Data Science Project Pipeline.

EDA

- EDA stands for Exploratory data Analysis. it refers to the process of Examining and analyzing dataset to summarize their main characteristics and gain insight into the data
- it typically involves the use of statistical graphics and data visualization technique to help identify pattern, relationship and anomalies in the data.
- the main objective of EDA is to get better understanding of the data and to uncover any hidden insight or relationship that may not be immediately appearant which can help in formulating hypothesis and developing predictive models as wellas 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 to to provide overview of the dataset and to help identify pattern or trends that may be present which can help better understand data and to make informed decision about how to analyze it further.
- there are number of techniques can be used to describe data. like tabular summaries, 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 sumarize 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 indicate 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 methode such as pearson correlation coefficient, spearman's rank correlation coefficient.

Covariance :-

it is measure of joint variability between two variables a positive co-variance indicates both vary in some direction whereas negative covariance tend to vary in opposite direction, unlike correlation. covariance is not normalized so it can be difficult to interprete the magnitude of value.

VIF (Vaiiance inflation factor) :-

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

In summary Correlation and covariance are measure of relationship bet^n two variables while VIF measure is used to detect multicollinearity in regression analysis.

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

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

Date: / /
Page

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
- Example of univariate analysis technique 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>'])

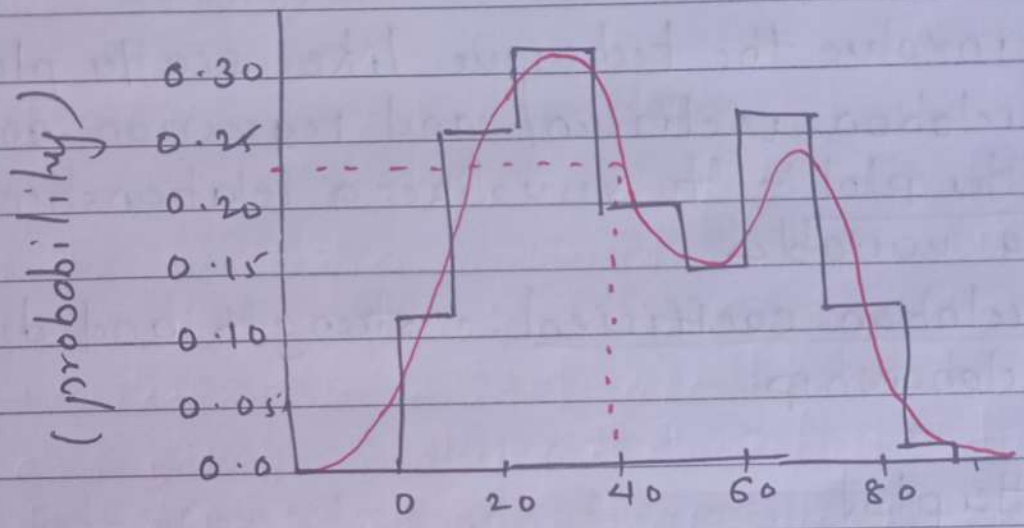
- Pie chart
  fig, ax = plt.subplots()
  ax.pie(sizes, labels = labels, autopct = '%.1.1F%%')

for Numerical data

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

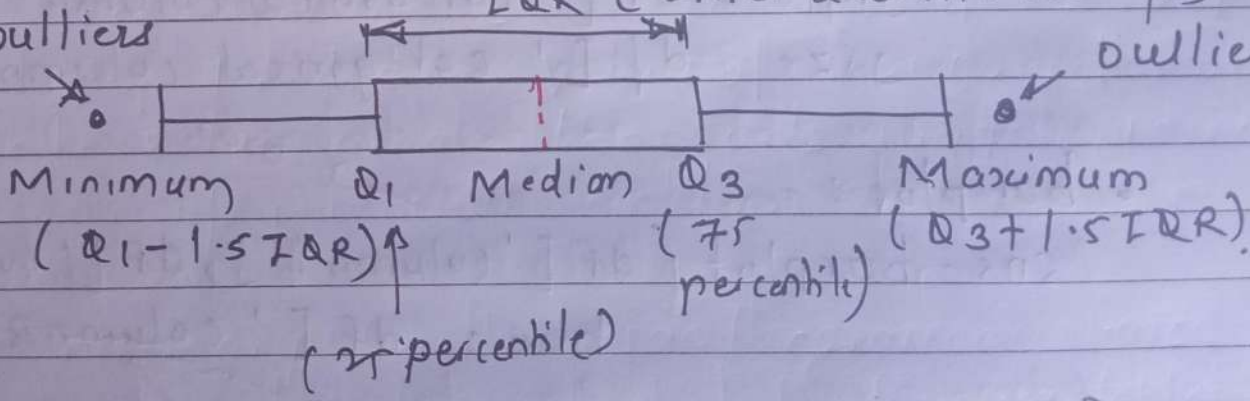
plt.hist(df['<age column>'])

b. Distplot:
it is almost similar to histogram only difference is curved line which called as kde (kernal density estimation) or pdf (probability density function)



probability of $x=40$ is between 0.20 to 0.25

c. Box plot:- it gives five number summary



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

Date: / /
Page

Bivariate Analysis :-

- it is analysis of relationship between two variables. for examining the relationship beth 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['column'], df['column'],
                (x-axis)      (y-axis)
    , hue = df['categorical column'],
    style = df['categorical column'],
    size = df['categorical column'].
```

- bar plot :-

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

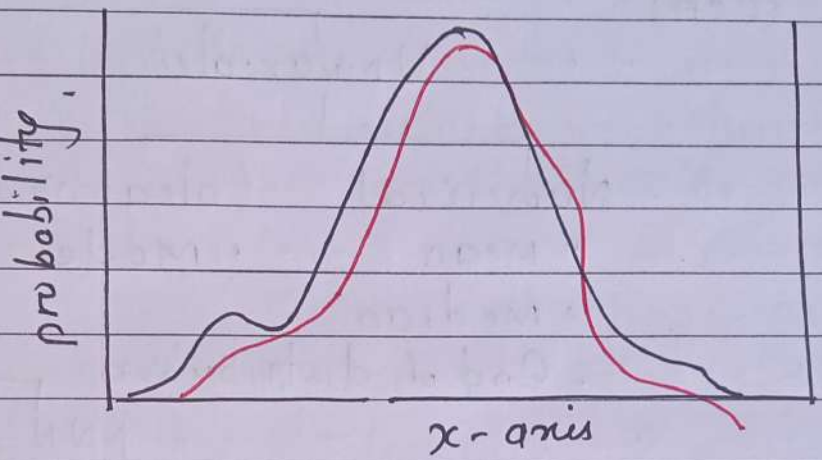
x-axis, y-axis, Categorical

- 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['column2']==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')
```

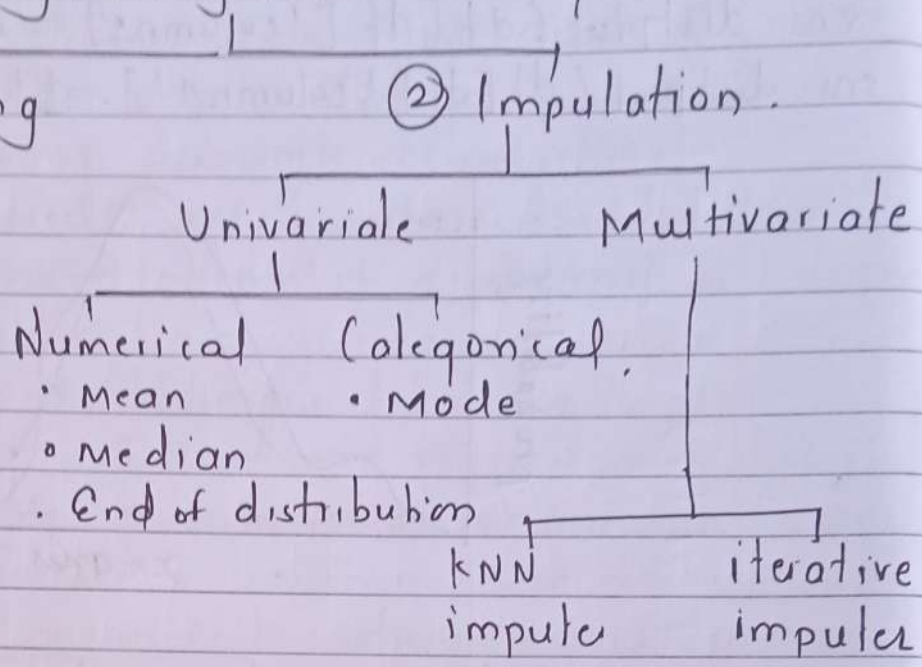
7. Lineplot - (Numerical - Numerical)

```
sns.lineplot(df['column1'], df['column2'
```

Date: / /
Page

Missing Value treatment

Handling Missing Values technique



① Dropping or Removing.

• Complele Case Analysis (CCA) also called as list wise deletion it consist of discarding (rows) Observation where value in any of the (column) Variable are missing

• it means literally allowing only those observation for which there is information in all of variable in the dataset

missing

Assumption :- data should be randomly distributed.

Advantages :- Easy to implement
2) preserve variable distribution

Disadvantages :- 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.
0. univariate → 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.
- outlier emerges.
- Effect on covariance/Correlation.

when to use:- when data completely
① missing at Random.
② missing value < 5%

Date: / /
Page

Arbitrary value impulation.
filling missing value with any arbitrary number
Numerical ⟶ 99, -1,
Categorical ⟶ -1, 0, 999.
it creat difference in distribution of dataset

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

End of distribution Imputation :-
it like arbitrary value impulation but diff is instead of putting any value we put extreme value of dataset

Normally distributed
mean + 3σ
or mean - 3σ

for skewed.
$Q_1 - 1.5\,IQR$
$Q_3 + 1.5\,IQR$
where $IQR = Q_3 - Q_1$
and Q_3 = 75 percentile
missing. Q_1 = 25 percentile

when to use :- when data is not randomly distributed

univariate → categorical.

Handling missing categorical data.

Most frequent category	or create new category "missing"
• Mode. • simple to implement	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 fill the missing value with any random number from some column.

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

Missing indicater.

- for the column which has missing values we create new column and whenever 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 differentiate between missing and non missing value and could improve accuracy.

Automatically select value for impulation
GridSearchCV - this is available in scikit learn which itself try all the method and give best suitable method and parameter.

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_1			-		
R_2	-				
R_3			-		

- 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.

Outlier 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.
- its 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 outlier.

- Effect of outlier in ML Algorithm
  outlier has very significant effect in weight based algorithm like
  1) Linear regression 2) Logistic Regression
  3) Adaboost 4) Deep learning
  whereas it is not much harmful for tree based algorithm
  1) Decision Tree 2) Random forest
  3) XGBoost 4) Bagging & Boosting.

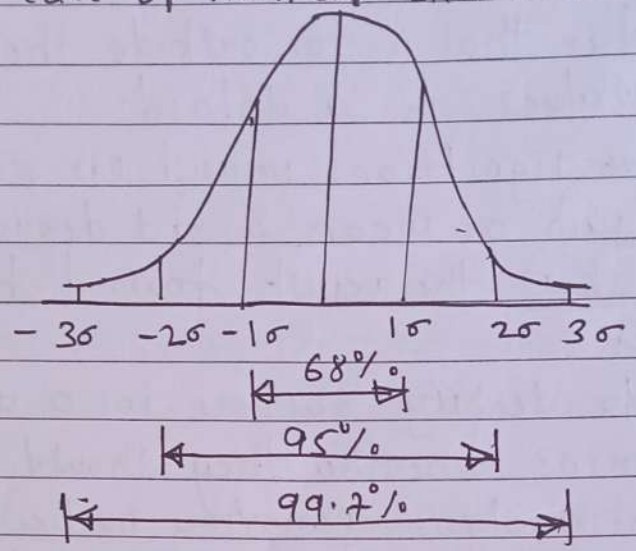
- How to treat outliers.

  Trimming | Capping

- makes data thin
- Fast

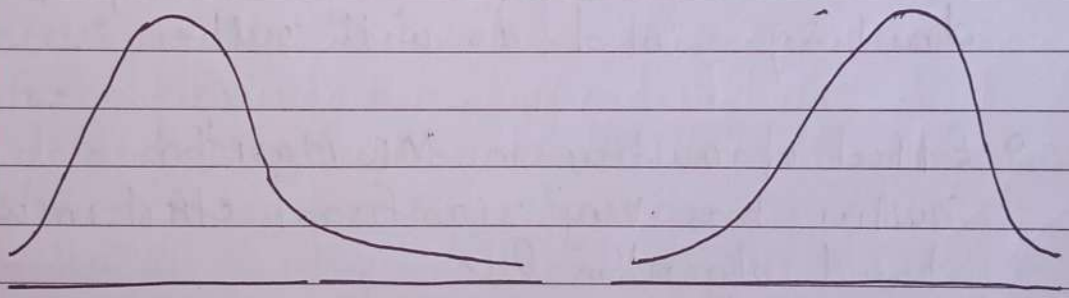
How to detect outliers ?

1). In case of normal distribution.

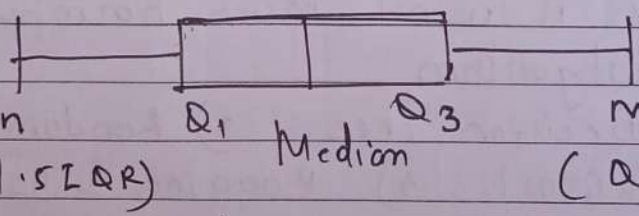


if $(\mu - 3\sigma) < x < (\mu + 3\sigma)$ then it is outlier

2) if the data is skewed distribution.

