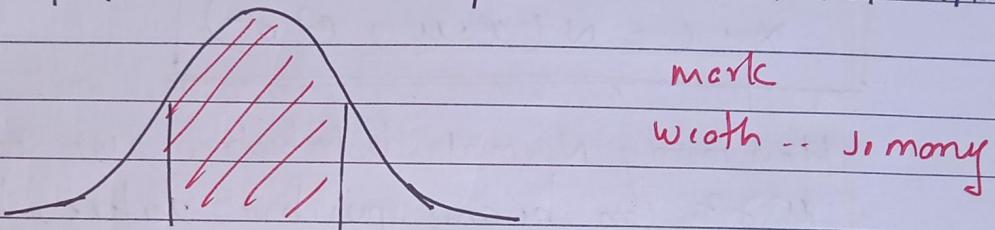
Date: / /

Page

- In inferential statistics we try to find some factor of population with the help of sample provided sample should be kind of random
- In inferential statistics large population mean is very very close to sample mean also variance and standard deviation

## Normal distribution

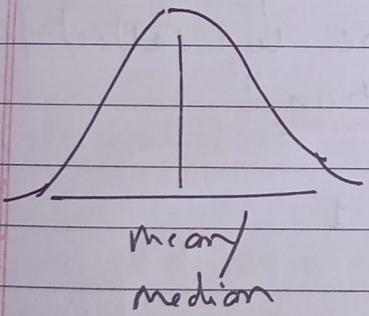
Majority of data tends to follow normal distribution.



Max value in middle

Normal distribution has two properties

1) mean & variance



Variance is nothing but spread of data if variance is more spread is less if variance is less spread is less

Total area under bell curve is 1

$$std = \sqrt{\text{variance}} \quad \therefore \sigma = \sqrt{\text{variance}}$$

$$\sigma^2 = \text{variance}$$

Normal distribution denoted by

$$n \rightarrow N(\mu, \sigma^2) \quad \text{for population}$$

$\mu$ : mean,  $\sigma^2$ : variance

what will be change in distribution for add or subtract any constant from mean.

$$x + s \rightarrow N(\mu + s, \sigma^2)$$

only mean will change with added constant & variance still remain same.

and for data multiply by any constant  
 For ex  $x \rightarrow (u, \sigma^2)$  &  $n=2x$  then

$$\text{mean} \rightarrow 2u$$

$$\sigma^2 \rightarrow 2^2 \sigma^2 = 4\sigma^2$$

Conclusion :-

$$x = N(u, \sigma^2)$$

$$x+c = N(u+c, \sigma^2)$$

$$x \cdot c = N(x \cdot u, c^2 \sigma^2)$$

How many Normal distribution we can have.

- $u, \sigma^2$  can be any positive value, hence we can have infinite number of normal distribution.
- So in order to put all these many normal distribution or convert in one we called it as "standard Normal distribution".
- and its significance is

$$\text{mean} = u = 0 \quad \sigma^2 = 1$$

how to convert ND  $\rightarrow$  std

$$ND : x \rightarrow N(u, \sigma^2)$$

$$x = N(0, 1)$$

to make  $u=0$  &  $\sigma^2=1$

$$x-u = N(u-u, \sigma^2)$$

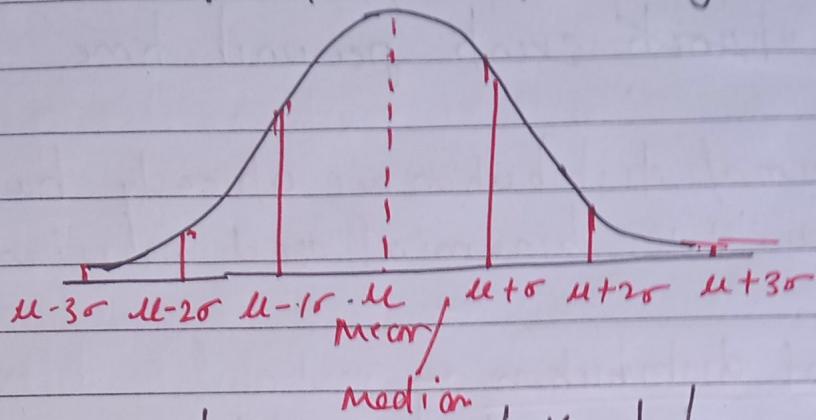
$$= N(0, \sigma^2)$$

$$\frac{1}{\sigma}(x-u) = N\left(0 \cdot \frac{1}{\sigma}, \frac{1^2}{\sigma^2} \cdot \sigma^2\right)$$

$$\frac{x-u}{\sigma} : N(0, 1) \Rightarrow \text{which is nothing}$$

$$\text{but} = \frac{\text{data-mean}}{\text{std.deviation}}$$

- Significance of std dev. :- measurement of dispersed data in relation with mean.
- low std dev data clustered closer around mean & high means spread out or far away from mean



$\mu \pm \sigma$  contains 68% of the data

$\mu \pm 2\sigma$  will contain 95% of the data

$\mu \pm 3\sigma$  will have 99.7% of the data

Normal distribution, binomial distribution and Poisson distribution are three important probability distribution in statistics and data analysis.

**Normal distribution** also called as gaussian distribution it is continuous probability distribution that is often used to describe a natural phenomenon such as height weight. it is bell shaped curve which symmetric & centered around the mean

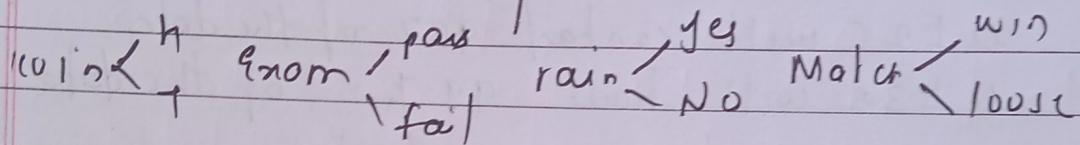
**The binomial distribution** it is discrete probability distribution that used to model number of success in fixed number of independent trial. It is characterized by two parameters - the probability of success in single trial number of trials

Poisson distribution :- it is discrete probability distribution that is used to model number of events that occur in fixed interval of time. It is characterized by single parameter - the avg number of such events per unit time.

Since Normal distribution we already have seen how we will see binomial and poisson in detail.

Binomial distribution :-

- in this we have only two outcome



in case of binary sum one will be zero and another will be one.

$X = 1$  = success

$X = 0$  = fail

$$P(X=x) = p^x (1-p)^{1-x} \quad \text{--- } ①$$

**event**  $\rightarrow$  this is called probability mass function since outcome is discrete value

{ Note: in case of outcome or continuous value it is called probability density function }

for  $n=0$

$$\begin{aligned} P(X=0) &= p^0 (1-p)^{1-0} \\ &= (1-p) = q \end{aligned}$$

for  $n=1$

$$\begin{aligned} P(X=1) &= p^1 (1-p)^{1-1} \\ &= p \end{aligned}$$

$$\text{Hence } P(\text{success}) = p$$

$$P(\text{failure}) = 1 - p = q$$

$$P(X) = \begin{cases} q = 1-p & \text{if } n=0 \\ p & \text{if } n=1 \end{cases}$$

Let try to understand with ex of fair coin

$$P(H) = P(\text{success}) = p = \frac{1}{2}$$

$$P(T) = 1 - p = \frac{1}{2} = \frac{1}{2}$$

There could be another scene

$$P(H) = 0.4 = p$$

$$P(T) = 1 - 0.4 = 0.6 = 1 - p.$$

How we will compute mean & variance and std deviation

To calculate mean

$$\Sigma(x) = \sum_{i=1}^n x \cdot P(x)$$

$$= 0 \times 0.4 + 1 \times 0.6$$

$$= 0.6 = p$$

$$= 1 \times 0.4 + 0 \times 0.6$$

$$= \underline{0.4} = p$$

$$\text{for } P(x=0) = 0.4 = p$$

$$P(x=1) = 1 - p = 0.6$$

$$P(x=1) = p$$

$$= 0.4$$

$$P(x=0) = 1 - p$$

$$= q = 0.6$$

Note for Bernoulli distribution your mean is equal to  $p$ .

