__/__/____

Elastic Net [L_1 & L_2 Norm)

Cost fun^c : $\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x)^{(i)} - y^{(i)}\right)^2 + \underbrace{\lambda_1 \sum_{i=1}^{m} (slope)^2}_{\text{Ridge}} + \underbrace{\lambda_2 \sum_{i=1}^{m} |slope|}_{\text{Lasso}}$

Statistics

→ find the average size of shark throughout the world

→ Amazon Big Billion Day save → which month should you choose.

Life cycle of Data science project.

Requirement Gathering

[EDA] → [FE] → [FS] → [Model Training] → [Hyper parameter Tunning]

- EDA ↓ Exploratory Data analysis
- FE ↓ Feature Engineering
- FS ↓ Feature selection
- Model Training ↓ Training with ML algorithm
- Hyper parameter Tunning ↓ [Deploying]

Statistics will be used almost every stage in the life cycle of a Data science project

Sonai

_/_/___

Average ⇒ Measure of central Tendency.

⇓

Descriptive stats

Statistics:- Statistics is the science of collecting, organising and analysing the data.

Data:- "facts or pieces of information"

Statistics

Descriptive stats [EDA + FE] | Inferential stats

It consists of organising and summarizing the data.

- Histogram
- Barchart
- pie
- Distribution
- Candlestick
- Box plot.

It consist of collecting sample data and making conclusion about population data using some experiments.

Population (N) & Sample (n)

Sampling Techniques

- Simple Random sampling: Every member of the population (N) has an equal chance of being selected for your sample (n).

Stratified Sampling:- Strata → Layers → charaters → Group.

_/_/__

Stratified Sampling :-

Population

↓

<18	≥18 (voting)
	↓ Random Sampling

- Systematic Sampling :- ~~St~~ select every n^{th} individual out of population (N)

- Convenience Sampling: Only those who are interested in the survey will only participate.

  ↓

- Survey Regarding new technology → Convenience Sampling

Variable: A variable is a property that can take any values.

Two different types of Variable

- Quantitative variable → Measured Numerically {Mathematical operation}.
  ex:- Age, weight, height, rainfall, temp, distance.
- Qualitative variables → Categorial Variables
  Eg:- Gender, Types of ~~Variable~~ movies.

_/_/_

Quantitative Variable

Descrete Variable

eg - whole number
No. of Bank account
No. of children

Continuous Variable.

Eg:- Height, weight, ages, Rainfall.

→ ~~Ass~~ Gender — Categorical
→ Marital Status — Categorical
→ Ganga River length — Continuous
→ Movie duration — Continuous
→ Picode — discrete.
→ IQ — ~~continuous~~ descrete

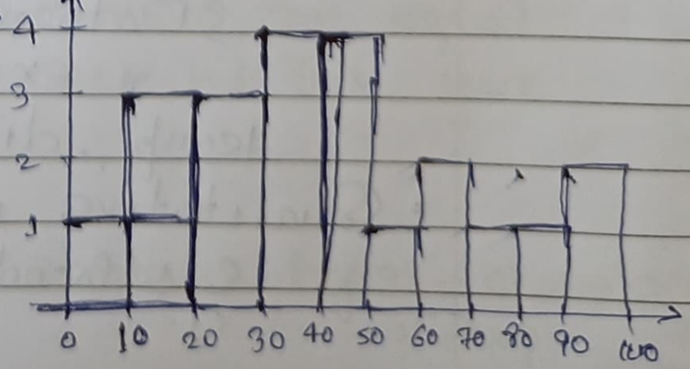
Histogram:-

Ages:- = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

① Sort the numbers
② Bins → No. of groups
③ Bin size : size of Bins

min = 10
max = 100

Bin size = $\frac{100}{10}$ ← No. of Bins
= 10



__/__/____

Measure of Central Tendency:

- Mean
- Median
- Mode

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position.

population (N) Sample (n)

population mean (μ) = $\sum_{i=1}^{N} \frac{x_i}{N}$

$N >> n$

sample mean ($\bar{x}$) = $\sum_{i=1}^{n} \frac{x_i}{n}$

Practical Application (Feature Engineering)

Age	Salary	Family size
–	NAN	–
–	–	NAN
NAN	–	–
–	–	NAN
–	NAN	–

NAN - Not a number

we have some datas that are missing, if we remove whole row data if any one data is missing in row, we will loose data or information that's why measure of central tendency is important, we can replace missing data with one of the measure of central tendency.