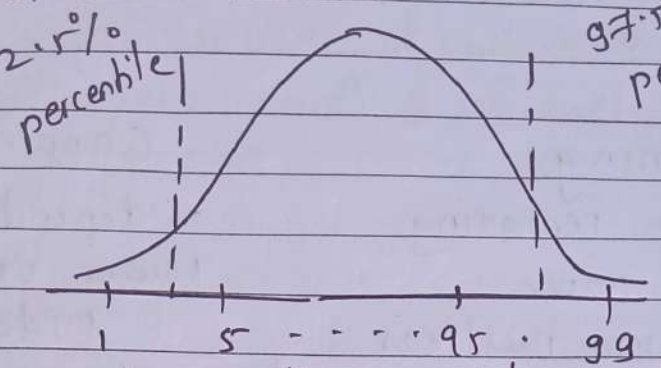


We can use : box plot



Anything less than min and greater than max is outlier.

In other distribution



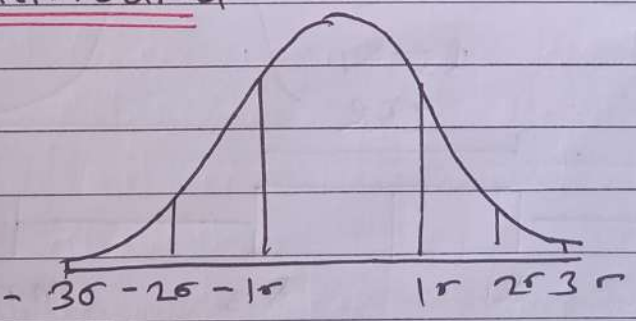
beyond 2.5th percentile and 97.5th percentile ir outlier.

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.



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

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

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

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

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

this is way of bringing any data distribution under standard normal

Date: / /
Page

outlier treatment

Trimming
- directly removing those rows containing outliers

Capping
- depending upon lower or higher side. we cap the value with lowest/highest restricted value

Ex. -3σ 3σ

8 8

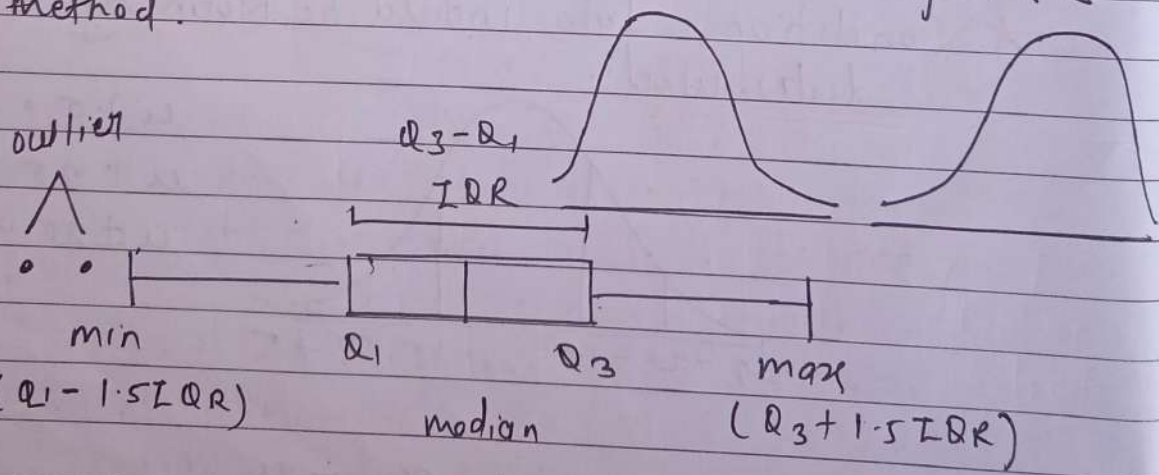
1 → 5

85 → 80.

Np. 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 IQR method.



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

Date: / /
Page

Summary:
if data is normally distributed then we can convert it into standard normal distribution by

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

- this is nothing but Z-score from it we can find upper limit and lower limit

lower limit : $\mu - 3\sigma$
upper limit : $\mu + 3\sigma$
} beyond this range anything is outlier, so 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.

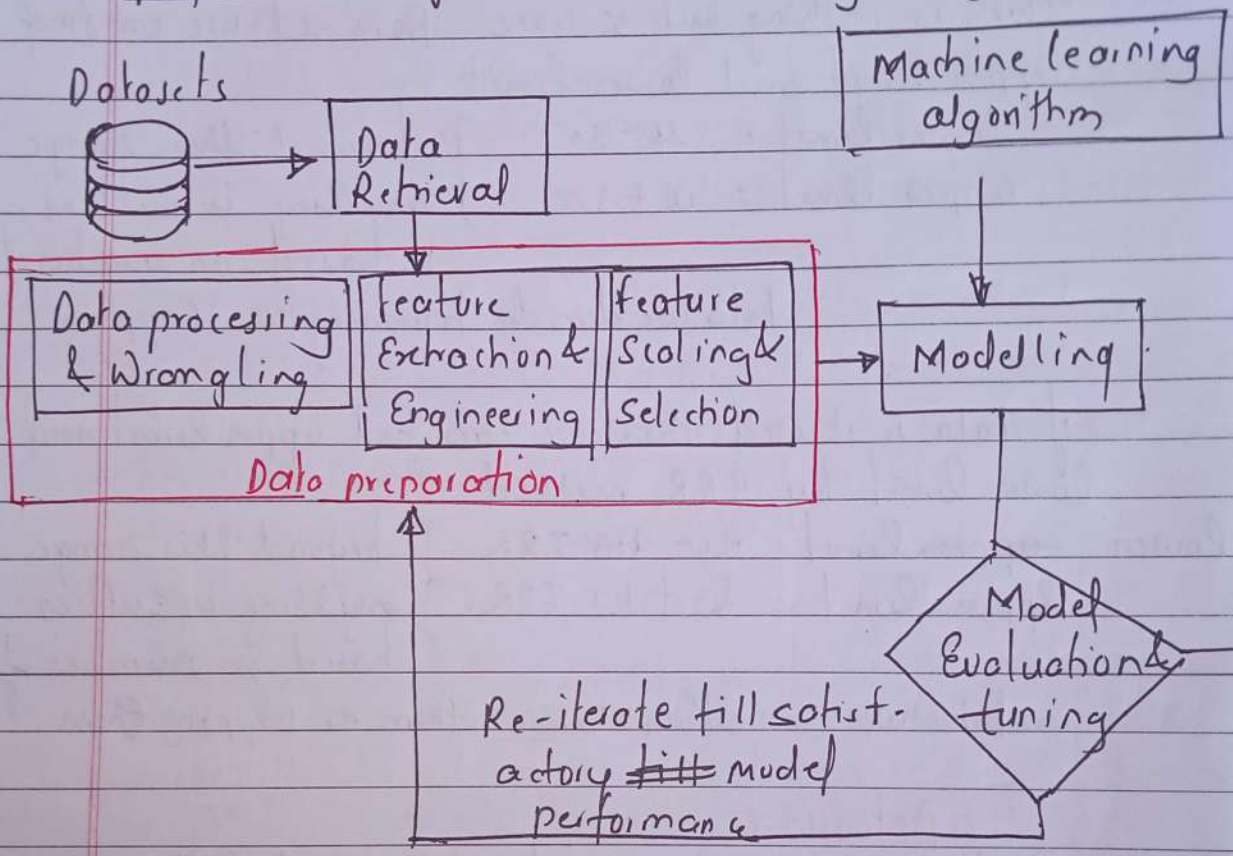
Lower ~~upper~~ limit : $Q_1 - 1.5 \text{ IQR}$
upper limit : $Q_3 + 1.5 \text{ IQR}$.
} beyond this range anything is outlier based on number of data we can either remove them or capping them.

Outlier Detection Using Percentile Method (Winsorization Technique)

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

Feature Engineering

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



Feature Engineering

- Feature Transformation
  - missing value impulation
  - Handling categorical features
  - outlier detection
  - Feature Scaling.
- feature Construction
- feature Selection
- feature Extraction

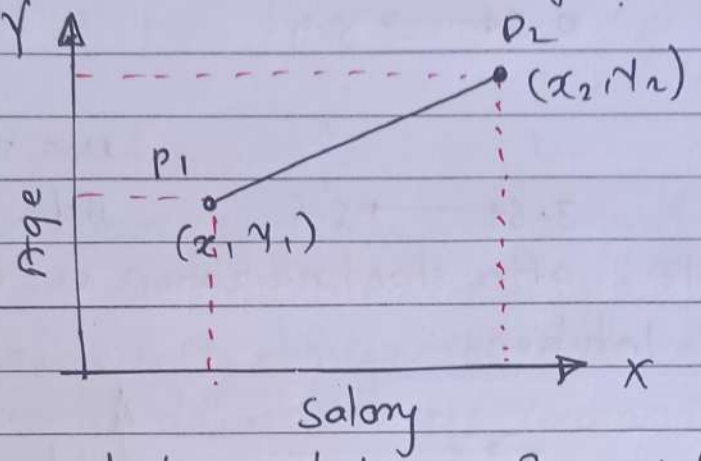
Feature Scaling

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

for ex

IQ.	GGPA	LPA
(10-100)	(1-10)	(5LPA - 20LPA)

why we do need feature scaling ?



Deployment & Monitoring.

Euclidian distance between P_1 and P_2

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

$(x_2-x_1)^2 = (83000-48000)^2$ } this will dominate

$= 1225000000$

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

$= 529$

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

Types of feature scaling

1. Standardization
2. Normalization
   - Min-Max scaler
   - Robust scaler.
   - mean normalization
   - Max absolute scaling.

standardization

Also called as Z-score Normalization.

$$x_i \longrightarrow x_i' = \frac{x_i - \mu}{\sigma} = \frac{\text{data} - \text{mean}}{\text{std deviation}}$$

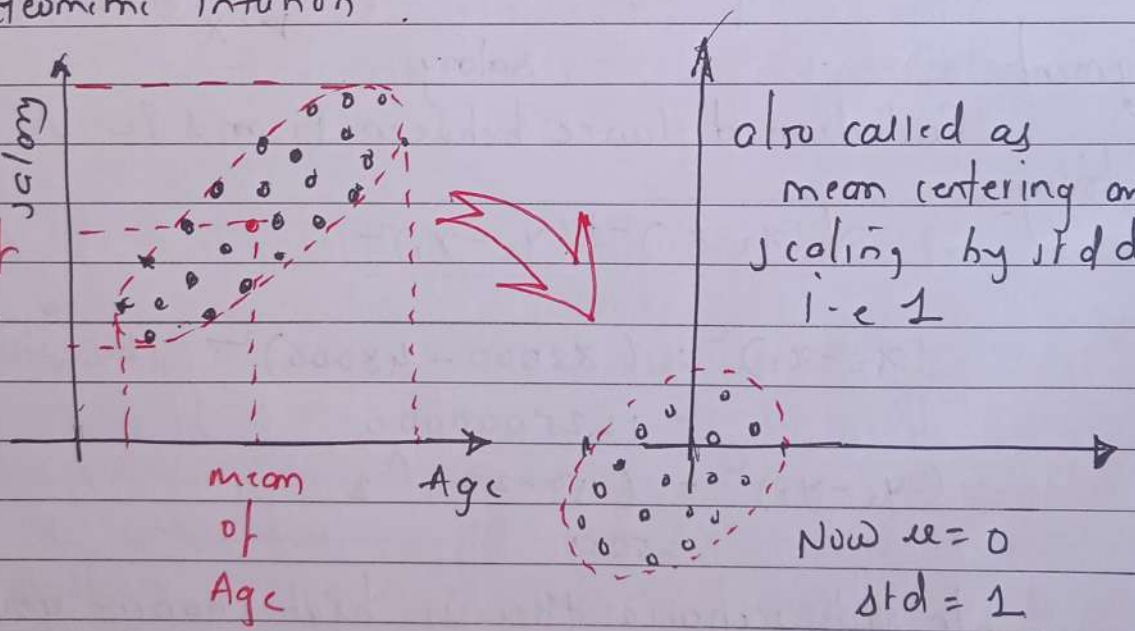
Salary	Age	Age'	Salary'
38000	27	-0.5	2.4
40000	35	0.3	2
23000	38	0.6	5.4
⋮	⋮	⋮	⋮
78000	58	3.6	5.6

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

$\mu = 50000$, std $= 5000$

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

Geometric Intution :



when to use standardization ?

	Algorithm	Reason for applying feature scaling
(1)	k-means	use Euclidian distance to measure
(2)	k-Nearest Neighbor	Measure the distance between pair of samples and these distance are influenced measurment units.

