DATE ___/___/___

# Statistics

## Data Types

- Categorical
  - Normial
    - This is without order
    - e.g. {This is place, play, Jump}
    - There is not in order form
  - ordinal
    - This is with The format or order
    - e.g. {child, young, old}
- Numerical
  - Discreate
    - {0, 1, 2, ...}
    - without Decimal Value
  - Countious
    - hight 193.2cm
    - with decima Valu

DATE / /

# Statistics → Statistics is the discipline that concerns the collection, organization, analysis interpretation and presentation of data

## Descriptive

① Analyze data, Summarizing data
organizing data that is the form of Number & Graph

② Bar plot, Histogram, Pie chart
pdf, cdf, Normal Distribution

③ Measure of Central tendency
(Mean, median, Mode)

④ Measure of variance
Standard deviation

## Inferential Statistics

(Confidence interval)
- Z-test
- T-Test
- hypotesis test
- Chi-square Test

Actually taking the some sample doing some inferential Testing.
(Take make a conclusion)
e.g exit poll,

DATE __/__/____

# Inferential Statistics →

Based on sample data of population we will do some conclusion or decision that is called inferential statistics where there we made decision for this.

e.g → Exit poll etc

some time called this to confidence interval

# Random Variable → (RV) →

every feature is present in dataset is a Random variable

Random → Numerical Random → Discreate RV, Countine R. variable
→ Categorical Random variable

x = 24

x → RV → RV is storing 24 (value)

RV is any type of data store, some time is called place folder.

| Age | Gender | |
|---|---|---|
| f1 | f2 | f3 |
| 23 | M | |
| 24 | F | |
| 25 | M | |

Categorical RV → It is repeating The Records in many times

discreate RV → It is whole number

e.g → family member, no. of people

Countine → loan amount, Salary

Basically decimal level value.

DATE __/__/____

# Population & Sample Understand

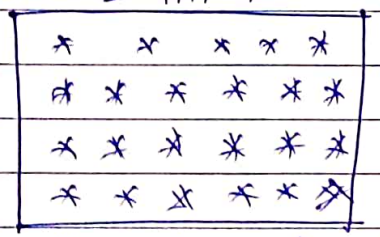
1- Population (denoted by N)
2- Sample (denoted by n)

1- Population mean ($\mu$) (Myu)
2- Sample mean ($\bar{x}$) (x bar)

We calculate the Height of any State Population →

Population (TOTAL No. of people denoted by N)

1 Milion



(Population Size)

$$\text{Average (population mean } (\mu)) = \frac{1}{1M} \sum_{i=1}^{M} x_i$$

got

Note:- But We can not check each & every persons of night. because its very difficult to find it.

So Picture Come

DATE __/__/____

## Sample

$$\text{Average (Sample mean } \bar{x}) = \sum_{i=1}^{10000} \left( x_i \times \frac{1}{10000} \right)$$

We can some sample of data from population in different diffent place

e.g exit poll (we can find the which ponties is going to win)