Now let calculate variance =  $P(1-P) = pq$ .  
 $= SD = \sqrt{pq} = \underline{\text{standard deviation}}$

Note  $P > 0.5$  then  $P$  is median or else ' $q$ ' will be the median.

2. If  $p \neq q$ , ' $p$ ' will be mode or else ' $q$ ' will be the mode.

### Poisson distribution :-

- Poisson distribution is probability distribution that measures how many times that event is likely to occur within specified period of time.
- Poisson distribution is used to understand independent events that occur at constant rate within given interval of time.

Example of Poisson distribution.

- Number of accident occurring in a city from 6pm to 10pm
- Number of patient arriving in an emergency room between 10pm to 11pm

Inference of first statement :- in a day if there were 500 accident then what is probability of 100 accident in specific time period for an evening 6pm to 10pm

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

$x \rightarrow$  Number of times the event occurs.

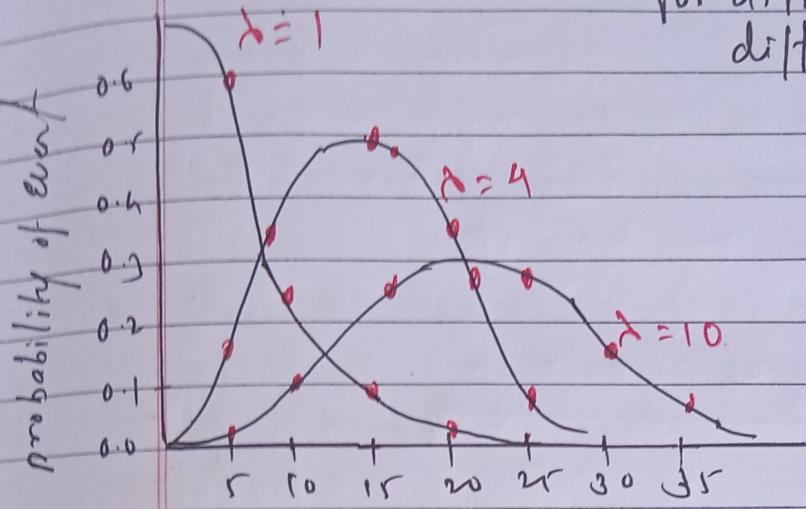
$P(x)$  Probability

$\lambda$  : Mean Number of Events

$n!$  = factorial of  $n$

$e$  = Euler's Number ( $2.71828$ )

for different value of  $\lambda$  we have  
different shape of curve



### Central limit theorem:-

- the central limit theorem says that the sampling distribution of mean will always be normally distributed as long as sample size is large enough regardless of whether the population has normal, Poisson, binomial or any other distribution. The sampling distribution of the mean will be normal.
- Assume you have Random Variable  $X$  which can have any distribution but  $X$  must have a finite mean and variance.
- Step 1 : you randomly pick sample of size  $n$  from  $X$  and you do it all total  $m$  times. At last you get  $m$  samples, each of size  $n$ .
- Step 2 : - we calculate the mean of each individual sample (size  $= n$ ) and end up with  $m$  sample means to be more clear now you  $m$  numbers and each of them represent mean of certain samples.

Step 3: then we plot distribution of m samples  
means and we are done.

# Probability.

Date : / /

Page

- Probability means possibility it is branch of mathematics that deals with occurrence of random event the value is express from 0 to 1 Probability has been introduced to math to predict how likely events are to happen,

$$= \frac{\text{No of favorable Event}}{\text{Total no. of outcomes}}$$

- There are two types of events
  - a) MEE (Mutually Exclusive Event)
  - b) Independent Event
- a) MEE : when A is happen B can't happen and when B is happen A can't happen  
Ex on dice we can either get even number or odd number.
- if any event is MEE then it follows following property.  
$$P(A \cup B) = P(A) + P(B)$$
- b) Independent Event :- if A is happen B can happen or can't happen. or outcome of one will not affect other.
- if any event is Independent Event then it follows following property.

$$P(A \cap B) = P(A) \cdot P(B)$$

## Hypothesis testing.

**hypothesis :-** it is an idea that is suggested as the possible explanation as the possible explanation for something but has not yet been found to true or correct.

**Testing of hypothesis :-** It means whenever somebody come up with assumption or claim we will do proper rigorous test statistical test before accepting or rejecting it.

There are two type of hypothesis

- |                   |                        |
|-------------------|------------------------|
| Null hypothesis   | Alternative hypothesis |
| • Existing Belief | • Somebody claiming    |

Example sun rises in south

$H_0$  : Sun rises in the east (Existing Belief)

$H_1$  : Sun rises in south (claim)

There are three types in testing of hypothesis

Type 1	Type 2	Type 3
$H_0: \mu = 60$	$H_0: \mu \leq 25$	$H_0: \mu > 3$
$H_1: \mu \neq 60$	$H_1: \mu > 25$	$H_1: \mu < 3$

how to type 1

duration of life before and after covid.

$$H_0: \mu = 60$$

$$H_1: \mu \neq 60$$

1) collect a sample data

$\{x_1, x_2, x_3, \dots, x_n\}$  — from population  $N(\mu, \sigma^2)$

if  $\sigma^2$  is known

for the sample

distribution of  $\bar{x} \rightarrow N(\mu, \sigma^2/n)$

to convert into standard normal.

$$\frac{\text{data-mean}}{\text{std dev}} = \frac{x - \bar{x}}{\sigma/\sqrt{n}} \quad \bar{x} - \mu \rightarrow N(0, 1)$$

Now  $Z$  is coming from std Normal distribution  $N(0, 1)$

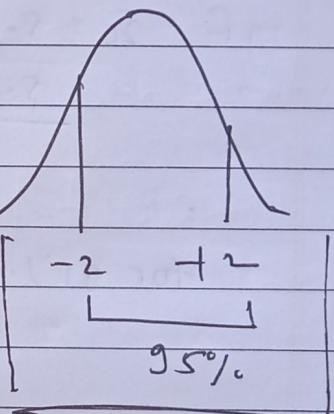
Since let assume  $H_0$  is true  $\mu = 60$

$$Z = \frac{\bar{x} - 60}{\sigma/\sqrt{n}}$$

$$Z = \frac{\bar{x} - 60}{\sigma/\sqrt{n}}$$

Assume

[95% chance if  $Z$  is coming  
from  $-2$  to  $+2$ ]



If calculated value of  $Z$  is in range of  $-2$  to  $+2$  Eq. 1.578

then will accept the null hypothesis  $\mu = 60$  for 95%. otherwise will reject the null hypothesis.

let solve one problem to understand more better

call duration avg =  $\mu = 4$   $\sigma = 3$ .