③ principle component Analysis (PCA) — Try to get feature with max variance

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 loosing information.

there are four types

① Min max scaling
② Mean Normalization
③ Maxabs scaling
④ Robust scaling

Min-Max Scaling - Intuition

weight
130
67
81
92
⋮
108

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

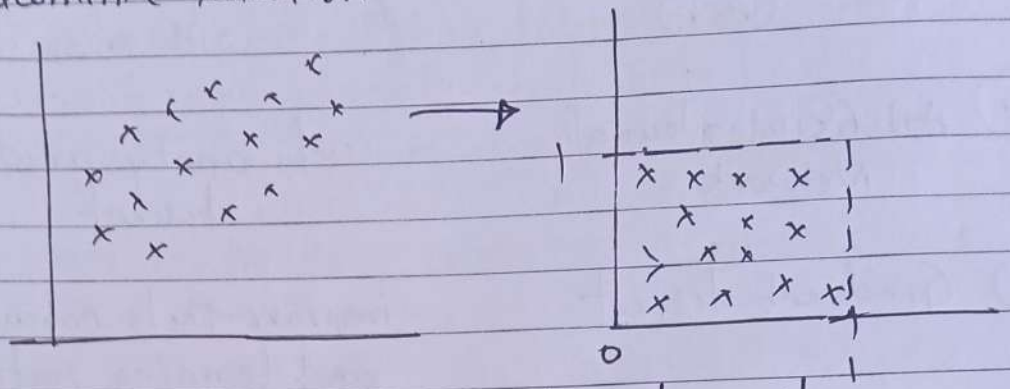
$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

$x_{min} = 32$, $x_{max} = 130$

$$\frac{130 - 32}{130 - 32} = 1$$

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

Geometric Intuition



here we try to squeez the data into unit square

Code

from sklearn.preprocessing import MinMaxScaler

Mean Normalization

$$x_i' = \frac{x_i - x_{mean}}{x_{max} - x_{min}}$$ { mean centering }

- Max Abs scaling

$$x_i' = \frac{x_i}{|x_{max}|}$$

from sklearn.preprocessing import MaxAbsScaler

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

Robust scaling

$$x_i' = \frac{x_i - x_{median}}{IQR}$$ ← { Q3 - Q1 }

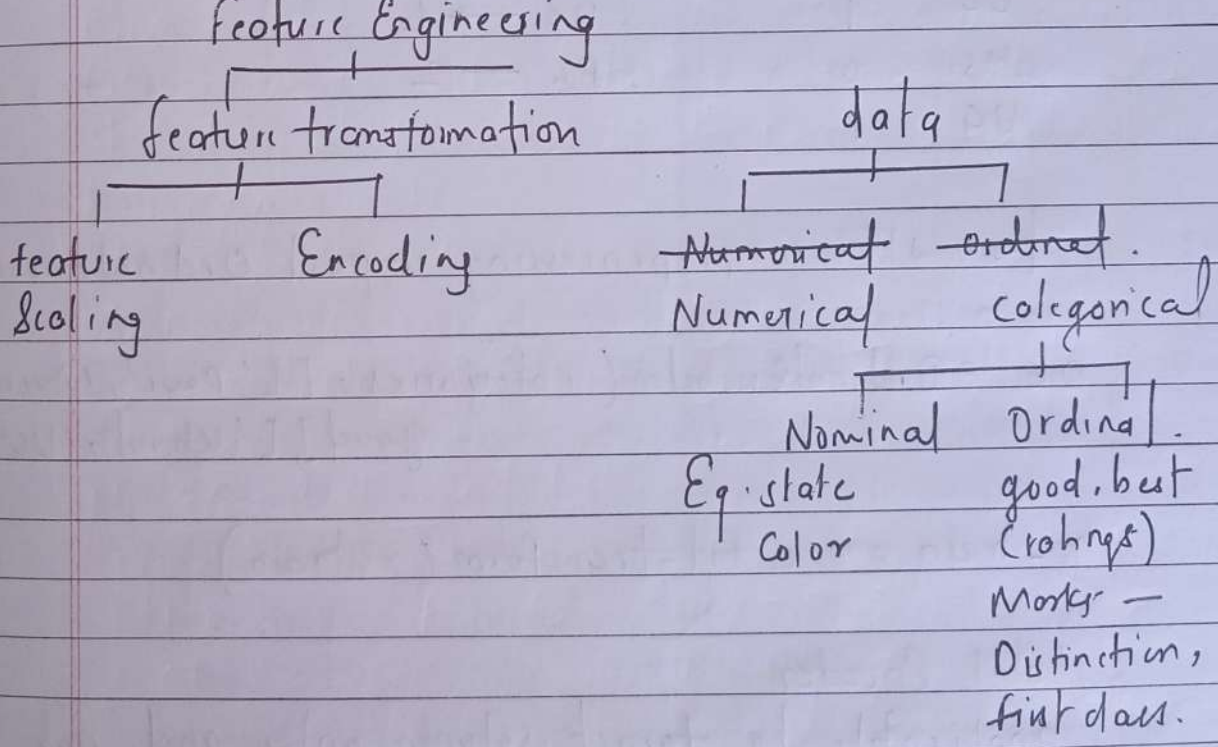
Robust scaler.

- Note : it is Robust to outlier

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 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 :-

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

Ordinal Encoding

Education	
HS	0
UG	1
PG	2
PG	2
UG	1
HS	0
UG	1

= PG > UG > HS — ordinal data

so here you give order to category

HS → 0

UG → 1

PG — 2

```
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)
```

Date: / /
Page

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
yellow	0
yellow	1
Blue	1
yellow	0
red	0
yellow	1
Red	1
red	1
Blue	0

Color Y	Color B	Color R	Target
1	0	0	0
1	*	0	1
*	1	*	1
1	0	0	0
0	0	1	0
1	0	0	1
0	0	1	1
0	0	1	1
0	1	0	0

Dummy Variable Trap

if we have n number category
then keep 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, column=['fuel', 'owner'])

2) dummy trap variable trap (1 hot Encoding)
pd.get_dummies(df, column=['_____'], drop_first = True)

3. One Hot Encoding using sklearn.

Code :-

- from sklearn.preprocessing import OneHotEncoder
- ohe = OneHotEncoder(drop='first', sparse=False, dtype=np.int32)
- X_train_new = ohe.fit_transform(X_train[['_', '_']])

Column Transformer.

in a dataset we have different feature column and need to perform different operation on it is difficult to handle for each column seperately.

Numerical {Nominal categorical} {Ordinal categorical}

for Age | city | gender | review |

Simple Imputer — OHE — OE

[ ] array [,] array [ ] array

[ ] array

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
2) Reciprocal transformation
3) power (sq/sqrt)

BOX-COX
Yeo-Johnson

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

funition Transformer

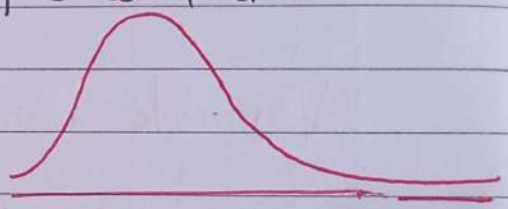
sklearn

- function transformer
  - log transformer
  - Reciprocal
  - sq/sqrt
  - Custom
- power transformer
  - BOX-COX
  - Yeo-Johnson
- Quantile transformer

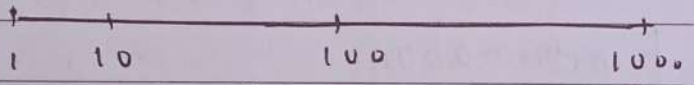
Since we apply these transformation when data is not normally distributed How we will get to know whether data is normal dist or not?

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

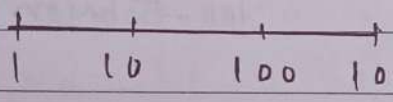
① Log transform.
- it is not applicable to the features containing negative values
- it is applicable to right skewed data



Note:- working of log.



after ↓ log



bring it at equidistant point

② Reciprocal transformation. $1/x$
small value ⇌ large ~~dot~~ value

③ Square (x^2)
can be used when data is left skewed.

④ sqrt $\sqrt{x}$

Code:

```
from sklearn.preprocessing import function.Transformer
trf = functionTransformer(func=np.log1p).
```

trf x_train_transformed = trf.fit_transform(x_train)
x_test_transformed = trf.transform(x_test)

Power transformer
- 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 (λ) that varies over the range of -5 to 5 and in the process of searching we examine all value of λ 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

$\therefore$ ~~$n \neq$~~ $n > 0$ & $n \neq -ve$

this restriction is overcome by Yeo Johnson

Yeo-Johnson Transform

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

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

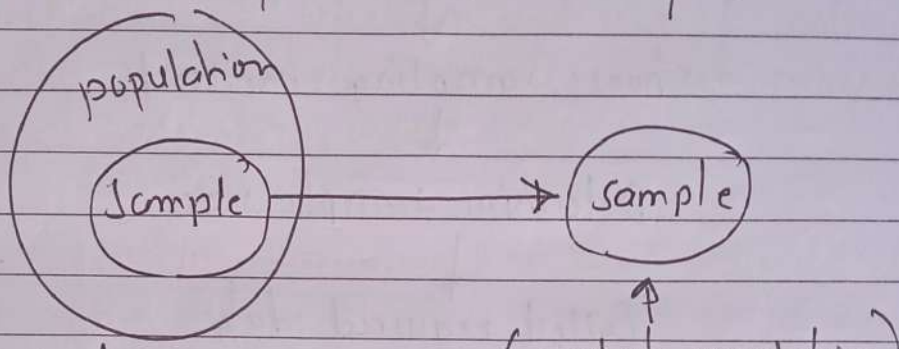
```
from sklearn.preprocessing.import PowerTransformer

pt = PowerTransformer(method='box-cox')
```

Analytics Vidhya
from ~~edureka~~.

~~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 invistigate every individual.



(Data under invishgation)

(part of population)

(Very large, difficult to invishgate)

for Ex: to calculate avg haigh of Men in Bombay its 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 quehon may arise how to take the samples.
- like if we take the sample of only basketball playes this would not consider a good sample because basketball playes are taller than Avg men

why do we need sampling?

- 1) Sampling is done to draw conclusion about population from samples
- 2) It require 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 targel population

↓

Select sampling frame

↓

Choose sampling method.

↓

Determine sample size

↓

Collect required data

for public opinion poll for election

sample frame
- (1) person above 18
- (2) his name should be in list of voter list

sample size
- (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	Non probability Sampling
1. Simple random	1. Convinience
2. systematic	2. Quota
3. stratified.	3. Judgment
4. cluster	4. Snowball

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

Non probability Sampling :- all element 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 choosen by chance

Suppose we have to choose any five individual 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, 15+4=19.

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

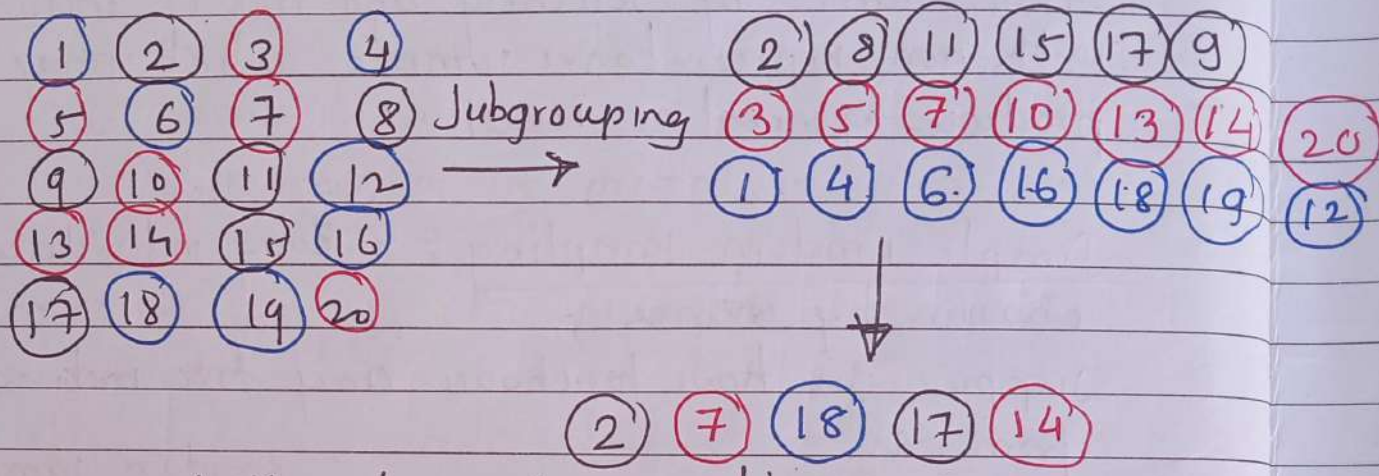
(3) (7) (11)
(15) (19)

it is more convinient than simple random sampling

Date : / /
Page

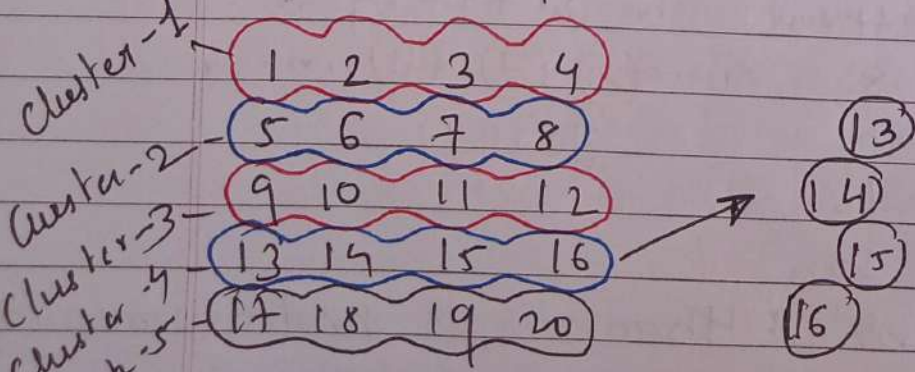
stratified sampling :-

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

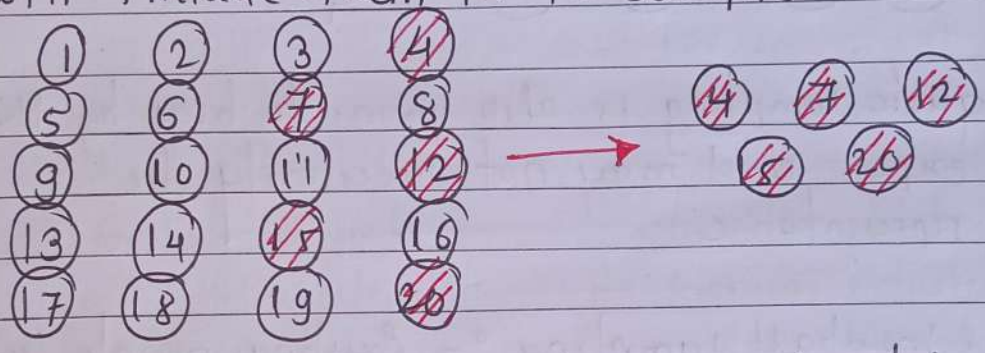
cluster sampling :- here we use the subgroups of population as sampling unit rather than individuals, the population divided into subgroups, known as cluster and whole cluster is randomly selected to be included in study.