## In ML Terms

** All the Records is the sample based on their we have the predication *

population count = N

Sample Count = n

population mean = $\mu$

Sample mean = $\bar{x}$

DATE / /

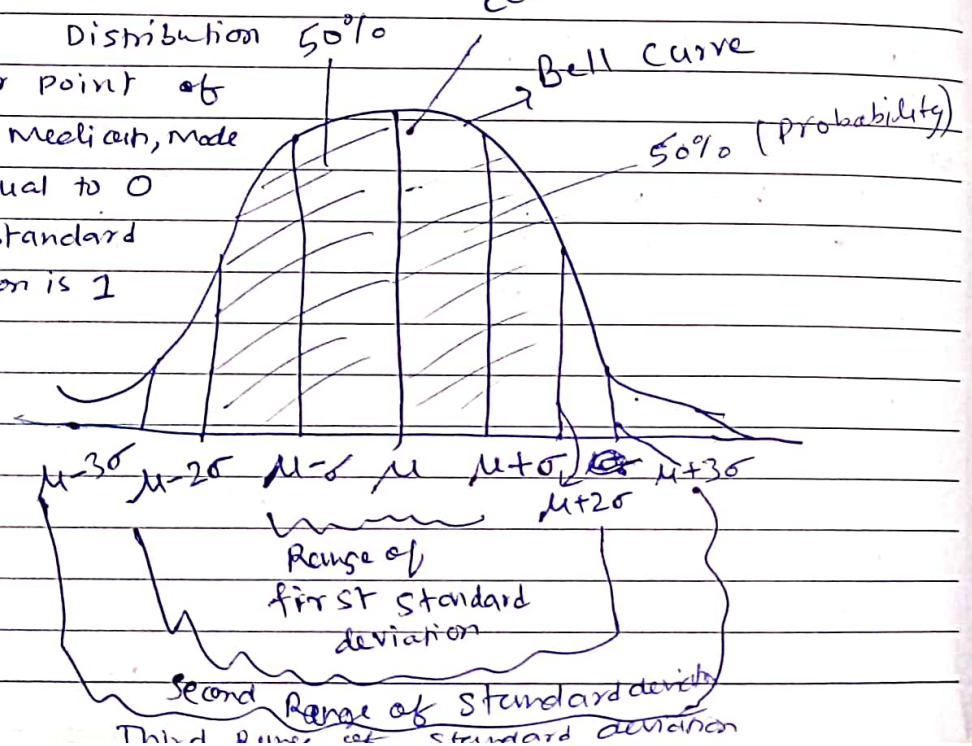
OR

# Normal / Gaussian Distribution

$$X \simeq GD(\mu, \sigma)^{\sigma^2}$$

- $X$ ↓ Random variable of $x$
- $\simeq$ ↓ belong to
- $GD$ ↓ Gaussian or Normal Distribution
- $\mu$ ↓ mean
- $\sigma$ ↓ Standard deviation (Sigma)
- $\sigma^2$ ↘ $\sigma^2$ = variance

When → Data is Normal Distribution




Center point of Mean, Median, Mode is equal to 0 and standard veriyon is 1

DATE / /

if $x$ (random variable) follow the Normal distribution, Then are Three property ~~these properties~~ is called Empirical formula

1- $P_r[\mu-\sigma \leq x \leq \mu+\sigma] \simeq 68\%$

↓ probability

Belonging to Random variable of $x$ & fall to the first deviation is almost 68%.

Because why saiy $x \simeq GD/ND$

↓ Belong to

2- $P_r[\mu-2\sigma \leq x \leq \mu+2\sigma] \simeq 95\%$

↓ Approx belong to

↓ These lies under second Range of standard deviation

3 - $P_\sigma [\mu - 3\sigma \leq x \leq \mu + 3\sigma]$ ≈ 99.7%

lies to Third Range of Standard deviation

In ML Terms

When we are doing EDA to plotting some of the feature it is having follow the same type of the curve

Internally follow of Empirical formula of all three Properties.

DATE / /

# Mean / Median / Mode

Mean ⇒ find the Average of feature are given is below

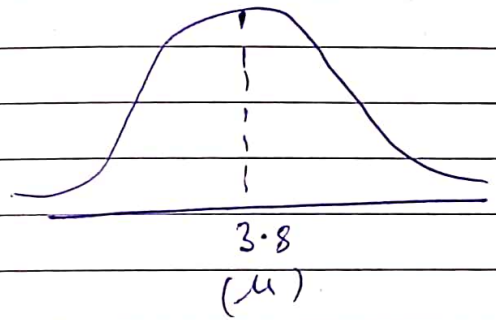
$x = [1, 2, 3, 7, 6]$

$\bar{x}$ or $\mu = \frac{1+2+3+7+6}{5}$

$\bar{x}$ or $\mu = 19/5$

$\bar{x}$ or $\mu = 3.8$

$\bar{x}$ or $\mu = 3.8$




Mean is measure of Centre tendency.

Mean/ Median/ Mode is help to Add some missing value.

$$\text{Media} = \sum_{i=1}^{n} \left(\frac{n_i}{2}\right) + 1$$

Mode = No. of times the particular value occour

DATE / /

# Range

The Range is the difference between The "lowest and highest values"

Example

4, 5, 6, 9, 3, 7

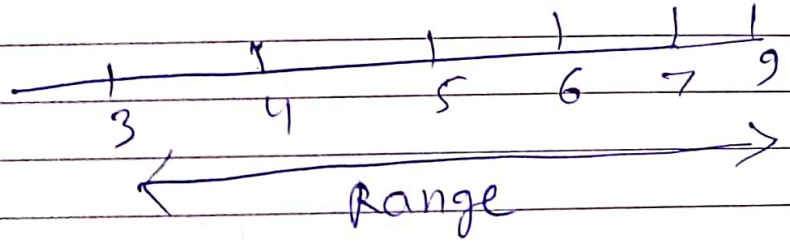
lowest value is = 3

highest value is = 9

$$Rang(X) = Max(x) - Min(x)$$

Range(x) = 9 − 3.