$$H_0: \mu = 4$$

$$H_1: \mu \neq 4$$

from  $\{x_1, x_2, x_3, \dots, x_n\}$  - sample

given  $\sigma = 3$   $\bar{x} = 4.6$   $n = 50$

$$\frac{\bar{x} - \mu}{\text{std dev}} = \frac{n/4.6}{3} = \frac{4.6 - 4.0}{3} = \frac{0.6}{3} = 0.2$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{4.6 - 4}{3/\sqrt{50}} = \frac{0.6 \times \sqrt{50}}{3}$$

$Z = 1.414$  which is between -2 to +2  
for 95% interval

will accept  $H_0$  and Reject  $H_1$   
if  $\bar{x} = 5.3$

$$= \frac{5.3 - 4.0}{3/\sqrt{50}} = \frac{0.3 \times \sqrt{50}}{3} = 3.06$$

for 95% and  $Z = 3.06$   $H_0$  will be rejected  
 $H_1$  will be accepted

### How to test type 2

$\{x_1, x_2, x_3, \dots, x_n\}$   $\rightarrow N(\mu, \sigma^2)$   
sample

$$\bar{x} - \mu$$

distribution of  $\bar{x}$  will be  $N(\mu, \sigma^2/n)$   
Now convert to std normal distribution.

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

upto this it will be same from type 1, 2, and 3.

$$H_0: \mu \leq 2r$$

$$H_1: \mu > 2r$$

for assuming null hypothesis is true

$$\mu \leq 2r \quad \text{multiply both sides by -1}$$

$$-\mu \geq -2r$$

$$\frac{\bar{x} - 4}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - 2r}{\sigma/\sqrt{n}}$$

$\downarrow$

$$z \geq \frac{\bar{x} - 2r}{\sigma/\sqrt{n}}$$

i)  $\bar{x} = 27$   $\sigma = 2$   $n = 100$

$$\frac{27 - 2r}{\sigma/\sqrt{n}} = \frac{27 - 2r}{2/\sqrt{100}} = \frac{27 - 2r}{2/10} = \frac{27 - 2r}{0.2} = 10$$

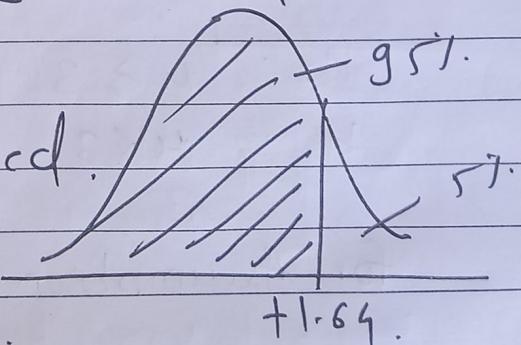
$$z \geq 10$$

but for 95%  $z \leq 1.64$ .

since anything greater is problematic

Since  $z$  is beyond 1.64

null hypothesis is rejected  
for considered assumption.



Solve the problem for test type 2

$$H_0: \mu \leq 120$$

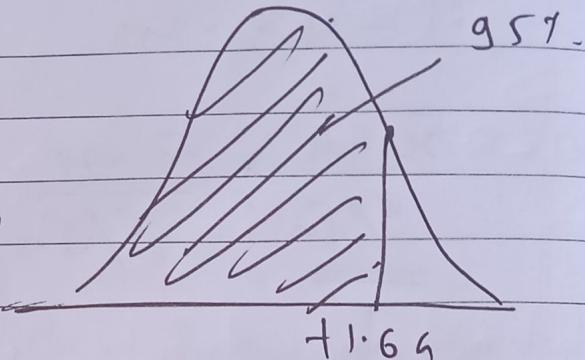
$$H_1: \mu > 120$$

$$n = 80 \quad \bar{x} = 130 \quad \sigma = 40$$

$$\mu \leq 120$$

$$-\mu > -120$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} > \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$$



$$z \geq \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$$

(anything greater is  
more plausible)

$$\frac{130 - 120}{40/\sqrt{80}} = \frac{10 \times \sqrt{80}}{40}$$

=

$$z \geq 2.23$$

and thus it is rejected. the null hypothesis.  
Since  $z$  supposed to be less than 1.64

To test type 3

$$H_0: \mu \geq 3$$

$$H_1: \mu < 3$$

$$\bar{x} = 2.9$$

$$\sigma = 0.2$$

$$n = 81$$

for Assumption  $H_0$  is true

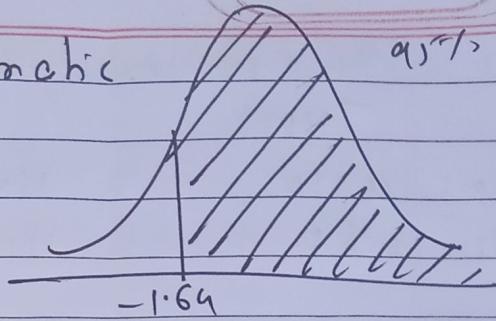
$$\mu \geq 3$$

$$-\mu \leq -3$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \frac{\bar{x} - 3}{\sigma/\sqrt{n}} \Rightarrow z \leq \frac{\bar{x} - 3}{\sigma/\sqrt{n}}$$

anything lesser is problematic

$$z \leq \frac{2.9 - 3}{0.2/\sqrt{8}}$$



$$\frac{-0.1 \times 9}{0.2} = \frac{-0.9}{0.2} = -4.5$$

thus null hypothesis is ~~is expected~~ since it is greater than -1.64  
lesser

### short Summary — $\sigma$ is known

Type 1

$$H_0: \mu = 10$$

$$H_1: \mu \neq 10$$

Type 2

$$H_0: \mu \leq 10$$

$$H_1: \mu > 10$$

Type 3

$$H_0: \mu \geq 10$$

$$H_1: \mu < 10$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

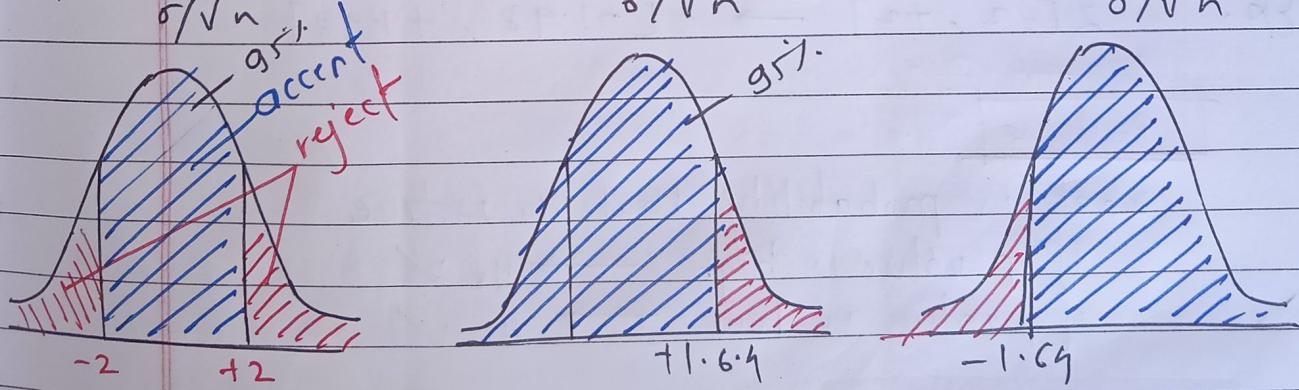
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$T_s$  = Test statistic (greater is problematic) (smaller is problematic)

$$T_s = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$T_s \leq \frac{\bar{x} - 10}{\sigma/\sqrt{n}}$$

$$T_s \geq \frac{\bar{x} - 10}{\sigma/\sqrt{n}}$$



anything in middle if  $T_s \leq 1.64$   
will accept it.

$$-2 < T_s < +2 \text{ NOV}$$

o/w reject and  
accept  $H_1$

$$H_0: \text{accpt}$$

$$H_1: \text{reject}$$

$$\text{if } T_s \geq -1.64$$

$$H_0: \text{accpt}$$

$$H_1: \text{reject}$$

if  $\sigma$  is unknown.