Non-Probability Sampling :-

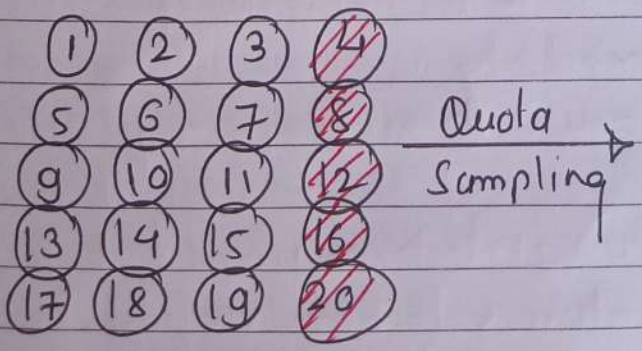
(1) **Convinience Sampling :-** this is perhaps easiest method of sampling because individual are selected based on their availibility and willingness to take part

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

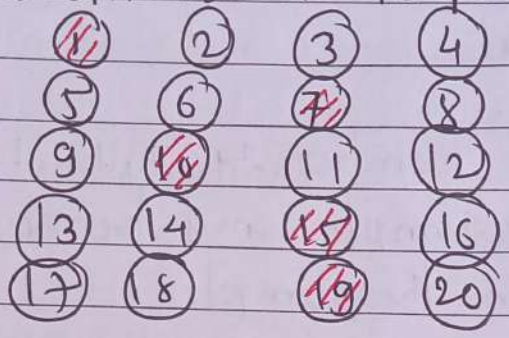


(2) Convinience sampling is prone to significant bias, bcoz the sample may not be the representation of the specific characteristic such as religion or say, the gender of the population

(2) **Quota Sampling :-** In this type of sampling, we choose item based on predermined 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 judgment of the expert when choosing whom to ask to participate.



- 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 increase in size like a rolling snowball.

1 → 6 → 11 → 14 → 19

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

Date: / /
Page

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 best set of features that allows one to build optimized model of studied phenomena.

- **Supervised Techniques**
  - used for labelled data & to identify the relevant features for increasing efficiency of the supervised model like classification & Regression
  - Linear Regression
  - decision Tree
  - SVM
- **Unsupervised techniques**
  - used for unlabelled data
  - Ex. k-mean, clustering, PCA, Hierarchical clustering

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

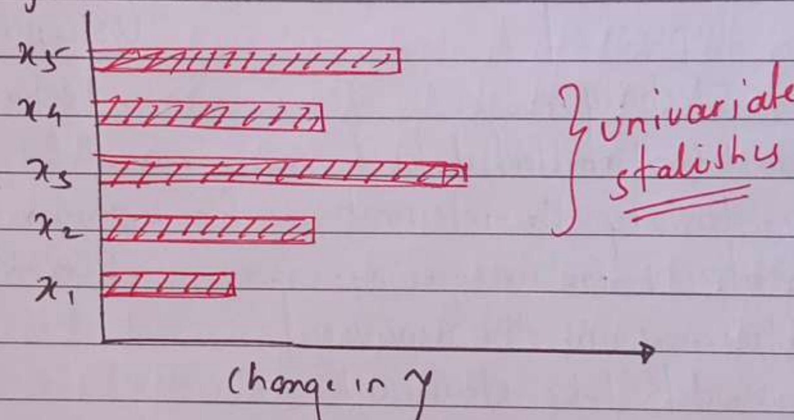
Date: / /
Page

filter Method:-

these method are 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 dataset

- 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 to test the relation between feature and target variable 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
- the 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 meet some threshold. By default it removes all zero variance features i.e feature with same value in all sample or observation

Wrapper Method :-

- The wrapper method usually result in better prediction 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.
- Types: 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 preset condition is achieved.

Backword feature Elimination :-

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

Recursive feature Elimination :-

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

Embeded Method :-

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

LASSO Regression (L1) :-

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

- In Unic 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 therefor 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
- Node 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 dataset 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
- f1score
- AUC
- Confusion Matrix
- Recall
- ROC

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

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

for 80 : $\frac{80}{100} = 0.80 = 80\%$ accuracy

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

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

Que: 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 some 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 is actually belong to zero but predict 1

So to overcome this we have another solution called as confusion Metrix.

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

from confusion Metrix we can calculate accuracy but from accuracy we cant calculate confusion Metrix.

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

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

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

for multi classification problem

Actual \	0	1	2	
0	7	0	5	
1	8	2	3	
2	1	4	7	

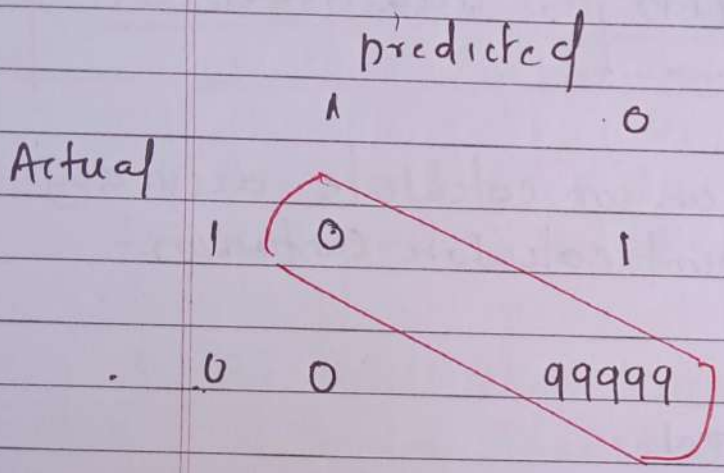
accuracy : $\frac{\text{diagonal element } (7+2+7)}{\text{sum of all element}}$

when accuracy is misleading ?

Ans :- Imbolanced 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 modely will highly trained on normal people and will predict everyone as not terrist with accuracy almost close to zero.



$$\frac{99999}{100000} = 0.9999$$

99.99 %

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

if type 1 Error are more dangerous then we we precison

if type 2 Error are more dangerous the we use recall.

Precision :- for type 1 Error → FP

which mean it is ↑ belong to 0 but predicted 1

model A

model B

Page

for Ex

Actual	predicted: spam	predicted: Notspam
spam	100 TP	170 FN
Notspam	30 FP	700 TN

Actual	predicted: spam	predicted: Notspam
spam	100 TP	190 FN
Notspam	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 why model B is good to finalize for deploy

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

model A

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

B

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

{ good }

My when type 2 is dangerous means FN is dangerous we take Recall

	model A: Cancer	model A: Not Cancer
Cancer	1000	200
Not cancer	800	800

	model B: Cancer	model B: Not cancer
Cancer	1000	500
Not cancer	500	8000

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

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

Model A

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

model B

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

model A is good to deploy since less no of mistake or FN (false negative)

but some time 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
	cat	0	30	4	34
	Rabbit	4	10	20	34
		29	45	34	108

Date: / /
Page

to calculate Precision

first we will calculate individual precision

TP / Predicted

$P_D = \frac{25}{29}$ $P_C = \frac{30}{45}$ $P_R = \frac{20}{34}$

$= 0.86$ 0.66 0.58

there two method of combining to mean

for balanced classes →

1) Macro precision

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

for imbalanced classes →

2) weighted precision

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

$= 0.71$

to calculate Recall

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

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

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

$0.22 + 0.20 + 0.17 = 0.59$

similar we do f1 score

$F1_{Dog} = \left(\frac{2PR}{P+R}\right)_{Dog}$; $f1_{cat} = \left(\frac{2PR}{P+R}\right)_{cat}$; $f1_{Rabbi} = \left(\frac{2PR}{P+R}\right)_{Ra}$

- macro f1 score
- weighted f1 score

Softmax Regression.

- accuracy, confusion matrix, Precison and Recall and f1 score also work very fine with Binary class, but what if multiclass problem is there, there 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?

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

placement → if yes → 1, No → 2, opt - 3

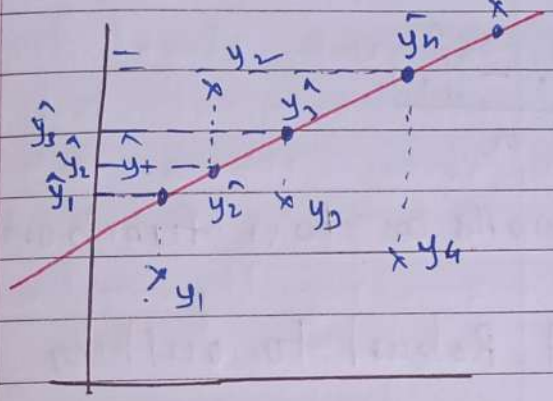
$$\sigma(z)_1 = \frac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}} \qquad e(z)_2 = \frac{e^{z_2}}{e^{z_1}+e^{z_2}+e^{z_3}}$$

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

Evaluation Metrix for regression Model.

1) MAE (Mean absolute Error)
2) MSE (Mean Jquare Error)
3) RMSE (Root Mean Jquare Error)
4) R2 - Score (coefficient of determination)
5) Adjusted R2 Score

1) MAE = Mean absolute Error



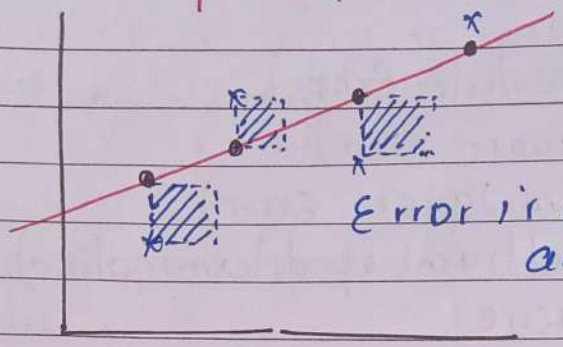
$$MAE = \frac{|y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + |y_3 - \hat{y}_3| + \cdots + |y_n - \hat{y}_n|}{n}$$

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

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

Disadvantages: graph of modulas function is not differential at zero.

Mean Squared Error (MSE)



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

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

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

Advantages :- it works as loss function

disadvantage :- Not Robust to outlier

Here since Error is in terms of square little difficult while Explaining you have to explain with undu root-

RMSE (Root Mean square Error)

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

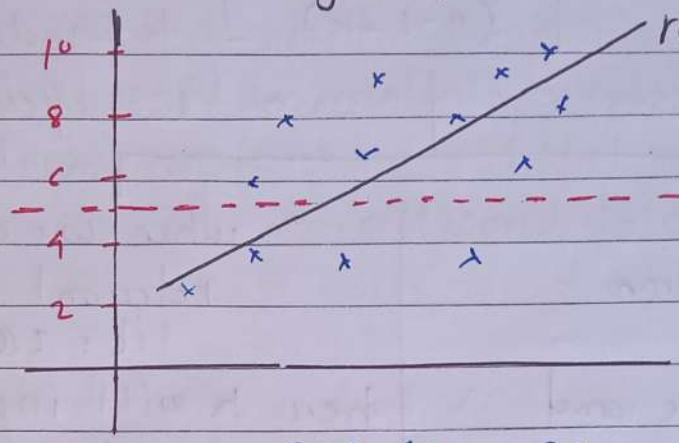
benefit :- Error you get in terms of y-unit

disadvantages. it is not Robust to outlier

Date: / /
Page

R^2 - score

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



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

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

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

if R2-score = 0.80

- it interpretation like this.
  variance in output column is coming 80% from Feature we selected for model training.
- R2-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 = R2-score
n = no of rows
k = independent feature

- 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 closed to zero rapidly. and overall nominator will decrease rapidly and overall fraction will be less..

- so overall adj R^2 will increase

classification mode Evaluation technique (free of cutoff value)

AUC - ROC Curve.
(Receiver Operating Characteristic Curve)

- for classification model Evaluation we have Metric like Recall, f1, accuracy and precision which is base 50% cutoff value, but it could not same for all the case in real world dataset sometime we may need to check for different range cutoff value [1% - 99%]

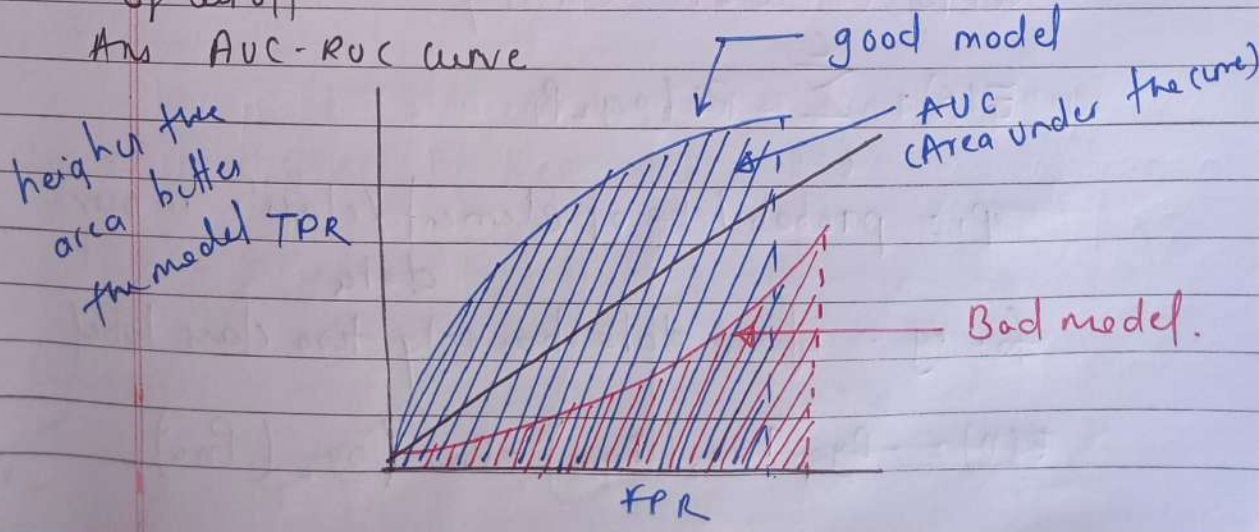
- Here we plot the graph between TPR and FPR rate for good model TPR supposed to be higher than FPR

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

Actual \ predicted	1	0
1	TP	FN
0	FP	TN

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

Ans AUC - ROC curve



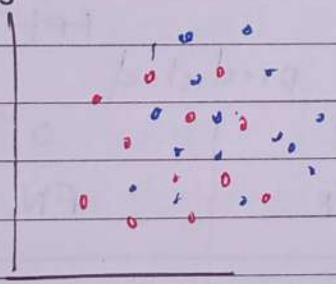
Date: / /
Page

- 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 called as measure of purity / impurity

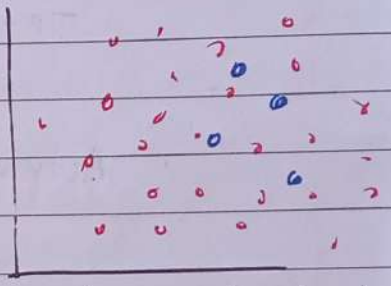
In physics term
Entropy of vapour > liquid water Entropy > Entropy of ice

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



to predict is difficult

high Entropy



to predict is quit 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.