_/_/___

Median

$\{1,2,3,4,5\}$ $\{1,2,3,4,5,100\}$

$\bar{x} = 3$

$\bar{x} = 19.16$

↳ due to outlier

Steps to find out median:-

- Sort the number
- find the central number
  - → if the no. of elements are even we find the average of central elements
  - → if the no. of elements are odd we find the central element

$\{0,1,2,3,4,5,6,7,8,100,120\}$

Median = 5

Mean = 25.6

If outlier is present, Use Median.

Mode: most frequent occuring element

$\{1,2,2,3,3,3,4,5\}$

↳ Mode is generally used for categorical data sets.

_/_/___

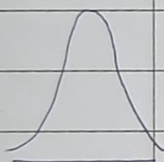
- Measure of Dispersion :-
  - Variance (σ^2)
  - standard deviation (σ)

- Variance :

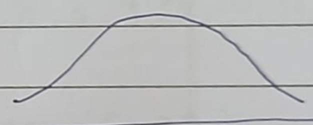
Population Variance (σ^2)	Sample Variance (s^2)
$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

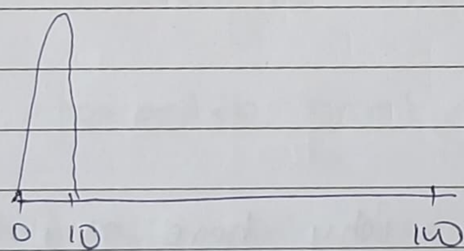
$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{1, 2, 3, 4, 50, 60, 70, 100\}$

$(\text{Variance})_A$ $(\text{Variance})_B$

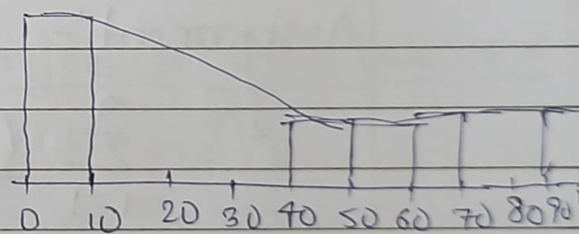


$(\text{Variance})_B > (\text{Variance})_A$





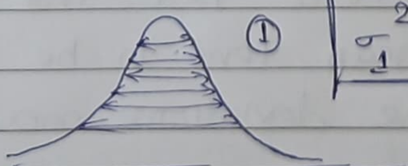
spread is less



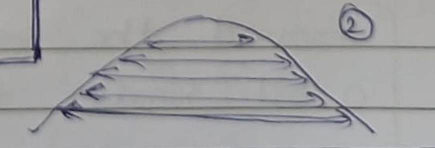
spread is more

$\sigma^2 = [$

- As spread increases, Variance increases.
- As spread is less, Variance is less.



① $\boxed{\sigma_1^2 < \sigma_2^2}$ ②



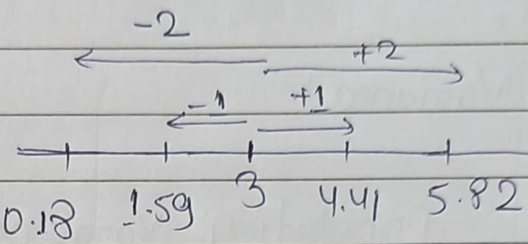
Standard Deviation :-

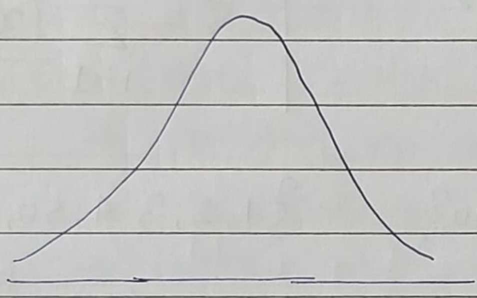
$\{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma^2 = 2$

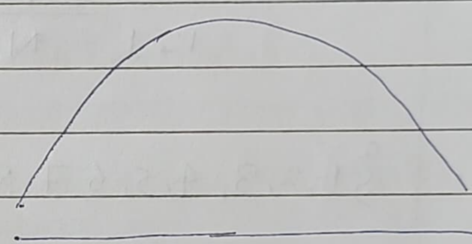
$\sigma = \sqrt{2} = 1.41$





σ^2 less

σ less



σ^2 high

σ high.