Type 1

$$H_0: \mu = 10$$

$$H_1: \mu \neq 0$$

Type 2

$$H_0: \mu \leq 10$$

$$H_1: \mu > 10$$

Type 3

$$H_0: \mu \geq 10$$

$$H_1: \mu < 10$$

Sample  $\frac{\bar{x} - \mu}{\text{std deviation}/\sqrt{n}} \rightarrow t_{n-1}$

this become  $t$ -distribution

$t_s = \frac{\bar{x} - 10}{s/\sqrt{n}}$

$t_s = \frac{\bar{x} - 10}{s/\sqrt{n}}$

$t_s = \frac{\bar{x} - 10}{s/\sqrt{n}}$

- As number of datapoint goes closer to population  $t$ -distribution becomes  $Z$ -distribution or std normal distribution. and 30 is threshold value. if  $n > 30$  then it follows  $Z$ -distribution.
- $t$ -distribution has big or thick tail and  $Z$ -dist has thin as i increase my number of sample thickness will be reduce.
- $t$  and  $Z$  distribution look alike but not same because value are different for  $Z$ -area

Ex.  $Z[-2, +2] \rightarrow t[-1.98, +1.98]$

### P-Value

what is probability that  $H_0$  is true

if p value is high  $\rightarrow H_0 \checkmark$

if p value is low  $\rightarrow H_1 \checkmark$

Sensitivity varies from domain to domain hence we can't boldly predict or make decision instead we can give probability for each hypothesis to b, but our client take the decision.

so "p-value" is nothing but probability of  $H_0$  or  $H_1$  is true

let see how p-value is calculated for type I?

$$H_0: \mu = 60$$

$$H_1: \mu \neq 60$$

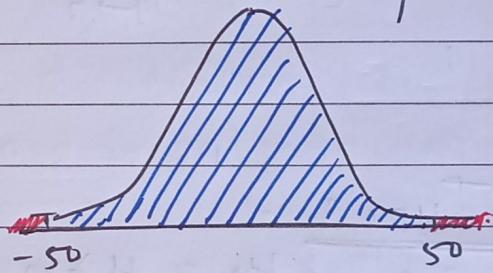
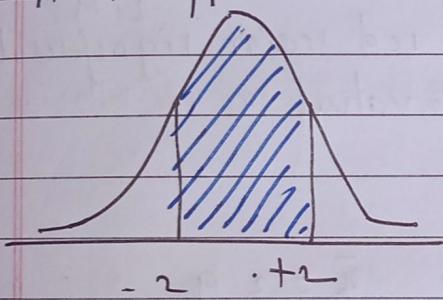
tatstistics

$$t \rightarrow \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\bar{x} = 70, n = 100, \sigma = 2$$

$$\frac{70 - 60}{2 / \sqrt{100}} = \frac{10 \times 10}{2 \times} = \underline{\underline{50}}$$

critical for 95%. we would have reject if far null hypothesis so p value must be very less.



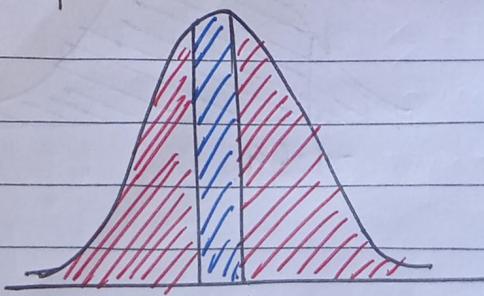
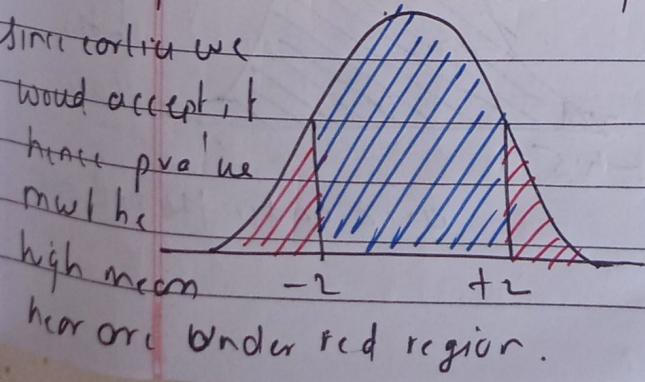
probability of  $H_0$  to be true is area under red region.

let check with acceptance

$$\text{if } \bar{x} = 60.005, n = 100, \sigma = 2$$

$$t \rightarrow \frac{60.005 - 60}{2 / \sqrt{100}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{60.005 - 60}{2 / \sqrt{100}} = \frac{0.005}{2 / \sqrt{100}} = \underline{\underline{0.025}}$$

for this critical we would accept if.



since critical we would accept, hence p-value must be high near mean hear or in under red region.

Let calculate p-value for type 2

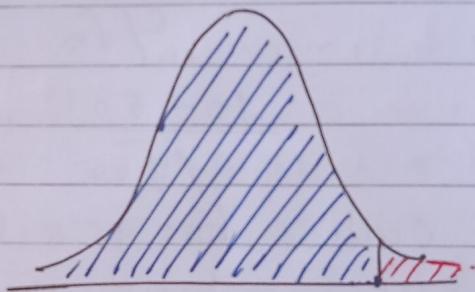
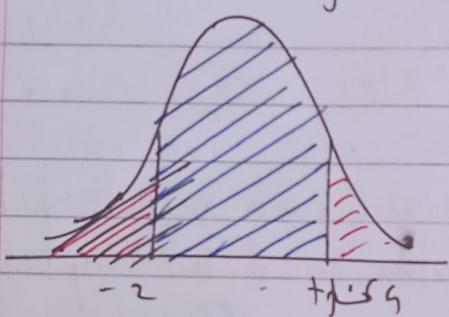
$$H_0 : \mu \leq 120$$

$$H_1 : \mu > 120$$

$$\bar{x} = 130, \sigma = 40, n = 80$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{130 - 120}{40/\sqrt{80}} = 2.23$$

Since, earlier we would reject it. hence p-value would be very less.