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

for Example

Salary	Age	Purchase
20k	24	Yes
10k	45	NO
60k	27	Yes
15k	31	NO
12k	18	NO

Salary	Age	Purchase
34k	31	No
15	25	No
69	57	Yes
25	24	No
32	28	No

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

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

$= -\frac{2}{3}\frac{\log(\frac{2}{3})}{\log 2} - \frac{3}{5}\frac{\log(3/5)}{\log(2)}$

$= (-0.66 \times -0.58) - (0.6 \times -0.73)$

$= 0.38 + 0.43$

$= 0.81$ wrong

$= 0.97$

{ Entropy is high }

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

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

$H(d) = 0.72$

{ Entropy is low }

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

O

ly for multiclass / 3 class problem

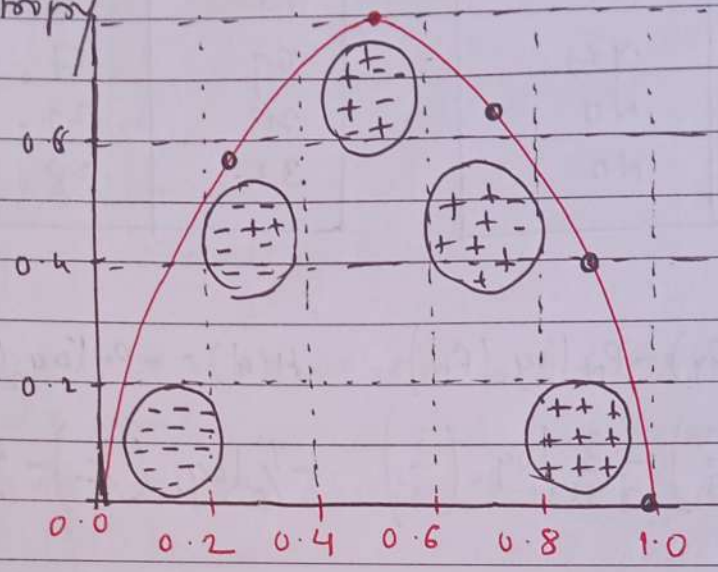
$H(d) = -P_Y \log_2(P_Y) - P_N * \log_2(P_N) - P_M \log_2(P_M)$

Conclusion: More the uncertainty More the entropy.

2) For 2 class problem the min entropy is zero and max would be 1

3) More more than 2 classes the min entropy is 0 but max can be greater than 1

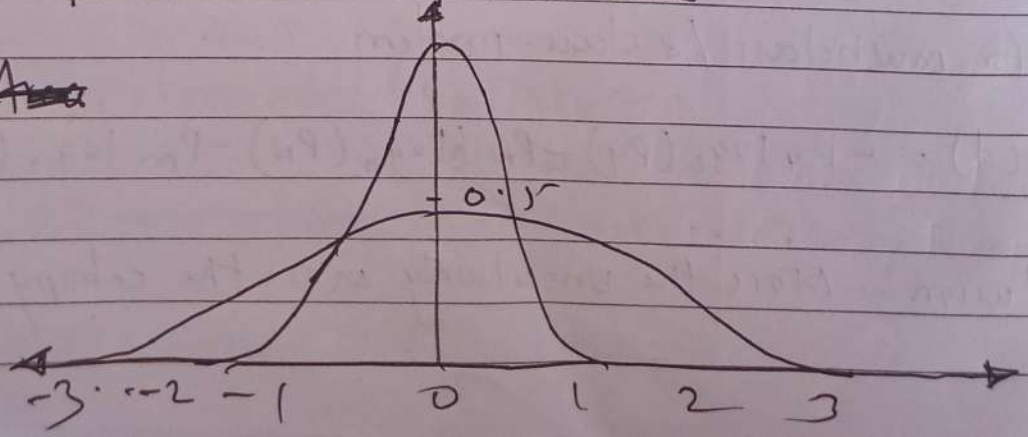
3) $\log_2$ or $\log_e$ both can be used to calculate entropy



Entropy for Continuous Variable

x_1	x_2	y	x_1	x_2	y
-	-	3.5	-	-	4.6
-	-	5.6	-	-	6.5
)	.	-			
:	-	7.3	-	.	12.8

will plot for both output variable



whichever 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.

Information Gain = E(parent) - {weight Avg} x E{child}

Consider a dataset

outlook	temp	Humidity	windy	playtennis
Sunny	Hot	High	false	No
sunny	Hot	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

$= -P_y \log(P_y) - P_N (\log(P_N))$ — 0.05

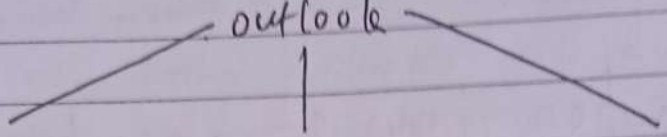
$= -9/14 \log\left(\frac{9}{14}\right) - \left(5/14 \log(5/14)\right)$

$= 0.12 - (-0.79)$

$= 0.12 + 0.79$

$= 0.91$

Now we will calculate entropy of children based on outlook



Outlook	Humid	Wind	Play
Sunny			N
"			N
"			Y
"			Y

E(S) = 0.97 E(OV) = 0 E(R) = 0.97

step 3 :- calculate weighted entropy of children

$5/14 \times 0.97 + 4/14 \times 0 + 5/14 \times 0.97$

$W \cdot E(\text{children})_{\text{outlook}} = 0.69$

Note: whereas entropy is 0 it is leaf Node }

step 4 :-

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 hieghest information gain algorithm will select that column to split the data

step 6: find information Gain ~~successfully~~ recursively.

Gini Impurity :-

formula for Entropy

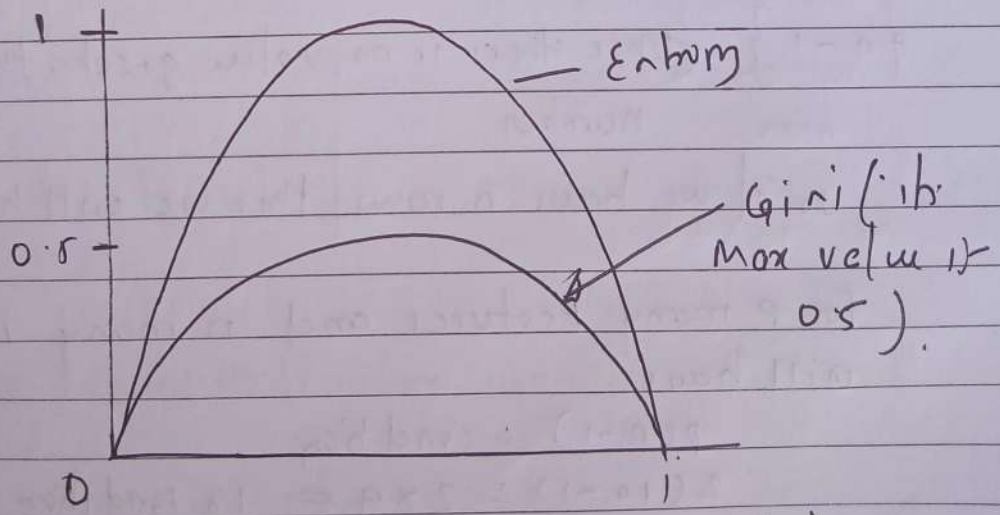
$$E = -P_Y \log(P_Y) - P_N \log(P_N)$$

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 lets ice any continuous data

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

calculate gini impurity :. $1 - (p_1^2 + p_2^2)$

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

in such condition we have to test the model for different conditions for each colum 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 paribl splitting condition we will select that split where ~~least impurity~~ information gain is higest. or less entropy.

Cross Validation (kfold)

In kfold 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 tel 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) Bat suitable for small dataset
2) Better for multiclass classification
3) more reliable
4) useful for model selection

Hyperparameter Tuning

- GridSearchCV
- RandomizedSearchCV

Parameter

Model parameters	Hyperparameters
These are the parameter of the model that can be determined by training with training data. These can be considered as internal parameters. weight & bias $Y = wX + b$	Hyper parameters are the parameter 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_estimators

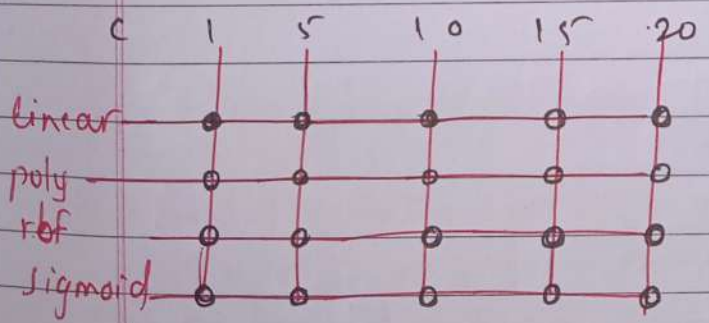
optimum values means for that perticular set of hyperparameter we will get highest result

- we do model training ⟶ Best model parameter
- hyper paramita tuning ⟶ Best hyper parameters
  ↓
  so that we can get highest accuracy or efficiency.

Hyperparameter tuning :-

hyperparamita 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' value and 4 kernal.
so in grid search CV we will try all 20 combination of C and kernel and among that whichever combinat will give highest result will be analize for best hyperparameter but in Randomized instead of trying all 20 we randomly choose few and among them whoever give highest result will

finalize as best hyperparameters

when to use gridsearch cv :- Based on its way of working we can easily get to know it will take lot of time for large datasets so if we are time constraint than 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.

Date: / /
Page

* 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
i) 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.

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

population mean : μ
sample mean : $\bar{x}$

{ outlier influence the central tendency of data }
↑
Extreme behaviours

Page

Median :-

- Median is 50th 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 number and take their mean.

Mode: most frequently occured element

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

mode can be found out for both quantative and Qualative 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.
- insimple terms - it measures of for the data deviates from mean value
- std dev formula varies for population & Sample

sample std deviation =

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

population std deviation =

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

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

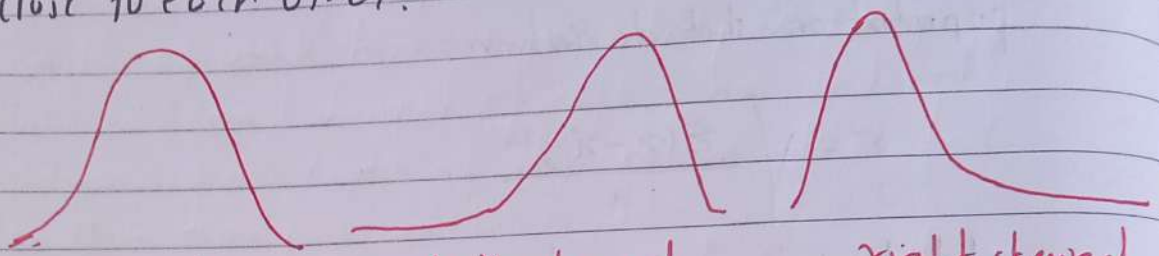
σ^2 = population Variance
s^2 = sample variance

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$ — population variance

$$= \frac{\Sigma(x-\bar{x})^2}{n-1}$$ — 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.

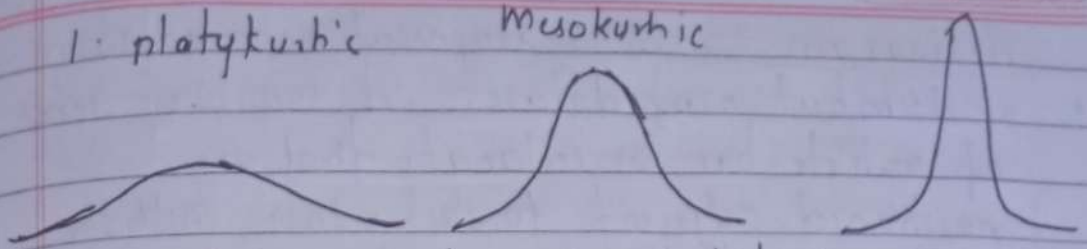


skewness :- it is measure of asymmetry in the distribution of data
a) positively skewed
b) negative skewed

1. Positively skewed :- here data value 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 value 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.
1. Platykurtic
2. Mesokurtic
3. Leptokurtic.