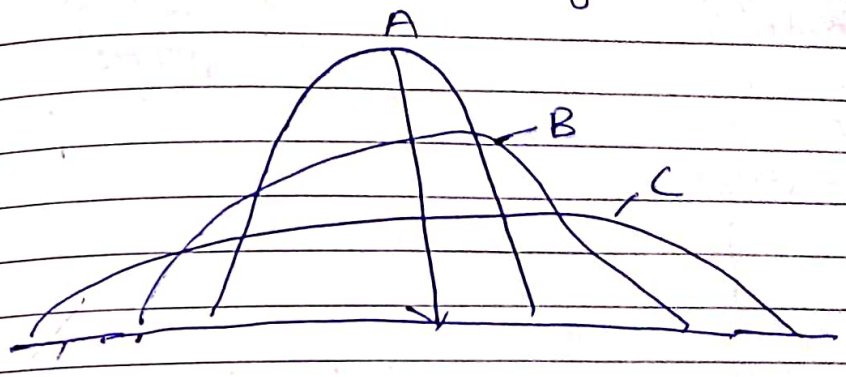
Range(x) = 6




DATE / /

# Measure of Dispersion

Dispersion helps to understand the distribution of data.




Measure of dispersion help to interpret the variability of data.

↓

In ML terms, it describe the variation of "observe variables"

↓

To find How much homog or hetrogenous the data

A simple terms →

It Show How Squeezed The Variable is.

# Type of Measure of dispersion

- Absolute Measure of Dispersion
  - Range
  - Variance
  - Standard Deviation
  - Quartiles and Quartiles Deviation
  - mean & Mean deviation
- Relative measure of Dispersion
  - Co-efficient of Range
  - Co-efficient of Variation
  - Co-efficient of Standard deviation
  - Co-efficient of Mean Deviation
  - Co-efficient of quartiles deviation

DATE __/__/__

## Variance →

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$$

where $\mu$ → mean, $N$ = Total Number of Population

$$\text{variance} = (\text{standard deviation})^2$$

$$\text{Variance} = \sigma^2 = (\text{standard deviation})^2$$

$$\text{Standard deviation} = \sqrt{\sigma}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