The red region is significant less  
p-value

Type 3 :-

$$H_0 : \mu \geq 3$$

$$H_1 : \mu < 3$$

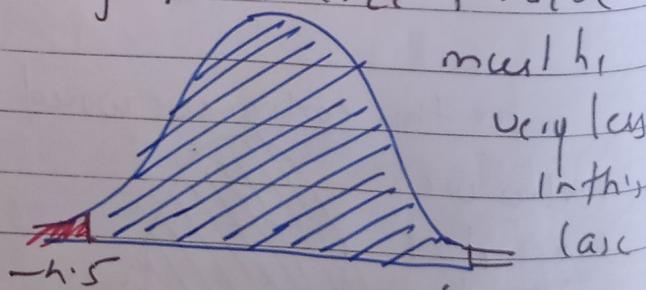
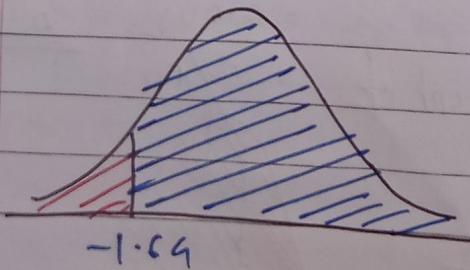
$$\bar{x} = 2.9$$

$$\sigma = 0.2$$

$$n = 81$$

$$t \rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{2.9 - 3}{0.2/\sqrt{81}} = \frac{-0.1}{0.2/9} = -4.5$$

earlier we would have rejected it hence p-value



much less  
very less  
in this  
case

area under the red one

Note :- type 1 Pvalue is on both extreme side  
 for type 2 pvalue is on right extreme side  
 for type 3 pvalue is on left extreme side.

## Significance level & Confidence level.

Significance level → it is probability with which we will reject null hypothesis when it is true in the significance level. or in our simple word it probability of rejecting null hypothesis when it is actually True

Confidence level : → it is probability of accepting null hypothesis when it is actually true

Significance level

$\alpha$

0.01

1%

0.05

5%

0.1

10%

Confidence level

$1 - \alpha$

0.99

99%

0.95

95%

0.90

90%

Level of significance

$\alpha$

0.01

0.05

1%

Corresponding confidence interval in term of z-value

-1.645σ to +1.645σ

-1.96σ to +1.96σ

-2.58σ to +2.58σ

## Confidence Level Interval

- it is applicable to continuous data.
- Since probability of continuous data is always zero for height, weight, temperature
- so whenever we have to estimate about continuous data we give interval with surely in terms of percentage
- Let try to understand with Example

$\{x_1, x_2, x_3, \dots, x_n\}$  ←  
(Sample)

Normal distribution  
population  
 $N(\mu, \sigma^2)$

To here we have to find  
mean height of men in Bangalore.

- we will calculate confidence interval for  $\mu$  which is avg ht of population.

population  $\mu \rightarrow$  sample  $\bar{x}$

$$= \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

$P(\text{sample mean}) = 0 \rightarrow [a, b] \text{ -- Interval}$

Step 1 if  $\sigma^2$  is known.

$$\underbrace{x_1, x_2, x_3, \dots, x_n}_{\text{sample}} \rightarrow N(\mu, \sigma^2)$$

population

whenever we consider data we don't consider individual data but mean.

distribution of population  $\rightarrow N(\mu, \sigma^2)$

distribution of sample drawn from population  $\rightarrow N(\mu, \frac{\sigma^2}{n})$

of mean will be same; variance will be reduced by factor of  $n$

Note: if each data coming from normal distribution  $N(\mu, \sigma^2)$  from same, if  $\bar{x}$  is going to come then

$$\bar{x} = N(\mu, \sigma^2/n) \text{ where } \bar{x} = \text{sample mean.} \quad (1)$$

Let eqn (1) convert it into standard Normal distribution

data - mean  
std dev

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

for 95% distribution of data it must be lie between  
-2 to +2

$$-2 < z < +2$$

$$-2 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < +2$$

$$\frac{-2\sigma}{\sqrt{n}} < \bar{x} - \mu < \frac{+2\sigma}{\sqrt{n}}$$

Multiply both sides by -1

$$\frac{2\sigma}{\sqrt{n}} > \mu - \bar{x} > -\frac{2\sigma}{\sqrt{n}}$$

$$\bar{x} + \frac{2\sigma}{\sqrt{n}} > \mu > \bar{x} - \frac{2\sigma}{\sqrt{n}}$$

So here interval for  $\mu$  at 95% confidence is

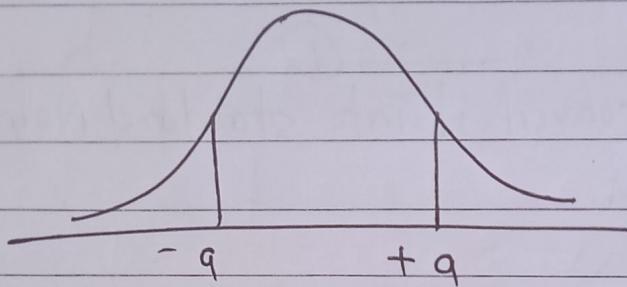
$$\mu \rightarrow \left[ \bar{x} + \frac{2\sigma}{\sqrt{n}}, \bar{x} - \frac{2\sigma}{\sqrt{n}} \right] - 95\%$$

$$\mu \rightarrow \left[ \bar{x} + \frac{\sigma}{\sqrt{n}}, \bar{x} - \frac{\sigma}{\sqrt{n}} \right] - 68\%$$

$$\mu \rightarrow \left[ \bar{x} + \frac{3\sigma}{\sqrt{n}}, \bar{x} - \frac{3\sigma}{\sqrt{n}} \right] - 99.7\%$$

Now general formula.

$$\left[ \bar{x} - \frac{\alpha \sigma}{\sqrt{n}}, \bar{x} + \frac{\alpha \sigma}{\sqrt{n}} \right]$$



for  $\sigma^2$  is unknown which is more realistic scenario

Since in real world nobody is going to tell us variance or  $\sigma^2$

- So when I don't know population variance we will calculate sample variance

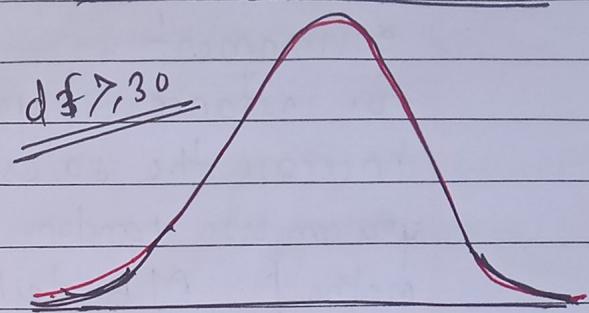
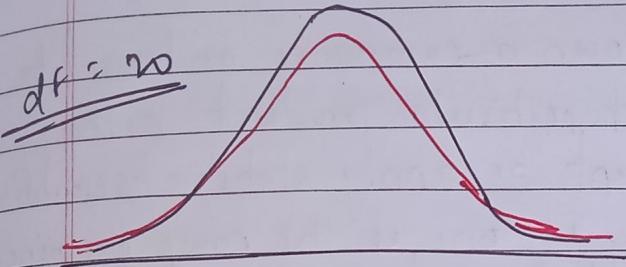
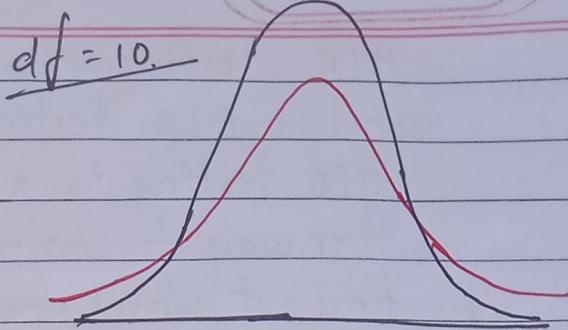
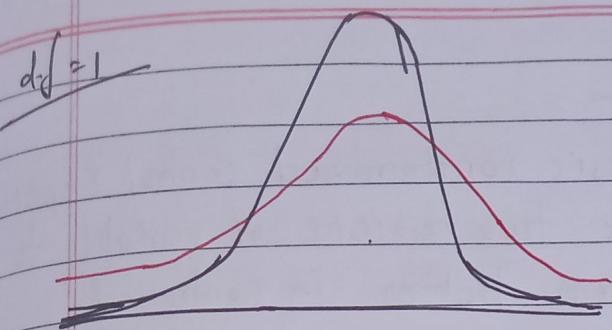
Population Variance  $\rightarrow$  Sample Variance  
 $\sigma^2$   $\rightarrow$   $s^2$