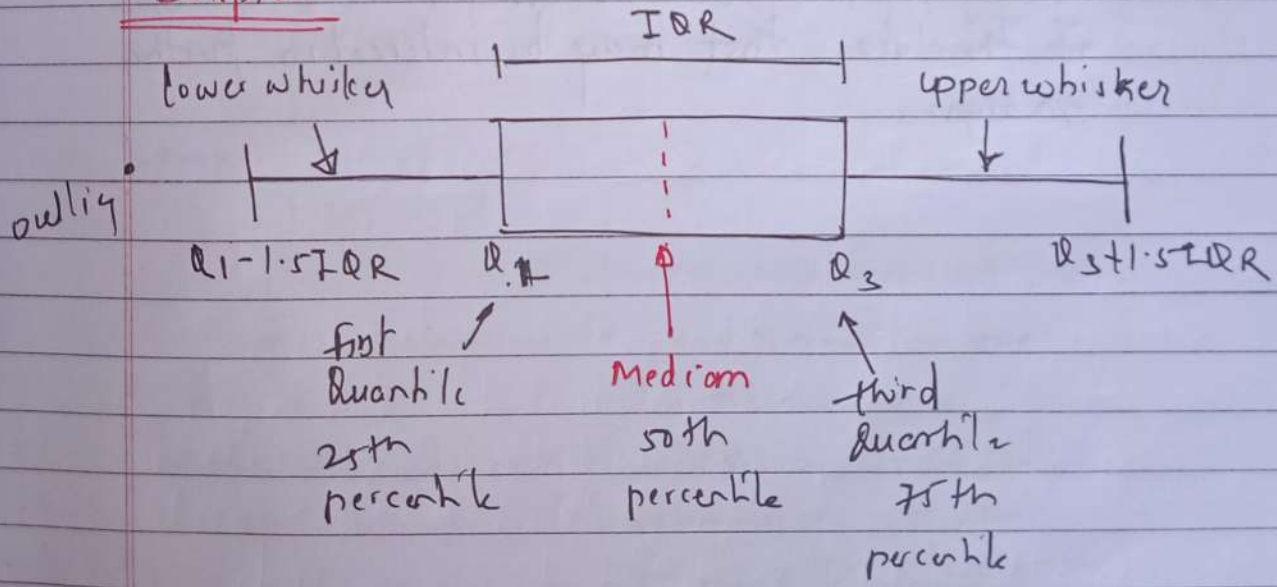


1. platykurtic show the distribution with flat tails Here data is distributed flatly it indicates small outlier in the distribution

2. Mesokurtic : it is normal distribution where data spread widely

3. lepto kurtic :- the data very closely distributed. the highest of peak is greater than width of peak

Boxplot :-



box plots are useful tool for summerizing 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.

Date : / /
Page

Histogram :- it is graphical representation of distribution of dataset. it consist 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

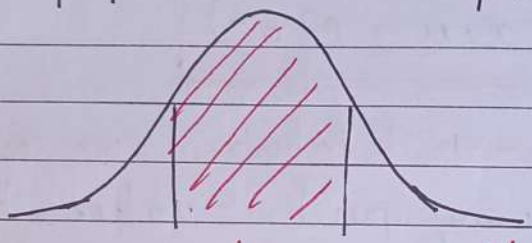
- it is commonly used in analysis and statistics to visualize the distribution of continuous variable, such as, wt, height by dividing range into equal interval or bins.
- Basically it is useful tool for visualizing the distribution of data and gaining insight into characteristics of dataset they can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

Inferential statistics

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

Normal distribution

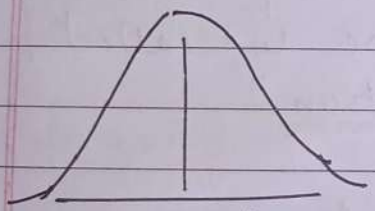
Majority of data tends to follow normal distribution



mark width .. is many

max value in middle

Normal distribution has two properties
1) mean & variance



mean/median

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

- total area under bell curve is 1

$std = \sqrt{variance}$ ∴ $\sigma = \sqrt{variance}$

σ^2 : variance

Normal distribution denoted by

$x \rightarrow N(\mu, \sigma^2)$ — for population

μ : mean , σ^2 : variance

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

$x+5 \longrightarrow N(\mu+5, \sigma^2)$

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

and for data multiply by any constant
for ex $x \longrightarrow (\mu, \sigma^2)$ & $x = 2x$ then

mean $\longrightarrow 2\mu$

$\sigma^2 \longrightarrow 2^2\sigma^2 = 4\sigma^2$

Conclusion :-

$X = N(\mu, \sigma^2)$

$X + c = N(\mu + c, \sigma^2)$

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

How many Normal distribution we can have

- μ, σ^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

mean $= \mu = 0$ $\sigma^2 = 1$

how to convert ND $\longrightarrow$ STD

ND : $X \rightarrow N(\mu, \sigma^2)$

$X = N(0, 1)$

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

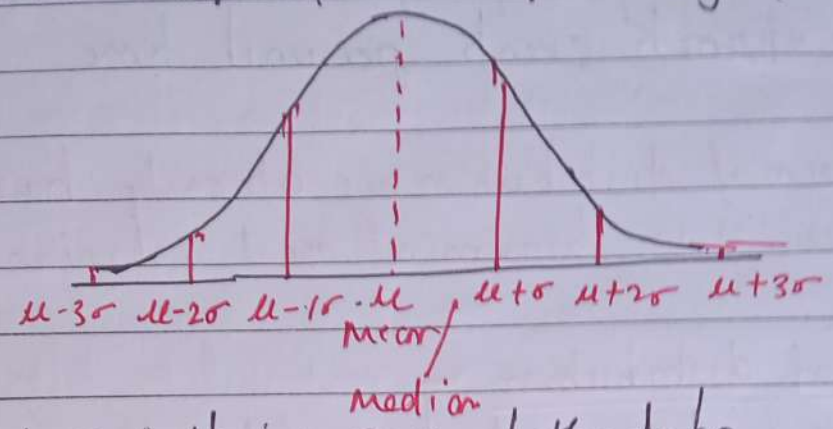
$x - \mu : N(\mu - \mu, \sigma^2)$

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

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

$\frac{x - \mu}{\sigma} : N(0, 1)$ => which is nothing but : $\frac{\text{data} - \text{mean}}{\text{std. deviation}}$

- significance of std dev. : measurement of dispersed data in relation with mean.
- low std dev data clustered closed 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 poission distribution are three important probability distribution in statistics and data analysis.

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

- **The binomial distribution** it is descrete probability distribution that used to model number of success in fixed number of idipedent trial. it is characterized by two parameter. - the probability of success in single trial number of trials

Poission 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 events per unit time

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

Binomial distribution :-

- in this we have only two outcome

coin < H / T ; exam < pass / fail ; rain < yes / No ; Match < win / loose

in case of binary two 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

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 $x = 0$

$$P(X=0) = p^0 (1-p)^{1-0}$$
$$= (1-p) = q$$

for $x = 1$

$$P(X=1) = p^1 (1-p)^{1-1}$$
$$= p$$

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

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

$$\text{pmf} = \begin{cases} q = 1-p & \text{if } x=0 \\ p & \text{if } x=1 \end{cases}$$

let try to understand with ex of fair coin

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

$p(T) = 1-p = 1-1/2 = 1/2$

there could be another scene

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

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

How we will compute mean & variance and std deviation

to calculate mean

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

~~for $p(x=0) = 0.4 = p$~~

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

$p(x=1) = p$

$= 0.4$

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

$= q = 0.6$

$= \cancel{0 \times 0.4 + 1 \times (0.6)}$

$= \cancel{0.6 = p}$

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

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

Note for Bernoulli distribution your mean is equal to p.

Date : / /
Page

Now let calculate variance : $p(1-p) = pq$.
$= SD = \sqrt{pq}$: standard deviation

Note $p > 0.5$ - then p is median or else 'q' will be the median

2. if $p > q$, '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
- poission distribution is used to understand independent events that occur at constant rate within given interval of time

Example of poission distribution.

- Number of accident occuring in a city from 6pm to 10pm
- Number of patient arriving in an emergency room between 10pm to 12pm

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.

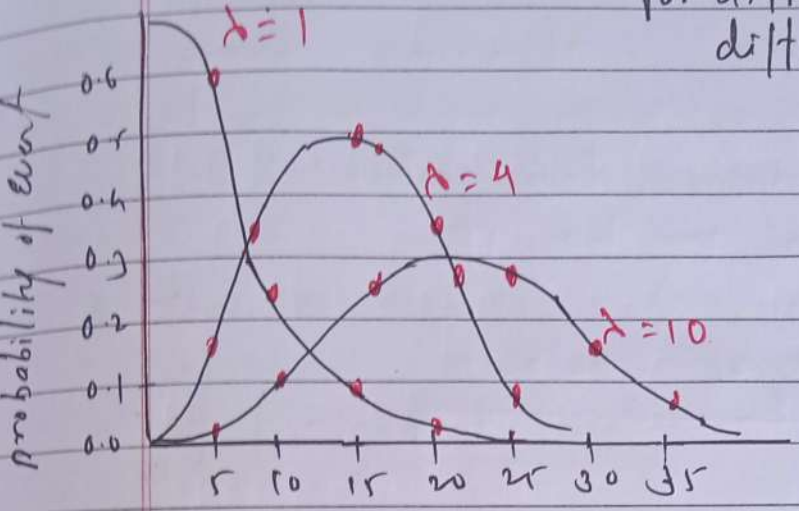
$P(x)$ Probability

λ : Mean Number of Events

$x!$ = factorial of x

e = Euler's Number (2.71828)

for different value of λ we have different shape of wave



Central limit therom :-

- the central limit therom 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 number and each of them represent mean of each sample

Date : / /
Page

step 3: then we plot distribution of n samples means and we are done.

Probability.

Date: / /
Page

- Probability means possibility it is branch of mathematics that deals with occurance 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
Eg: 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 not affect other.

- if any event is Independent Event then it follows following property.

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

Date : / /
Page

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 befor expecting or rejecting it

There are two type of hypothesis

Null hypothesis	Alternative hypothesis
• Existing Belief	• Somebody claiming

Example sun rise in south
H_0 : Sun rise in the east (Existing Belief)
H_1 : Sun rise in south (claim)

There are three type in taking of hypothesis

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

how tal type 1
duration of life before and after covid.

$H_0 : \mu = 60$
$H_1 : \mu \neq 60$.

1) collect a sample data

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

if σ^2 is known

for the sample

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

to convert into standard normal.

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

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

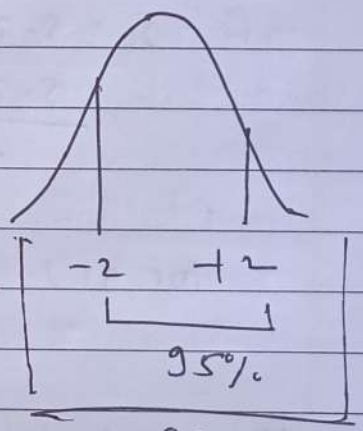
Let assume H_0 is true $\mu = 60$

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

$$z = \frac{\bar{x} - 60}{6/\sqrt{n}}$$

Assume

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



if calculated value of z is in range of -2 to $+2$ Eg. 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 : $\mu = 4$ $\sigma = 3$.

$H_0 : \mu = 4$

$H_1 : \mu \neq 4$

Date: / /
Page

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

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

~~$\frac{\bar{x} - \mu}{\text{std dev}} = \frac{x - 4.6}{3} \quad \frac{4.6 - 4.0}{3} = \frac{0.6}{3}$~~

~~$z = 0.2$~~

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

$Z = 1.414$ which is betwe -2 to $+2$

for 95% interval

will accept H_0 and Reject H_1

If $\bar{x} = 5.3$

$$= \frac{5.3 - 4.0}{3/\sqrt{50}} = \frac{0.43 \times \sqrt{50}}{3}$$

$$= 3.06$$

for 95% and $z = 3.06$ H_0 will be rejected

H_1 will be accepted

How to test type 2

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

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

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 for type 1, 2, and 3.

$H_0 : \mu \leq 25$

$H_1 : \mu > 25$

for Assuming null hypothesis is true

$\mu \leq 25$ multiply both sides by -1

$-\mu \geq -25$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - 25}{\sigma/\sqrt{n}}$$

$$z \geq \frac{\bar{x} - 25}{\sigma/\sqrt{n}}$$

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

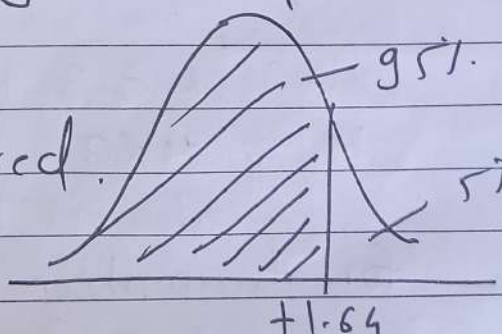
$$\frac{27 - 25}{2/\sqrt{100}} = \frac{2}{2} \times 10 = \frac{20}{2} = 10$$

$z \geq 10$

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.