when $n$ = no. of observation in Sample set

$\bar{x} = \mu$ = Mean

$x_i$ = each Records of the data

e.g → Find the variance of the Number

3, 8, 6, 10, 12, 9, 11, 10, 12, 7

Sol →

$$\text{mean} = \frac{3+8+6+10+12+9+11+10+12+7}{10}$$

$$= \frac{88}{10}$$

$$\mu = 8.8$$

$x_i - \mu = 3 - 8.8 = -5.8$

$8 - 8.8 = -.8$

⋮

$7 - 8.8 = -1.8$

$(x_i - \mu)^2 = (-5.8)^2 = 33.64$

$= (-0.8)^2 = .64$

$= (-2.8)^2 = 7.84$

⋮

$= (-1.8)^2 = 3.24$

DATE

Adding of $(x_i - \mu)^2$

$= 33.64 + 0.64 + 7.84 + 1.44 + 10.24 + 0.04 + 4.84 + 1.44 + 10.24 + 3.24$

$= 73.6$

$$\sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N}$$

$$= \frac{73.6}{10} \quad \text{(Total No. of Sample)}$$

$$\boxed{\sigma^2 = 7.36}$$

DATE / /

# Co-efficient of dispersion

Range of co-efficient = $\frac{X_{max} - X_{min}}{X_{Max} + X_{min}}$

Quartile deviation = $\frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$

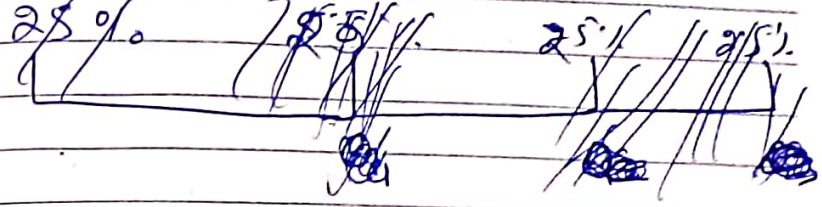
SD = $\frac{SD}{Mean}$

Mean Deviation = $\frac{\text{Mean deviation}}{\text{Average}}$

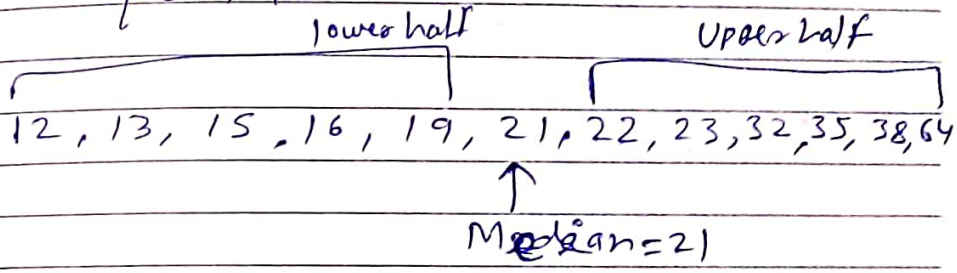
DATE / /

# Quartile formula

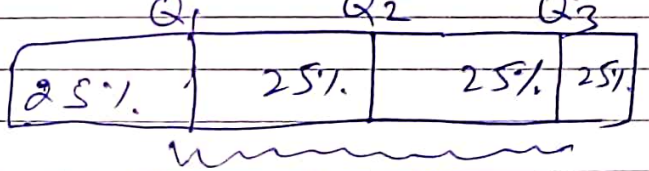
A quartile divide the set of observation into 4 equal parts.




QD → Mean the semi variation between the upper quartiles ($Q_3$) and lower quartiles




Interquartile Range (IQR) →




IQR = Q3 − Q1

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{ term}$$

(25 Percentile)

$$Q_2 = \left(\frac{n+1}{2}\right)^{th} \text{ term}$$

(50th percentile)

$$Q_3 = \left(\frac{3(n+1)}{4}\right)^{th} \text{ term}$$

IQR = Upper Quartile − lower Quartile

$$IQR = Q_3 - Q_1$$

DATE / /

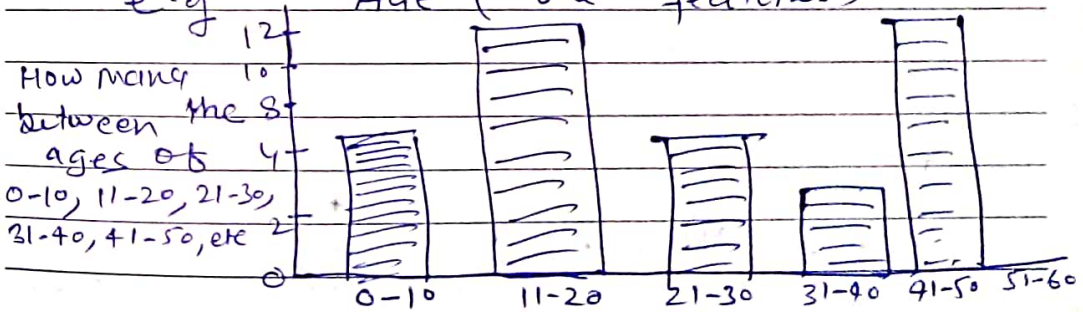
# Histrogram

A histogram provide a visual representation of distribution of a dataset

- Skewness of the data
- It helps to visualize whether the distribution is symmetric or skewed left or Right

- Histogram is used to summalize discreate or continuous data
- Histrogram is represent mostly by Bar Chart.
- Histogram are commonly used in statistics to demonstrate how many of a certain type of variable occurs within a specific Range

e.g Age ( one features)



DATE __/__/__

— Histogram, It provide a visual representation of numerical data by showing the number of data points fall within a "specified range of values" (bin is called)
It is similar to vertical graph

## CENTRAL LIMIT THEOREM

$X \not\sim GD(\mu, \sigma^2)$

may be / may not be
belong to

Where $n \geq 30$

n is data point is greater than or equal to 30 or more.