$\sigma^2$ known	$\sigma^2$ unknown
$x_1, x_2, x_3, \dots, x_n$	$x_1, x_2, x_3, \dots, x_n$
$\bar{x} \rightarrow N(\mu, \sigma^2/n)$	$\bar{x} \rightarrow N(\mu, \sigma^2/n)$
$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$	$\frac{\bar{x} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1}$

this is called t-distribution

- When sample standard deviation ( $s$ ) will become close to population standard deviation ( $\sigma$ ) t-distribution will become standard Normal distribution.

- t-distribution and Z-distribution look like.



as the degree of freedom increases t-distribution going closer to standard Normal distribution

degree of freedom is nothing but  $n-1$  (because only one variable is free).

### Summary:

$\sigma^2$  is known

$\sigma^2$  is unknown

$$\bar{x} \rightarrow N(\mu, \sigma^2/n)$$

$$\bar{x} \rightarrow N(\mu, \sigma^2/n)$$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$$

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightarrow t_{n-1}$$

if  $n$  is large  
z distribution

small  $n$

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightarrow N(0, 1)$$

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightarrow t_{n-1}$$

## A/B testing :-

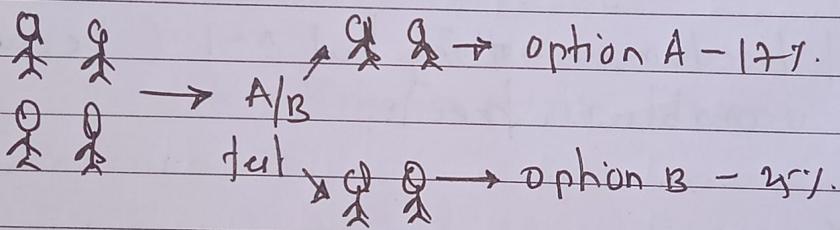
what is A/B testing :-

A/B testing is a basic randomized control experiment. It is a way to compare two versions of variable to find out which performs better in controlled environment.

For instance if you own a company and want to increase the sales of particular product either you can use random exp or apply statistic/scientific method. A/B testing is one of the most prominently and widely used statistical tools.

Ex option A : same old product

option B : updated old product.



Now based on this response we can take further decision.

## When we should use A/B testing :-

- A/B testing works best when testing incremental changes such as UX changes, new features, ranking and page load time. Here you may compare pre and post modification results to decide the changes are working as desired or not.
- A/B testing doesn't work well when testing major changes like new product, new branding or completely new user experience. In these cases there may be effect that drive higher than the normal engagement or emotional response that may cause users to behave in different manner.

## statistical test

Z-test :- it is statistical way of testing null hypothesis when either

- 1) we know population variance  $\sigma^2$  or
- 2) or we don't know population variance  $\sigma^2$  but our sample size is large than  $n > 30$ .
- if we have  $n < 30$  and don't know population variance (or) then we use t-test this is how we judge when to use Z-test or t-test.
- it assumed that Z-statistic follows standard normal distribution. In contrast t-statistic follows the t-distribution with degree of freedom equal to  $n - 1$  where  $n$  is sample size.

### One Sample Z-test :-

We perform one sample Z-test when we want to compare sample mean with population mean.

sample mean

$$Z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \begin{matrix} \text{population mean} \\ \text{deviation} \end{matrix}$$

Let try to understand with Example.

Ques: we need to determine if girls, on average score higher than 600 in the exam we have the information the std deviation of girl's score is 100, so we collect the data of 20 girls by using random sample and record their marks finally we also set our  $\alpha$  value (significance level) to be 0.05

Given sample mean score is  $\bar{x} = 641$ ,  $n = 20$ ,  $\mu = 600$   
 $\sigma = 100$ .

$$\text{Z-score: } \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{641 - 600}{100/\sqrt{20}} = \frac{41 \times 4.47}{100} = 1.8327$$

for it p value = 0.9664

$$1.8327 - 0.9664 =$$

$$1 - 0.9664 = \underline{\underline{0.0336}}$$

critical value is 1.645

Z score > critical value

p value < 0.05

we will reject null hypothesis

and accept alternative hypothesis

$$H_0: \bar{u} \leq 600 \times$$

$$H_1: \bar{u} > 600 \checkmark$$

since the p-value is less than 0.05 we can  
reject null hypothesis and conclude based on  
our result girl's avg score is higher than 600.

### Two sample Z-test

We perform two sample Z-test when we want  
to compare the mean of two samples

diff betn two sample mean  $\rightarrow$  difference between population mean

$$\text{Z-score: } \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad \begin{matrix} \downarrow \\ \text{population std deviation} \end{matrix} \quad \begin{matrix} \downarrow \\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{matrix} \quad \begin{matrix} \downarrow \\ \text{sample size} \end{matrix}$$

Let understand with example.

If we want to know if girls on avg score 10 marks more than boys we have info std dev ( $\sigma$ ) of girls is 100 and for boys is 90. Then we collect the data of 20 girls and 20 boys using random sample and record their marks. Finally we also set our  $\alpha$  value to be 0.05.

Solutions - Mean score for girls (sample mean) is  $\bar{x}_{\text{girls}} = 641$

- Mean score for boy (sample mean) is 613.3
- std deviation for the population of girls is 100
- std deviation for the population of boys is 90.
- sample size is 20 for both girls and boys.
- difference between mean of populations is 10.

$$\bar{z} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

type 2 problem

$$H_0: \mu_1 - \mu_2 \leq 10$$

$$H_1: \mu_1 - \mu_2 > 10$$

$$= \frac{(641 - 613.3) - (10)}{\sqrt{\frac{100^2}{20} + \frac{90^2}{20}}}$$

$$Z\text{-Score} = 0.588 < \text{critical value } 1.642$$

$$1 - 0.6844 = 0.3156$$

$$p\text{-value} > 0.05$$

thus we can conclude based on p-value that we fail to reject null hypothesis, we don't have enough evidence to conclude that girls on average score of 10 marks more than boys.

## What is the T-test?

T-test are statistical way of testing hypothesis when

- we don't know population variance.
- our sample size is less than 30.

## One-Sample T-test:-

We perform one sample t-test when we want to compare a sample mean with population mean. The difference from z-test is that we don't have information about population variance. We use sample std deviation instead of population std deviation in this case.

Sample

mean

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

population mean  
sample size

Sample std

deviation.

## Let understand with Example.

Let's say we want to determine if on average girls score more than 600 in the exam we don't have information related to variance (or std deviation) for girls' score. To a perfect t-test we randomly collect the data of 10 girls with their marks and choose our  $\alpha$  value (significance level) to be 0.05 for hypothesis testing.