Assignment :-

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

why here $n-1$?

Since, we compute deviations from the mean of all the items in the population, rather than the deviation from ~~which the~~ sample mean ~~was drawn~~.

However population mean is generally unknown, so the sample mean is used in its place.

It is a mathematical fact that the deviations around the sample mean tend to be ~~smaller than~~ a bit smaller than the deviations around the

__/__/____

population mean and that dividing by $n-1$ rather than n provides exactly the right correction. it is also known as Bessel's correction.

* Percentiles and Quartiles : A percentile is a value below which a certain percentage of observations lie.

99 percentile = It means the person has got better marks than 99% of the entire students.

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

$$\begin{array}{cccccc} 2 & 2 & 3 & 4 & 5 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 0 & 1 & 2 & 3 & 4 & 5 \end{array}$$

What is the percentile rank of 10

$$\text{percentile rank of } x = \frac{\#\ \text{No. of value below } x}{n}$$

$$\text{percentile rank of } 10 = \frac{16}{20} = 80 \text{ percentile.}$$

$$\text{percentile rank of } 8 = \frac{9}{20} = 45\%$$

* What is the value that exists at 25 percentile

$$\text{Value} = \frac{\text{Percentile} * n}{100}$$

Index start with '0'

$$\frac{25}{100} \times 20 = 5^{th} \text{ Index}$$

$\boxed{O/P = 5.}$

_/_/_

5 number Summary :-

- Minimum
- First Quartile
- Median
- Third Quartile
- Maximum

} Using Box plot ⇑ Remove the outlier

{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

0 1 2 3 4 5 6 7 8 9th 10th 11th 12th 13th 14th 15th 16th

[Lower fence ⟷ Higher fence]

Lower fence = $Q_1 - 1.5(IQR)$

$IQR = Q3 - Q1$

⇓

Inter Quartile Range.

$Q1 = \frac{25}{100} \times (n+1) = \frac{25}{100} \times 21 = 5.25$ ⇒ Index

O/P ~~5.25th index~~ = 3 [Average of 5th & 6th index value]

$Q3 = \frac{75}{100} \times 21 = 15.75$ index $= \frac{8+7}{2} = 7.5$

Lower fence = $3 - (1.5)(4.5) = -3.65$

Higher fence = $7.5 + (1.5)(4.5) = 14.25$

[-3.65 ⟷ 14.25]

so 27 is outlier in the above data set.