Suppose we take 30 random sample

$S_1 = x_1, \ldots, x_{30} = \bar{x}_1$

$S_2 = x_1, \ldots, x_{30} = \bar{x}_2$

$\vdots =$

$S_{100}$ $x_1, \ldots, x_{30} = \bar{x}_{100}$

$$X \simeq GD\left(\mu, \frac{\sigma^2}{n}\right)$$

that is is a deviation of 100 Random sample

DATE __/__/__

## Covariance → is a measure of the relationship between two Random variable

very important for Data preprocessing

Suppose we have a example of House pricing

| Size | price |
|---|---|
| 1200 Sqft | 10000 |
| 1300 Sqft | 15000 |
| 1400 Sqft | 20000 |

See

Size Price

Quantify Relationship between them

Size ↑ price ↑ (price increase)

Size ↓ price ↓ (price decrease)

Covariance equation

$$Cov(\text{Size}, \text{Price}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

x y

$$\text{Variance}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x) * (x_i - \mu_x)$$

x is relation to x

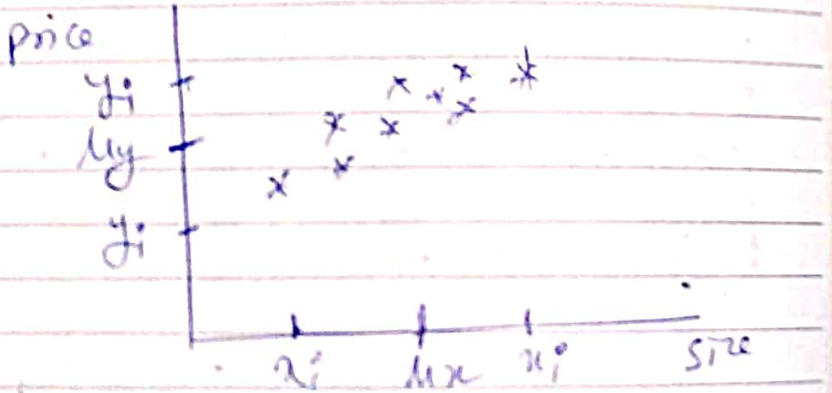
$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

where $\mu_y$ = Mean of y

$x\uparrow \; y\uparrow$ = [ ] +ve value

$x\uparrow \; y\downarrow$ = [ ] -ve value




Covariance → the change in one variable is equal to change in another variable

- $\text{cov}(x, z) = (+ve) * (+ve) = +ve$

$\text{cov}(x, y) = (-ve) * (+ve) = -ve$

Covariance

$x\uparrow \; y\uparrow$ = +ve How much positive

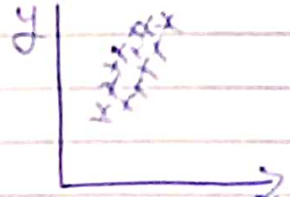
$x\uparrow \; y\downarrow$ = -ve How much negative

Type of Covariance

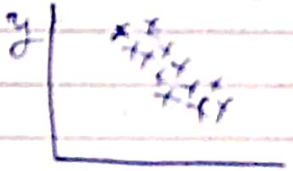
1 — Positive Covariance } Positive Covariance (moving both the value in same direction $x\uparrow \; y\uparrow$)

2 — Negative Covariance }

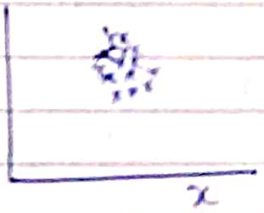
if




$\text{cov}(x, y) > 0$ positive




$\text{Cov}(x, y) < 0$ (less than zero) Negative




$\text{Cov}(x, y) = 0$ (is zero)

Then no relation

# Pearson Correlation Co-efficient

it only happen for linearly data

1- Co-variance = $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)*(y_i-\mu_y)$

2- Pearson CC = $\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

$\sigma_x$ = Standard deviation of x
$\sigma_y$ = Standard deviation of y

This is use in feature selection process.

3- Spearman ~~Correlation~~ Rank Correlation Coefficient.