In this Example

$$\text{Mean score for girl} = 606.8 = \bar{x} \quad | \text{type 2}$$

$$\text{size of sample} = 10$$

$$\text{population mean is} = 600.$$

$$\text{std deviation for sample} = \underline{13.14}$$

$H_0: \mu \leq 600$

$H_1: \mu > 600$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{606.8 - 600}{13.14/\sqrt{10}} = 1.54$$

for t-distribution critical value for 95% i.e.  
1.83

so t-score < critical value

(anything greater would be )  
problematic

$$p = 1 - 0.9495 = 0.0505$$

pvalue > 0.05

Since p-value > 0.05 we will reject the null hypothesis and don't have enough evidence to support hypothesis that on avg girl score more than 600 in the exam.

## Two sample - t-test

We perform a two sample t-test when we want to compare the mean of two samples.

diff betn  $\rightarrow$

sample  $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  Difference betn  
population mean

mean  $\bar{x}_1 - \bar{x}_2$

Sample size  $n_1, n_2$

standard deviation  $s_1, s_2$

Let try to understand with example

Let say if we want to determine if on average boy score is more than girl in the exam we don't have information related to variance (or std deviation) for girls and boy's score to perform t-test we randomly collect the data of 10 girls and boys with their marks we choose  $\alpha$ -value 0.05 (significance level) as criterion for hypothesis testing.

$$H_0 : \mu_1 - \mu_2 \leq 15$$

$$H_1 : \mu_1 - \mu_2 > 15$$

In this example.

$$\bar{x}_{\text{Boys}} = 630.1$$

$$\bar{x}_{\text{girls}} = 606.8$$

$$(\mu_1 - \mu_2) = 15$$

$$\sigma_{\text{Boys}} = 13.42$$

$$S_{\text{girls}} = 13.14$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}$$

$$= \frac{(630.1 - 606.8) - (15)}{\sqrt{\frac{(13.42)^2}{10} + \frac{(13.14)^2}{10}}}$$

$$= 2.23$$

$$\text{critical value} = 1.8333$$

$t$  exceed or the critical value

$$p\text{-value} = 1 - 0.9871$$

$$= 0.0129 < 0.05$$

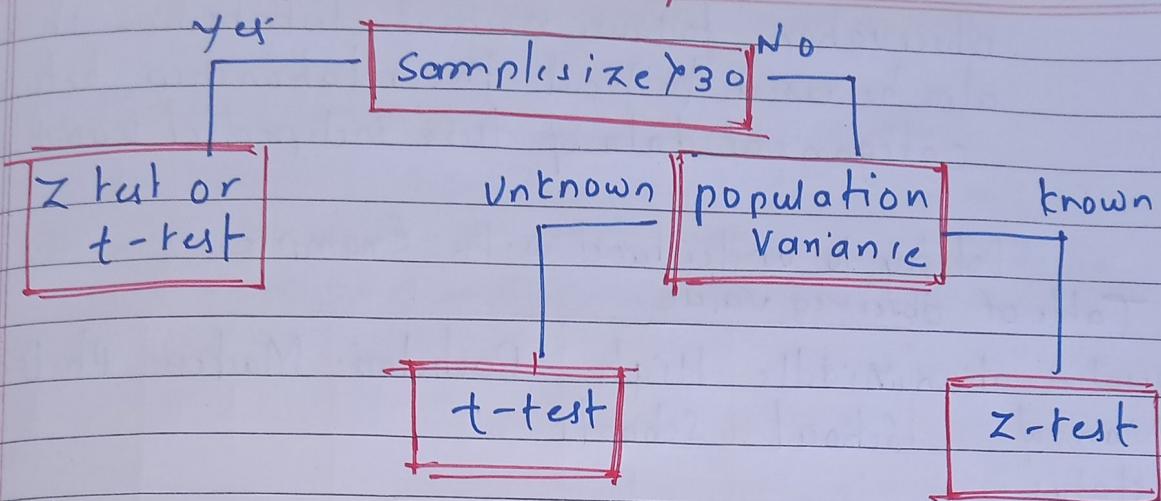
thus p-value is less than 0.05 we can reject the null hypothesis and alternate hypothesis

$$H_0 : \mu_1 - \mu_2 \leq 10 \quad \times$$

$$H_1 : \mu_1 - \mu_2 > 10 \quad \checkmark$$

and conclude that on average boys score 15 marks more than girls in exam

## Deciding between z-test / t-test



- i) if sample size is large enough the z-test or t-test will conclude with same result for large sample size. Sample variance will be better estimate of population variance so even if population variance is unknown we can use z-test using sample variance.
- similarly for large sample, we have degree of freedom and since t-distribution approaches the normal distribution the difference between the z-score and t-score is negligible.

## chi-square ( $\chi^2$ ) test

- chi-square test is hypothesis testing method used to test the hypothesis about the chi-square distribution of observation / frequency of different categories.
- How it is calculated. We first find the difference between observed (o) and expected (e) values then we take the square of that number and divide by expected value. Finally we add all of these values from various categories to get chi-square.
- What is chi-square used for :- It is used to predict probability of observation assuming null hypothesis is true. It is often used to determine if a set of

observation follows normal distribution it can also be used to find the relationship between categorical data of two independent variable.

Let's try to understand with Example

Table of observed value

Qualification	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Marital status						
Unmarried	18	36	21	9	6	90
Married	12	36	45	36	21	150
Divorced	6	9	9	3	3	30
Widowed	3	9	9	6	3	30
<u>Total.</u>	30	90	84	54	33	300

Aim:- Here our aim is to calculate or identify is there any relation between Education Qualification and Marital status

~~H<sub>0</sub>~~: there is no relation between two variable

H<sub>1</sub>: there is significant relation between two Variable  
Significant level ( $\alpha$ ) = 0.05.

Table of expected values.

Qualification	Middle School	High School	Bachelor's	Master's	Ph.D	Total
Marital status						
Unmarried	11.7	27	25.2	16.2	9.9	
Married	19.5	45	42	27	16.5	
Divorced	3.9	9	8.4	5.4	3.3	
Widowed	3.9	9	8.4	5.4	3.3	

Expected value ( $E$ ) =  $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$

Now our aim is to find how much deviation is there observed and expected value which is obtained by

$$\frac{(\text{observed} - \text{Expected})^2}{\text{Expected value}}$$

observed value ( $O$ )	Expected value ( $E$ )	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
19	11.7	6.3	39.69	3.39
36	27	9	81	3
21	25.2	-4.2	17.64	0.7
9	16.2	-7.2	51.84	3.2
.	.	.	.	.
.	.	.	.	.
3	5.3	-0.3	0.09	0.02
				$\sum \frac{(O-E)^2}{E}$

$$\chi^2 = 23.57$$

$\chi^2_{\text{calculated}} = 23.57$

We also need to compare value with tabular  $\chi^2$

For that we should to D.O.F

$$\begin{aligned} \text{D.O.F.} &= ((\text{col}-1)(\text{row}-1)) \\ &= (5-1)(4-1) = 4 \times 3 = 12 \end{aligned}$$

then from "percentage point of chisquare distribution" 0.5  $\rightarrow$  significance level.

$$\begin{array}{|c|c|} \hline \text{D.O.F.} \rightarrow 12 & 27.03 \\ \hline \end{array}$$

$$\chi^2_{\text{tabular}} = 21.03 \quad \chi^2_{\text{calculated}} = 23.57$$

$\chi^2_{\text{cal}} > \chi^2_{\text{tabular}}$  (or called as critical)

since it is greater than critical value our null hypothesis is rejected and alternative is accepted

H<sub>1</sub>: there is significant relation between two variable

### ANOVA : (Analysis of Variance)

Analysis of variance (ANOVA) is statistical technique used to check if the means of two groups are significantly from each other.