__/__/____

five number Summary :-

minimum = 1

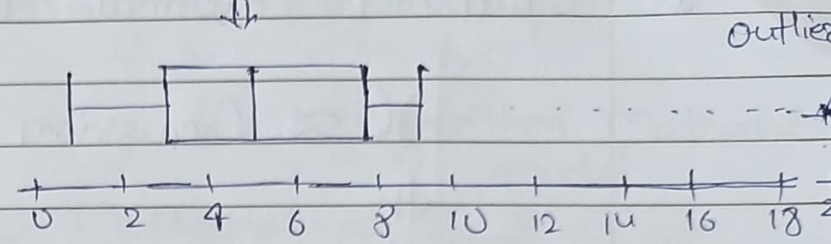
Q1 = 3

median = 5

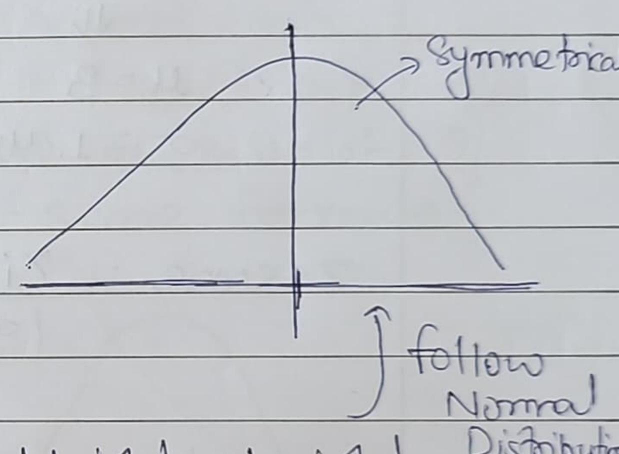
Q_3 = 7.5

maximum = 9

Box plot

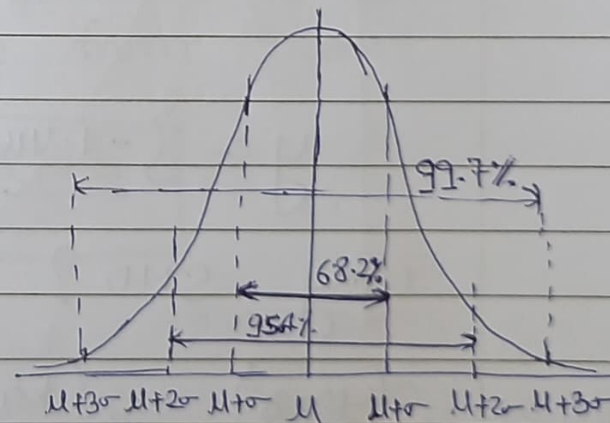


Gaussian / Normal Distribution



Age, Weight, height ↳ says (Domain Expert)

Emperical Rule of Normal Distribution



Q-Q plot for Distribution is Gaussian or Not ?

_/_/_

Standard Normal Distribution:

$X \approx$ Gaussian Distribution (μ, σ)

$\Downarrow$

$Y \approx$ Standard Normal Distribution $(\mu = 0, \sigma = 1)$

$x = \{1, 2, 3, 4, 5\}$

$\Downarrow$

$\mu = 3$

$\sigma = 1.41$

$$Z\text{-score} = \frac{x_i - \mu}{(\sigma/\sqrt{n})} \qquad \text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

Since we are applying Z-score formula for all the values so $n = 1$

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

$y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

Why SND? → Standardization

(year) Age	(kg) Weight	(cm) Height
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180

In Machine Learning → Math Required

⇓

Algorithm Mathematic model

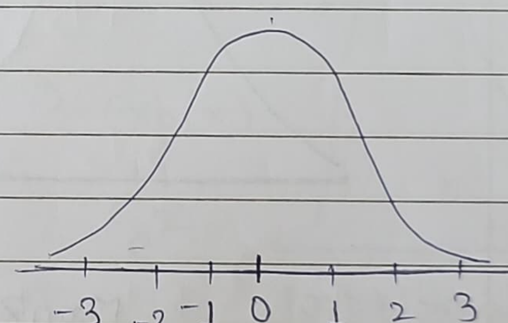
Time ↑↑ ⇐ Mathematical calculation ↵

With the help of standard scaling, we try to bring data into same scale because of which calculation become easy.

✓ ~~Nor~~ standardization {with the help of Z-score formula}

$\mu = 0, \sigma = 1$

almost All the data will range between [-3 3]



because almost 99.7% of data is between 3 standard deviation from mean.

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

✓ Normalization :- you decide the Range in between all the data will lie.