Covariance Revise

| | Height(X) | Weight(y) |
|---|---|---|
| Cov(x,y) = | x ↑ | y ↑ (increasing) |
| | x ↓ | y ↓ (Decreasing) |

What is the Relationship both x, y is say the "Direction of Relationship"

"

Covariance is just find The direction of Relationship of How positive, and How Negative"

# Pearson Corelation Coefficient

Actually help to find the both Pearson CC

1- Strength
2- Direction of Relationship

x ↑ y ↑

* Cov(x,y) says only the Relationship of two Random variable, How much positive, but never said How much Corelated?

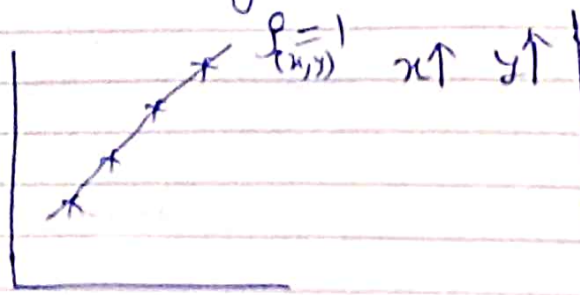
So this problem is solve by Pearson Corelation Coefficient.

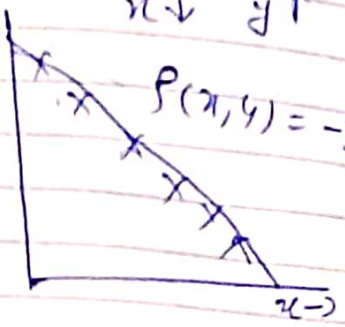
$\rho$ is lies $-1 \leq \rho \leq 1$

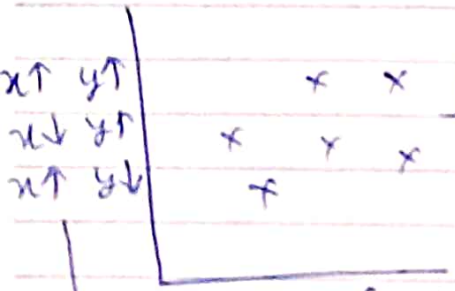
" $\rho$ (Pearson CC) is Range from -1 to 1 "

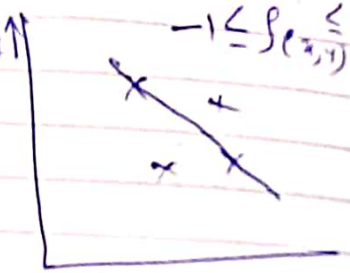
$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Range −1 to 1




(Linearly happend data)

$\rho(x,y) = 0$

These Technique use by feature Selection.

Suppose we have Three feature columns if

$x_1$ $x_2$ $x_3$

$x_1$ $x_2$ is $x_1 \uparrow$ $x_2 \uparrow$

$x_1 - x_2 \geq 1$

Then Remove one one feature. Now

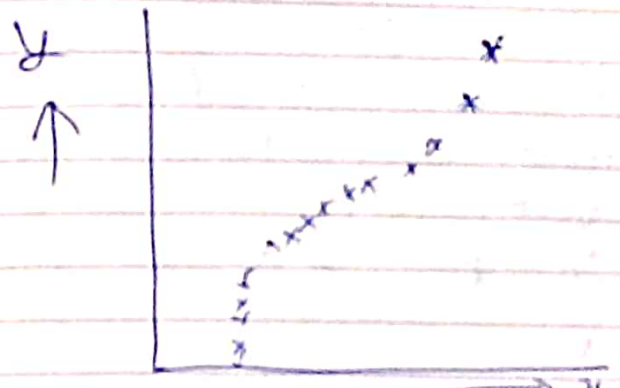
# Spearman's Rank Correlation Coefficient

DATE / /

This is worked on non-linear Structure Data.