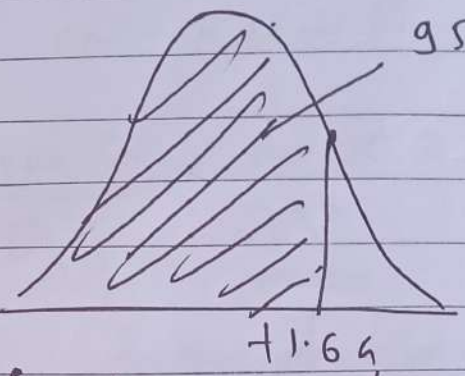


Solve the problem for test type 2

$H_0 : \mu \leq 120$

$H_1 : \mu > 120$

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



$\mu \leq 120$

$-\mu \geq -120$

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

$Z \geq \frac{\bar{x} - 120}{\sigma/\sqrt{n}}$ (anything greater is problematic)

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

$=$

$Z \geq 2.23$.

and this 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

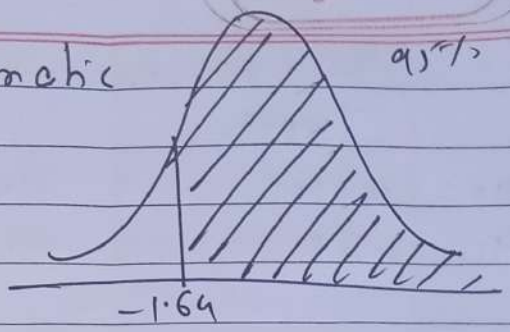
$\mu \geq 3$

$-\mu \leq -3$

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

anything lesser is problamatic

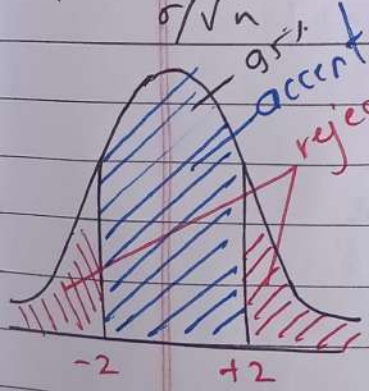
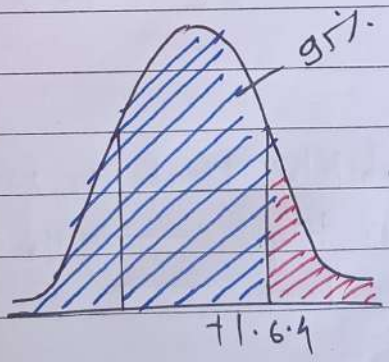
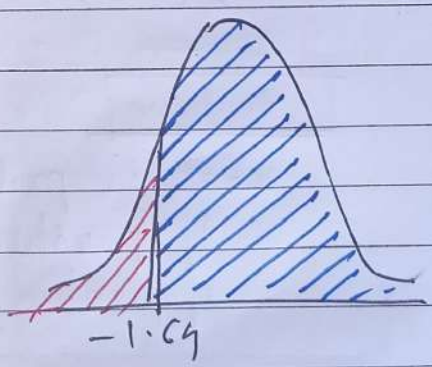
$$\therefore z \leq \frac{2.9 - 3}{0.2/\sqrt{8}}$$



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

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

Short Summary — σ is known

Type 1	Type 2	Type 3
$H_0: \mu = 10$	$H_0: \mu \leq 10$	$H_0: \mu \geq 10$
$H_1: \mu \neq 10$	$H_1: \mu > 10$	$H_1: \mu < 10$
$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
T_s = Test statistic $= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	(greater is problamatic) $T_s \leq \frac{\bar{x} - 10}{\sigma/\sqrt{n}}$	(small is problamatic) $T_s \geq \frac{\bar{x} - 10}{\sigma/\sqrt{n}}$
		
anything in middle will accept it. i.e $-2 < T_s < +2$ H_0 V o/w reject and accept H_1	if $T_s \leq 1.64$ H_0 : accept H_1 : reject	if $T_s \geq -1.64$ H_0 : accept H_1 : reject.

Date: / /
Page

if σ is unknown.

Type1	type2	type3
$H_0: \mu = 10$	$H_0: \mu \leq 10$	$H_0: \mu \geq 10$
$H_1: \mu \neq 0$	$H_1: \mu > 10$	$H_1: \mu < 10$

Sample std deviation $\frac{\bar{x} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1}$ this becomes t distribution

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

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

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

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

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

- As number of datapoint goes close to population t-distrib 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 x-axis

Eg. $Z[-2, +2] \longrightarrow t[-1.98, +1.98]$

P-Value

what is probability that H_0 is true

if p value is high $\longrightarrow H_0$ ✓

if p value is low $\longrightarrow H_1$ ✓

Sensitivity varies from domain to domain hence we can't boldly predict or make decision instead we can give probability for each hypothesis to be true and let client take the decision.

so "p-value" is nothing but probability of H0 or H1 is true

Let see how p-value is calculated for type 1 ?

$H_0 : \mu = 60$

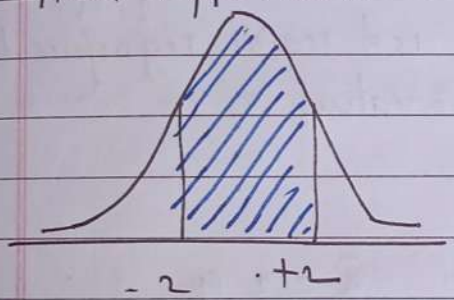
$H_1 : \mu \neq 60$

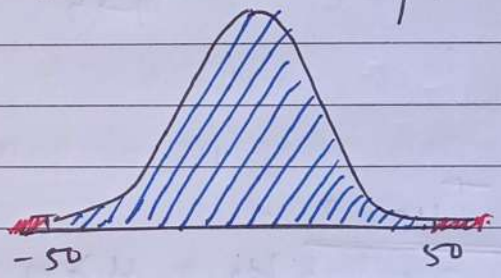
test statistics

$$tr \rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \qquad \bar{x} = 70, n = 100, \sigma = 2$$

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

earlier for q's. we would have reject if for null hypothesis so p value must be very less.





probability of H0 to be true is area under red region.

let check with acceptance

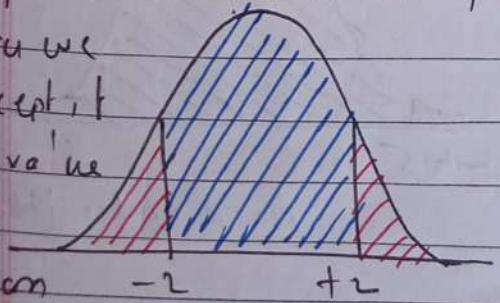
if $\bar{x} = 60.005$, $n = 100$, $\sigma = 2$

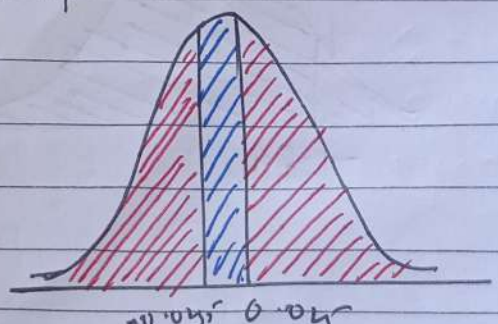
$$ts = \frac{60.005 - 60}{2/\sqrt{100}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{60.005 - 60}{2/\sqrt{100}}$$

$$= \frac{0.005}{2/10} = \underline{\underline{0.025}}$$

for this earlier we would accept it.

since earlier we would accept it have p value must be high mean near or under red region.





Let calculate p-value for type 2

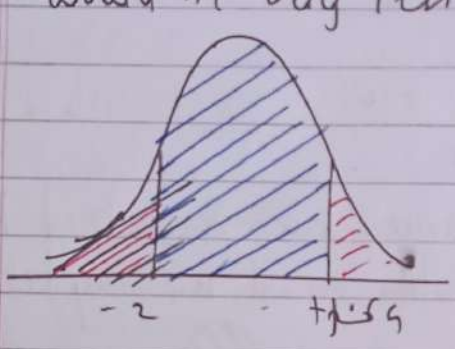
$H_0 : \mu \leq 100$

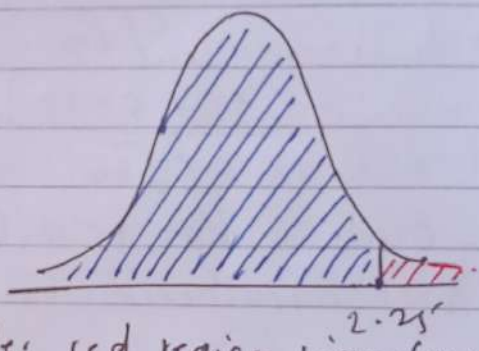
$H_1 : \mu > 100$

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

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{130 - 100}{40/\sqrt{80}} = 2.23$$

Hence earlier we would reject it. hence p value would be very less.





Here red region signifies less p-value

Type 3 :-

$H_0 : \mu \geq 3$

$H_1 : \mu < 3$

$\bar{x} = 2.9$

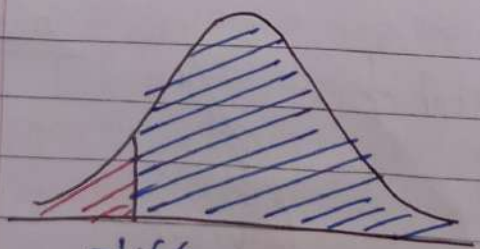
$\sigma = 0.2$

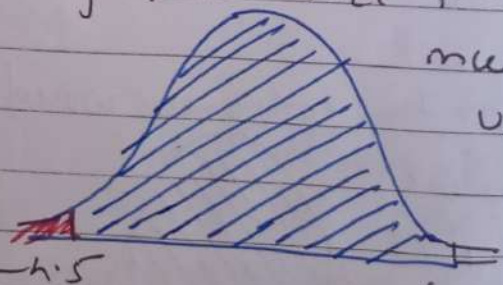
$n = 81$

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

$$= -4.5$$

earlier we would have reject it hence p value must be 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 rejed null hypothesis when it is true is 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 α			Confidence level $1-\alpha$	
0.01	1%	→	0.99	99%
0.05	5%	→	0.95	95%
0.1	10%	→	.90	90%

level of significance α	Corresponding confidence interval in term of z-value
0.01	-1.645σ to $+1.645\sigma$
0.05	-1.96σ to $+1.96\sigma$
$	-2.58σ to $+2.58\sigma$

Confidence ~~level~~ Interval

- it is applicable to continuos data.
- Since probability of continuos data is always zero for heigh, weigh temperature
- so whenever we have to estimate about continous data we give interval with surety in terms of percentage
- let try to understand with Example

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

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

so here we have to find mean height of Men in ~~b~~ Banglore.

- we will calculate confidence interval for μ which is avg ht of population.

population μ ⟶ sample μ

$$: \frac{x_1 + x_2 + x_3 + \cdots x_n}{n}$$

$P(\text{sample } \mu) = 0 \longrightarrow [a, b]$ -- Interval.

step 1 if σ^2 is known.

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

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

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

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

{ 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

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

for σ^2 if known

let eqn (1) convert it into standard Normal distribution

$$\frac{\text{data} - \text{mean}}{\text{std dev}} :$$

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

for 95% distribution of data it must be lie between

$$-2 \text{ to } +2$$

$$-2 < z < +2$$

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

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

multiply both side 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 μ at 95% confidence is

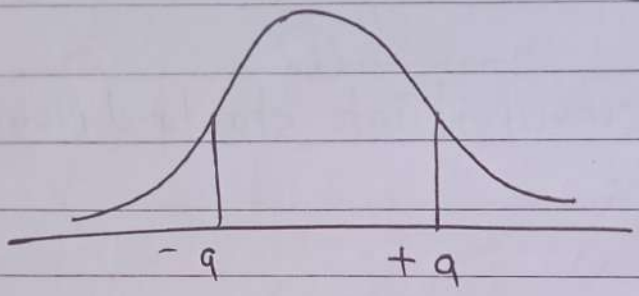
$$\mu \longrightarrow \left[\bar{x} + \frac{2\sigma}{\sqrt{n}}, \bar{x} - \frac{2\sigma}{\sqrt{n}}\right] \quad \text{— 95%.}$$

$$\mu \longrightarrow \left[\bar{x} + \frac{\sigma}{\sqrt{n}}, \bar{x} - \frac{\sigma}{\sqrt{n}}\right] \quad \text{— 68%.}$$

$$\mu \longrightarrow \left[\bar{x} + \frac{3\sigma}{\sqrt{n}}, \bar{x} - \frac{3\sigma}{\sqrt{n}}\right] \quad \text{— 99.7%.}$$

Now general formula.

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



for σ^2 is Unknown which is more realistic scenario in real word nobody is going to tells variance or σ^2

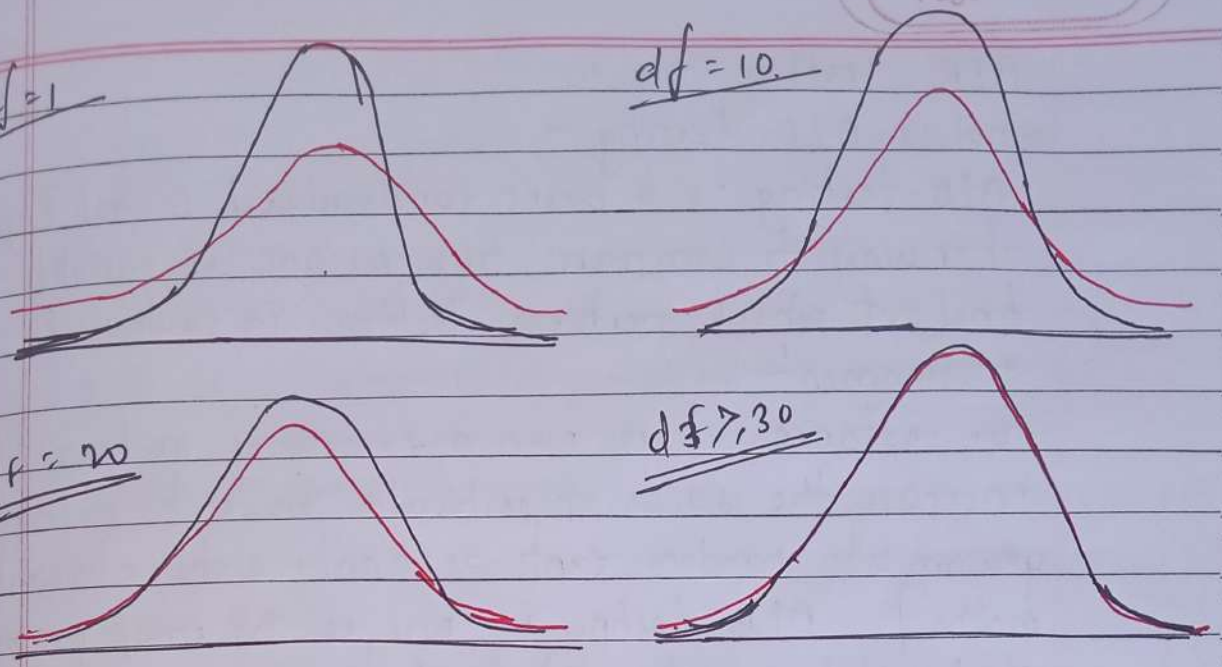
- so when i dont know population variance we will calculate sample variance

population Variance σ^2 ⟶ sample variance s^2

σ^2 known	σ^2 Unknown
$x_1, x_2, x_3 \cdots x_n$	$x_1, x_2, x_3 \cdots x_n$
$\bar{x} \longrightarrow N(\mu, \sigma^2/n)$	$\bar{x} \longrightarrow N(\mu, \sigma^2/n)$
$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1)$	$\frac{\bar{x}-\mu}{s/\sqrt{n}} \longrightarrow t_{n-1}$ ↑ this is called t-distribu

- when sample std deviation (s) will become close to population standard deviation (σ) t-distribution will become standard Normal distribution.