[lower scale ↔ Higher scale]

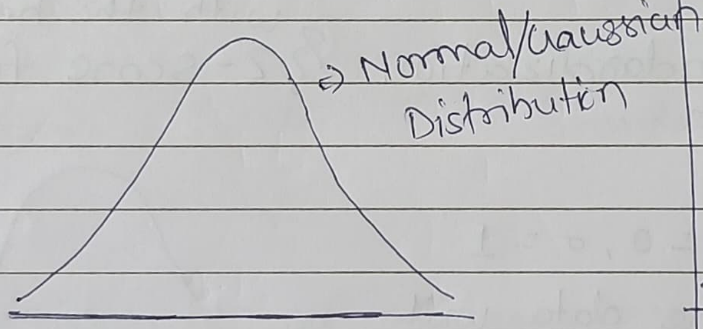
_/_/_

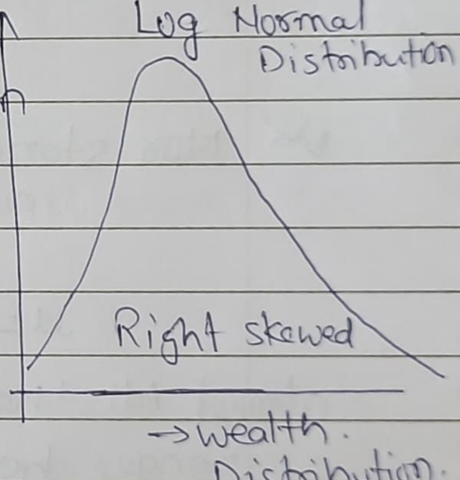
- Min-Max Scaler [0 - 1]

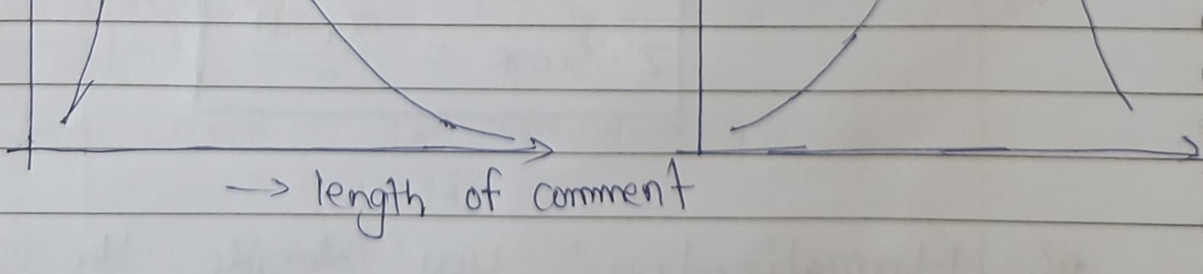
$$\boxed{x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}}$$

↳ Application: pixels values that lie b/w 0-255 can be brought between 0-1 scale.

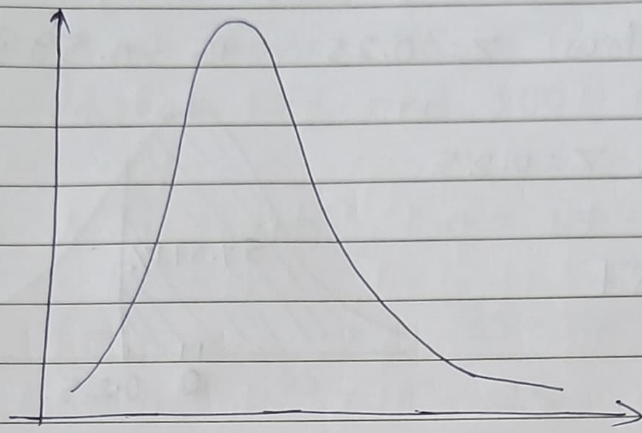
★ Log Normal Distribution:-



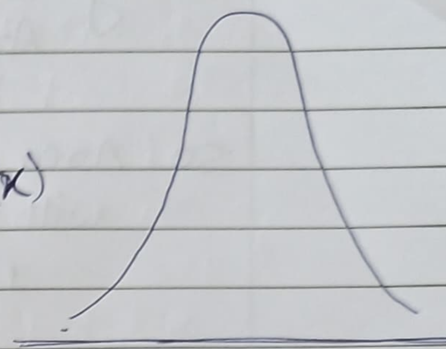




from Ascending order given the relation of mean, median & Mode.