Spearman Rank CC = 1

Pearson CC Result = 0.88




– its Non-linear image data

formula

$$\rho(x_{rank}, y_{rank}) = \frac{Cov(x_{rank}, y_{rank})}{\sigma_{rank} * \sigma_{rank}}$$

finding spearman find the $rank_x$, $rank_y$,

DATE / /

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Check the wiki

Spearson Rank C.C is help to find the non-linearty data to How Corelated to each if very close then result will be 0.96 or 0.85 kind and very low −0.0076 then Negatively

# Feature Scaling

- Standrization
- Normalization

**Normalization** → In this approch we Will Scale down the values of features between 0 to 1

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Using MinMaxScaler in Sklearn

few data sample

| Class | Alcohol | Malic |
|---|---|---|
| 1 | 14.23 | 1.71 |
| 1 | 13.20 | 1.78 |
| 1 | 13.16 | 2.26 |
| 1 | 14.37 | 1.95 |
| 1 | 13.24 | 2.59 |

We can Alcohol is very high so scale down

y have also malic column is Scale down using MinMaxScaler function of scikit learn.

Output after applied MinMaxScaler()

[0.860532, 0.233201]

[0.78 3261, 0.21834]

It mean both the column Alcohol & malic is Scale down to 0 to 1 form.

# Standardization (Z-score Normalization)

Here all the feature will be transformed in such a way that it will have the properties of standard normal distribution with mean ($\mu = 0$) and Standard deviation ($\sigma$) = 1

$$Z = \frac{x - \mu}{\sigma}$$

Using Sklearn Using StandardScaler
Value Considering the Mean=0
$\sigma = 1$
of both column (Alcohol & Malic)

[ 0.06732, -0.542316 ]
[ 1.03467, -0.326476 ]

Basically →
When we use lot in machine learning term is StandardScaler because perform well.
But MinMaxScaler is bit not much equal to StandardScaler so in ML term StandardScaler it's use more.
DL terms they use CNN (convolution NN) uses minMaxScaler because it's converted 0 to 1 form.
Why DL is Usage MinMaxScaler because image data to represent 0 to 255.
If grey scale is 0 to 1 to good to go with.

DATE __/__/__

Basically, its scale down the feature, it thing happen through the gradient decedant. Actully reduce the weights & gets the new weights.

Scaling feature
- CNN
- Linear Regression
- KNN
- K mean
- ANN
- K-nearest neighbour

Scaling Perform

No Scaling
- Decision tree
- Random forest
- XGBoot Boosting All technique
- Bagging tree

don't Perform Scaling based on decision they doing it. branch create

# Skewed Distribution

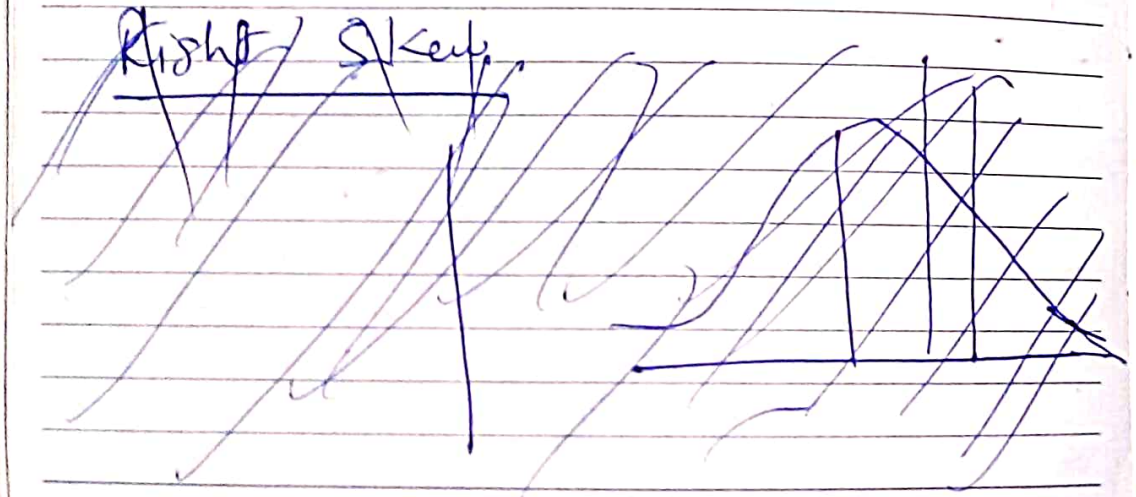
Q1- What is the Right skew distribution?

Q2- What is the Left Skew distribution?

Q3 - Classical example of Right Skewed distribution & left skewed distribution?