- t-distribution and z-distribution look like.



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

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

Summary :-

σ^2 is known	σ^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

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

small n

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

Date: / /
Page

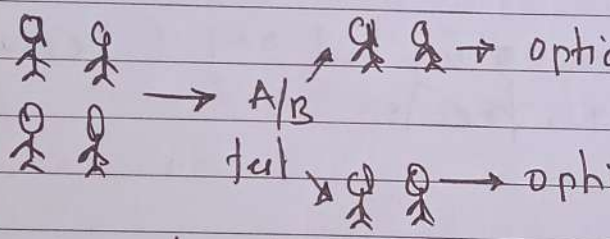
A/B testing :-

what is A/B testing :-

A/B testing is a basic randomized control Experiment it is 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 perticular product either you can use random expt or apply statistic/scientific method . A/B testing is one of the most prominently and widely used stochical tools .

Ex option A : same old product
option B : updated old product.



Now based on this reponce 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 change are working as desired or not.
- A/B testing doesn't work well when testing major changes like new product, new branding or for completly new use experience . in that cases there may be effect that drive highur than the normal engagment or emotional responce that may cause users to behave in different manner.

Statistical tests

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

1) we know population variance σ^2 or
2) or we don't know population variance σ^2 but our sample size is larger than $n \geq 30$.

- if we have $n < 30$ and don't know population variance (σ^2) 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.

$$z\text{-score} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$\bar{x}$ — sample mean; μ — population mean; σ — population std deviation; n — sample size.

let try to understand with example.

Que: we need to determine if girls on average score higher than 600 in the exam we have the information the std deviation of girl score is 100, so we collect the data of 20 girls by using random sample and record their marks finally we also select an α value (significance level) to be 0.05.

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

$\sigma = 100$.

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

$= \frac{41 \times 4.47}{100}$

$Z_1 = 1.8327$

for if p value = 0.9664

~~1.8327 → 0.9664 =~~

$1 - 0.9664 =$ 0.0336

critical value is 1.645

Z score > critical value

p value < 0.05

We will reject null hypothesis and accept alternative hypothesis

$H_0 : \mu \leq 600$ ✗

$H_1 : \mu > 600$ ✓

since the p. value is less than 0.05 we can reject null hypothesis and conclude based on our result. girls 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 bet^n two sample mean

difference between population mean

$$\text{Z-score} : \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

population std deviation.

sample size

Let understand with example.
if we want to know if girls on avg score 10 marks more than boys we have info std dev (σ) of girls is 100 and for boys is 90. the we collect the data of 20 girls and 20 boys using random sample and record their marks finally we also set our α value to be 0.05.

Solution :- • Mean score for girls (sample mean) is $\bar{x}_{girls} = 641$
• Mean score for boy (sample mean) is 613.3
• std deviation for the population of girls is 100
• std deviation for the populatio of boys is 90.
• sample size is 20 for both Girls and boys.
• difference between mean of populatio is 10.

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

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-Score = 0.588 $\leq$ critical value 1.642

$1 - 0.6844 = 0.3156$

pvalue > 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 avrage score of 10 marks more than boys.

Date: / /
Page

What is the T-test ?

T-test are statistical way of testing hypothesis when

- we dont 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 dont have information about population variance. we use sample std deviation instead of population std deviation in this case.

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

- $\bar{x}$: Sample mean
- μ : population mean
- s : Sample std deviation.
- n : sample size

let understand with Example.

lets 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 α value (significance level) to be 0.05 for hypothesis testing.

In this Example

Mean score for girl = 606.8 = $\bar{x}$
size of sample = 10
population mean is = 600.
std deviation for sample = 13.14

type 2
$H_0 : \mu \leq 600$
$H_1 : \mu > 600$

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

$$= \frac{606.8 - 600}{13.14/\sqrt{10}} = 1.64$$

for t-distribution critical value for 95% is 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.

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

- diff betn sample mean $\bar{x}_1 - \bar{x}_2$
- Difference betn population mean
- Sample size n_1, n_2
- Sample standard deviation s_1, s_2

Let try to understand with example

Let say if we want to determine if an average boy score 15 mks 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 mks we choose α-value 0.05 (significance level) as criteria for Hypothesis testing.

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

$H_1 : \mu_1 - \mu_2 > 15$

In this example.

$\bar{x}$ Boys = 630.1

$\bar{x}$ girls = 606.8

$(\mu_1 - \mu_2) = 15$

& $s_{Boys} = 13.42$

$s_{girls} = 13.14$

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

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

$= 2.23$

critical value = 1.8333

t exceed ~~are~~ the critical value

pvalue = $1 - 0.9871$

$= 0.0129 < 0.05$

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

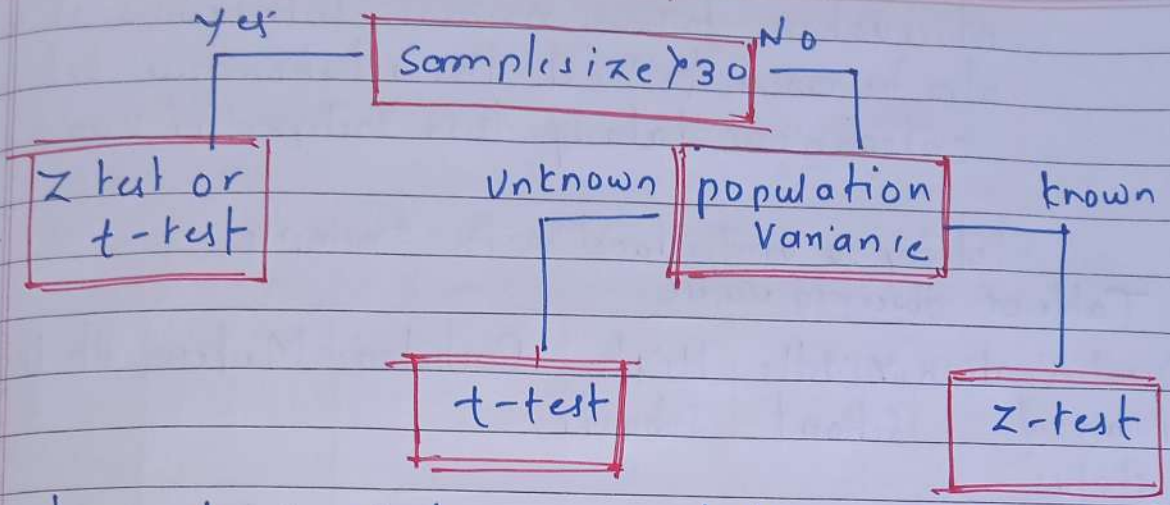
$H_0 : \mu_1 - \mu_2 \leq 10$ ✗

$H_1 : \mu_1 - \mu_2 > 10$ ✓

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

Date : / /
Page

Deciding between z test / T-test



- 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 approches the normal distribution the difference between the z-score and t-score is negligible

chi-Square (χ^2) test:

- chisquare test is hypothesis testing method used to test the hypothesis about the chi-square distribution of observation/frequency of different categories
- How 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 try to understand with Example

Table of observed value

Qualification / Marital status	Middle School	High School	Bachelor	Masters	Ph.D	Total
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
Total.	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_0: there is no relation between two variable

H_1: there is significant relation between two variable

Significant level (α) = 0.05.

Table of expected values.

Qualificahe / Marital status	Middle School	High School	Bachlor	Mash	ph.D	Total
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	

Date: / /
Page

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

Now our aim is to find how much deviation is their 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)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
18	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	3.3	−0.3	0.09	0.02

$$\sum \frac{(O-E)^2}{E}$$

$$\chi^2 = 23.57$$

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

we also need to shared value with tabular χ^2

for that we should to Dof

Dof : (col − 1) (row − 1)

: (5 − 1) (4 − 1) = 4 × 3 = 12

then from "percentage point of chi square distribution)

	0.5 ← significance level.
DOF → 12	21.03

$\chi^2_{tabular} = 21.03$ $\chi^2_{calculated} = 23.17$

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

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

H_1: 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 group 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 sample 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 these groups.

let try to understand with example.

standard deviation

$$S_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$ — how data is spread.

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

- we have three groups of student each contain 4. for whome 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. (x)	x^2	student	Questions (x)	x^2	student	Que (x)	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.	$\Sigma x_1 = 32$	$\Sigma x_1^2 = 266$		$\Sigma x_2 = 25$	$\Sigma x_2^2 = 165$		$\Sigma x_3 = 17$	$\Sigma x_3^2 = 77$

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.

$\Sigma x_1 = 32 \mid \Sigma x_1^2 = 266$ $\quad \Sigma x_2 = 25 \mid \Sigma x_2^2 = 165$ $\quad \Sigma x_3 = 17 \mid \Sigma x_3^2 = 77$

Correction term : $CT = \frac{(\Sigma x)^2}{N} = \frac{(32+25+17)^2}{12} = \frac{5476}{12} = 456.33$

Sum of squares of total :

$SS_T = \Sigma x^2 - CT = (266+165+77) - 456.33$

$= 508 - 456.33 = 51.67$

Sum of squares Among group :

$SS_A = \frac{(\Sigma x)^2}{n} - CT = \left(\frac{32^2}{4} + \frac{25^2}{4} + \frac{17^2}{4}\right) - 456.33$

$= 484.5 - 456.33 = 28.17$

Sum of squary within group

$SS_W = SS_T - SS_A = 51.67 - 28.17 = 23.5$

Mean of sum of square among group.

DOF for among group

$$MSS_A = \frac{SS_A}{K-1} = \frac{28.17}{3-1} = 14.085$$

Mean of sum of square within group :

DOF for within group

$$MSS_W = \frac{SS_W}{N-K} = \frac{23.5}{12-3} = 2.611$$

No. of student

category

$$F\text{-Ratio} = \frac{MSS_A}{MSS_W} = \frac{14.085}{2.611} = 5.394.$$

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	2.6111	

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

2

↓

9 → 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 ~~feed~~ effect of noise on number of question solved.

and if for $\alpha = 0.01$ 1%.

F value 8.0215

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

5.394 < 8.0215

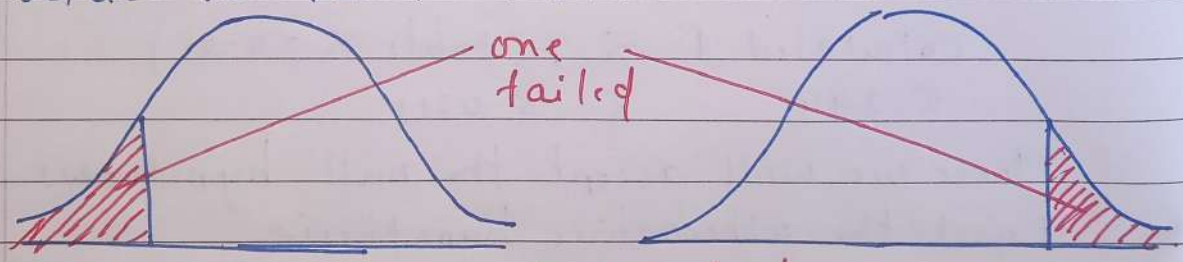
Hence we will accept the null hypothesis and Reject the alternative hypothesis

like Hence we are 99% confidence we can't say that there is significant effect of Noise on number of question solved.

Difference between One tailed and Two tailed test.

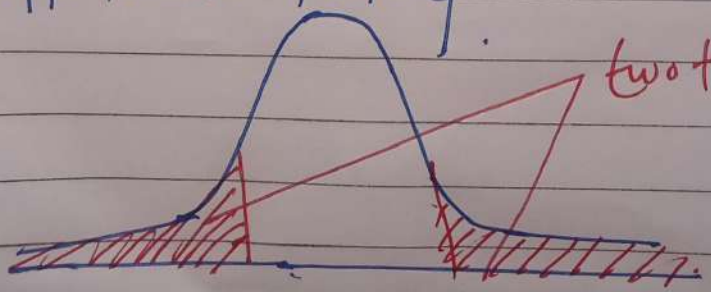
one tail and two tailed test are always identify relationship between statistical variable. for checking the relationship between variables in single direction (left or right) we use one-tailed test. A two tailed test is used for checking whether the relation between variable in any direction or not.

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



one tailed test

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.



Date : / /
Page

Difference between one and two tailed test

one tailed Test	Two-tailed test
• A test of any statistical hypothesis where H_1 is one direction anything greater or smaller.	A test of hypothesis where H_1 is non-directional.
• sign we used for one-tailed. $>$ or $<$ for the alternative hypothesis	sign we used is $\neq$ for alternative hypothesis.
• Critical region lies entirely on either left side or right side of sampling distribution	critical region is given by the portion of area lying on both the tails of sampling distribution.
• Her entire level of significance (α) test. is either in the left or right tail.	it splits level of significance (α) into half.