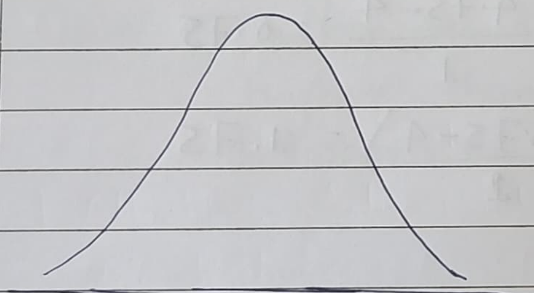


$y = \ln(x)$

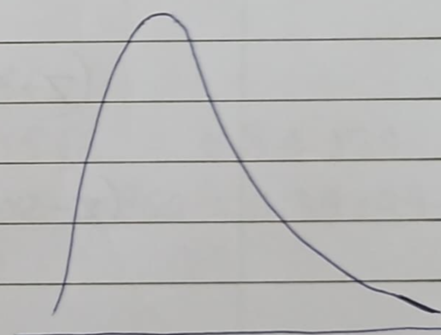


$y \approx$ Normal Distribution

$x \approx$ log Normal Distribution

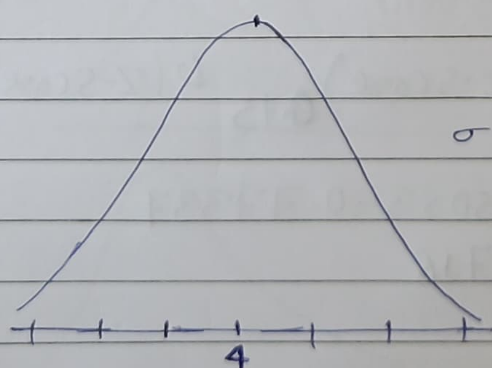


$\Rightarrow$ $\exp(x)$ or e^x



$x \approx$ Normal distribution (μ, σ)

log Normal Dist



$\sigma = 1$

What is the percentage of data that falls above 4.25?

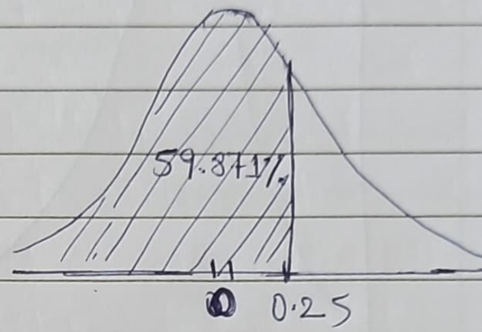
$$\text{Z-Score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Using z-table we can find the area left of $z = 0.25$.

_/_/__

Using Z-table.
Area below $Z = 0.25$ is 0.59871

So Area above $Z = 0.25$ will be:



$1 - 0.5987$

$= 0.4013\%$

Q Percentage of data lies b/w 4.75 & 5.75 if mean is 4 & median is 1.

$$(Z\text{-score})_{4.75} = \frac{4.75 - 4}{1} = 0.75$$

$$(Z\text{-score})_{5.75} = \left(\frac{5.75 - 4}{1}\right) = 1.75$$

Area below/left of Z-score 0.75 = 0.77337

Area below/left of Z-score 1.75 = 0.95053

So, Area b/w $(Z\text{-score})_{0.75}$ & $(Z\text{-score})_{1.75}$

$= 0.95053 - 0.77337$

$= 0.17716$

Q In India the average IQ is 100 with a standard Deviation of 15. What is the percentage of population would you expect to have ~~IQs~~ an IQ.

_/_/____

- lower than 85.
- Higher than 85.
- Between 85 and 100.

$\mu = 100 \qquad \sigma = 15$

$$(\text{Z-score})_{85} = \frac{85-100}{15} = -1$$

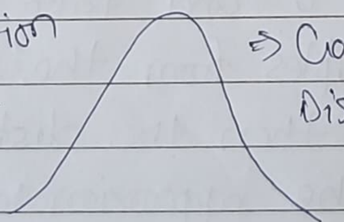
Using z-table:-

~~lower than 85: 0.84134~~
~~Higher than 8~~
lower than 85 : 0.15866
higher than 85 : $1 - 0.15866 = 0.84134$
Between 85 & 100 : $0.5 - 0.15866 = 0.34134$