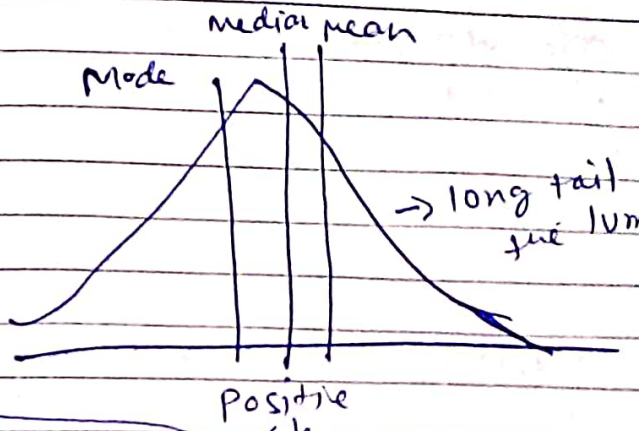
Q4 - What is Relationship bet^n mean, median, mode with Right skewed & left skewed distribution?

Right Skew



DATE / /

## Right Skewed

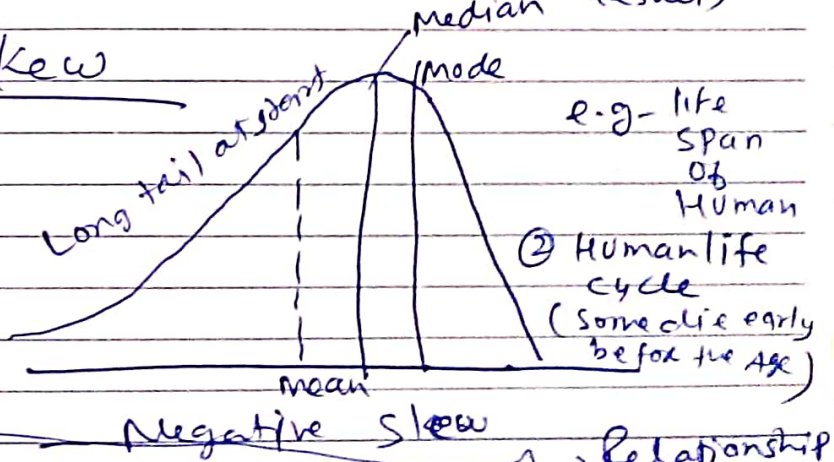



Right Skew mean > median > mode

Right hand side is tail aligned to x

e.g- wealth distribution
- it mean lot of peple wealth is more like Ambani, Tata to normal one
- cricket score (few score good others no)
- Exam Results (many low, less high result)

## Left Skew




e.g- life span of Human
② Human life cycle (some die early before the Age)

Left Skew = mode > median > mean → Relationship


Normal distribution

e.g age distribution
weight "
height "
IRIS "

Q1 What is sampling ?
Q2 What are different different Sampling technique ?

Q2 ~~What are differ~~

# Sampling techniques

mainly 4 types is useful →

1- Simple Random Sampling →

In Childhood, in cricket one person randomly choose number of batting position of every player, while another person pointing out the number by hand backside of 1st person, This is called simple Random sampling

2 - Stratified random Sampling

In school prayer in the morning there is particular line for every class and different line for boys and girls, so here there are n groups of particular class student. And here if teacher choose one boy form every class or one girl from every class this is Stratified Random sampling

3 - Cluster random Sampling →

Assume in ~~sudden~~ school suddenly one gest come and principal ordered every teacher to all students to come at play ground, this is an immediate order so teacher and student can't make proper line according to class wise, Here there can be or can't be students of different class are stood in one line, so here every line is called as cluster, How teacher order one student from every line, This is called cluster Random Samplig

4→ Systematic Random Sampling →

Assume everyday in school assembly, techers order students by their roll number, from every Line (here it can be stratified or cluster Random sampling), so take randomly from group, after nth position, This type of samplig is a systematic Random Sampling.

Whelt is aveage length of Shark in a Sea?

(1) Hypothesis testig → Let's assume we have a Theoritical value of shark length around 2 feet

H0 (Null hypothesis): The population mean (2feet) is same as sample mean.

Ha (Alternate hypotheis) The population mean is different

Set the P value as 0.05 and collect the shark length data by using Z-Test / T-test we will conclude whether we can accept the Null or Reject the Null.