ANOVA check the impact to prove or disprove whether all the of one or more factor by comparing the means of different samples  
or in simple word :- it is test for statistical difference between two or more groups to check if there is any significant difference between the means of those groups.

Let try to understand with example.

standard deviation

$$S_n = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad \text{-- how data is spread.}$$

$$\text{Variance } \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

- we have three groups of student each contain 4 for whom we are going to conduct test in three different condition of Low Noise, medium noise and high noise.

Low Noise			Medium Noise			High Noise		
student	Que <sup>(x)</sup>	$x^2$	student	Que <sup>(x)</sup>	$x^2$	student	Que <sup>(x)</sup>	$x^2$
1	10	100	5	8	64	9	4	16
2	9	81	6	4	16	10	3	9
3	6	36	7	6	36	11	6	36
4	7	49	8	7	49	12	4	16
Total.	$\Sigma x_i = 32$	$\Sigma x_i^2 = 266$		$\Sigma x_i = 25$	$\Sigma x_i^2 = 165$		$\Sigma x_i = 17$	$\Sigma x_i^2 = 77$

By the observation we can know there is negative correlation between student question solving ability and noise but by the ANOVA we will get to know the percentage difference of correlation do we have between them.

$$\Sigma x_i = 32 \quad \Sigma x_i^2 = 266 \quad \Sigma x_i = 25 \quad \Sigma x_i^2 = 165 \quad \Sigma x_i = 17 \quad \Sigma x_i^2 = 77$$

$$\text{Correction term} = c_n = \frac{(\Sigma x_i)^2}{N} = \frac{(32+25+17)^2}{12} = \frac{5476}{12} = 456.33$$

Sum of squares of total :

$$SST = \Sigma x^2 - c_n = \frac{(266+165+77)^2}{12} - 456.33 = 508 - 456.33 = 51.67$$

Sum of squares Among group :

$$SSA = \frac{(\Sigma x_i)^2}{n} - c_n = \left( \frac{32^2}{4} + \frac{25^2}{4} + \frac{17^2}{4} \right) - 456.33 = 484.5 - 456.33 = 28.17$$

Sum of squares within group

$$SSW = SST - SSA = 51.67 - 28.17 = 23.5$$

Mean of sum of square among group.

$$\text{Df for among group} \quad \text{MSSA} = \frac{SSA}{K-1} = \frac{28.17}{3-1} = 14.085$$

$$\text{Df for within group} \quad \text{MSSW} = \frac{SSW}{N-K} = \frac{23.5}{12-3} = 2.611$$

No. of student category

$$F\text{-Ratio} = \frac{\text{MSSA}}{\text{MSSW}} = \frac{14.085}{2.611} = 5.394$$

Summarize ..

Source of Variance	df	ss	MSS	F-Ration
Among Groups	$(K-1) = 2$	28.17	14.085	5.394
Within groups	$(N-K) = 9$	23.5	2.611	

Null hypothesis :

$H_0$ : No significant effect of Noise on number of question solved.

$H_1$ : Significant effect of Noise on number of question solved.

from table from F-Ratio

$\downarrow$

$$q \rightarrow 4.2565$$

We have to compare this with F-Ratio

calculated  $F > F$  (table  $\alpha = 0.05$ )

$$5.394 > 4.2565$$

that's why we will Reject the null hypothesis and accept alternative hypothesis.

Since  $\alpha = 0.05$  means 5%.

Interpretation Hence with 95% confidence, we can say that there is significant effect of noise on number of question solved.

and if for  $\alpha = 0.01$   $\gamma\gamma$ .

$$F \text{ value} = 8.0215$$

Calculated  $F < F$  (table  $\alpha = 0.01$ )

$$5.394 < 8.0215$$

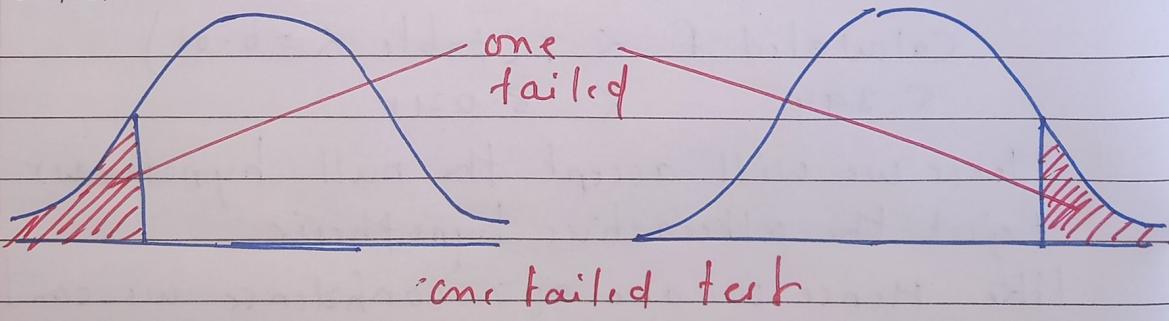
Hence we will accept the null hypothesis and Reject the alternative hypothesis

like. Hence we are 99% confidence we can say that there is significant effect of Noise on number of question solved.

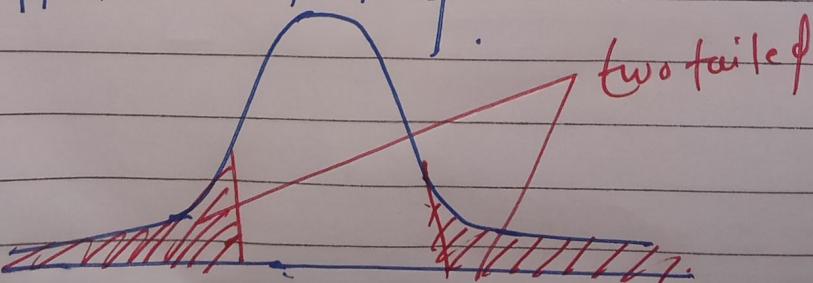
## Difference between One tailed and Two tailed test.

One tail and two tailed test are always identify relationship between statistical variable. For checking the relationship between variables in singl. direction (left or right) we use one-tailed test - A two tailed test is used for checking whether the relation between variables in any direction or not.

One tailed test :- one tailed test is based on unidirectional hypothesis where area of rejection is only on one side of sampling distribution. It determines whether the particular population parameter is larger or smaller than the predicted parameter it uses one single value critical value to test the data.



Two tailed test :- it is also called as non-directional hypothesis for checking sample is greater or less than a range of value we used two tailed. It is used for null hypothesis testing.



## Difference between one and two-tailed test

### One-tailed test

- A test of any statistical hypothesis where  $H_1$  is one direction anything greater or smaller.
- Sign we used for one-tailed.  $>$  or  $<$  for the alternative hypothesis.
- Critical region lies entirely on either left side or right side of sampling distribution.
- Half entire level of significance ( $\alpha$ ) i.e.  $\alpha$  is either in the left or right tail.

### Two-tailed test

- A test of hypothesis that where  $H_1$  is non-directional.
- Sign we used is  $\neq$  for alternative hypothesis.
- Critical region is given by the portion of area lying on both the tails of sampling distribution.
- It splits level of significance ( $\alpha$ ) into half.