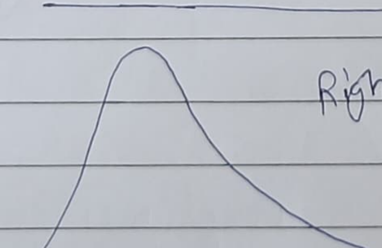
Central Limit Theorem :-

Population data

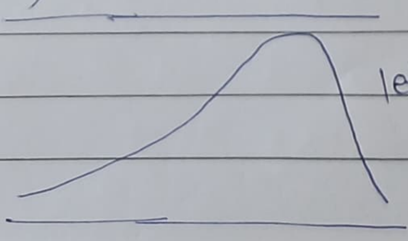
$\boxed{N}$



⇒ Gaussian/Normal Distribution.



Right skewed



left skewed.

$\{x_1^1, x_2^1, \ldots, x_n^1\} \to \bar{S}_1$

$\{x_1^2, x_2^2, \ldots, x_n^2\} \to \bar{S}_2$

$\{x_1^3, x_2^3, \ldots, x_n^3\} \to \bar{S}_3$

$\vdots$

$\{x_1^m, x_2^m, \ldots, x_n^m\} \to \bar{S}_m$

$\boxed{n}$ ↓ size of sample

$\boxed{m}$ ↓ No of sample.

Central limit theorem states that Given distribution may or may not be Gaussian/ Normal distribution of population data.
If we ~~at~~ are taking samples of size $n \geq 30$ from the population then ~~and calculate~~ ~~the mean of different~~ ~~different~~ n the distribution of the sample means will be approximately normally distributed.

✓ Size of shark through out the world?

10 different Region $n \geq 30$

↳ Many Assumptions can be made.

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of sample means will be approximately normally distributed.

_/_/_

Probability: Probability is a measure of the likelihood of an event.

Mutual Exclusive Event

Two events are mutually exclusive if they can't occur at the same time.

- Tossing a coin • Rolling a dice

$$P(A \text{ or } B) = P(A) + P(B)$$ → Addition Rule for ME

Non mutual Exclusive events

Two events can occur at the same time.

- picking randomly a cards, two ~~cards~~ events "hearts" and "king" can be selected.

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$$

↳ Addition Rule for Non mutual exclusive events.

Multiplication Rule:-

Dependent events: Two events are dependent if they affect one another.

Multiplication Rule for Independent Events

$$P(A \text{ and } B) = P(A) \times P(B)$$

_/_/_

- Probability of drawing a "orange" and drawing a "yellow" marble from the bag?

x x x x
o o o

$\rightarrow P(x) = \frac{4}{7} \rightarrow$

x x x
o o o

$P(o/x) \rightarrow$ Conditional probability

$3/6$

$$P(x \text{ and } o) = P(x) * P(o/x)$$

$$P(A \text{ and } B) = P(A) * P(B/A)$$

$$P(A \text{ and } B) = P(B) * P(B/A)$$

- Permulation & Combination

↳ with permutation, order matters. means all the possible arrangement matters.

$${}^{n}P_{r} = \frac{n!}{(n-r)!}$$

n = Total no. of object

r = No. of selection.

↳ with combination order doesn't matter. Only unique combination is considered.

$${}^{n}C_{r} = \frac{n!}{(n-r)!\,r!}$$

n = Total no. of object

r = No. of selection.

_/_/__

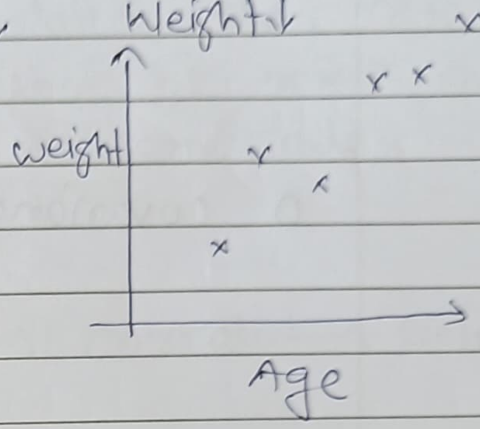
CO-Variance **:

↳ One of Feature selection is done with the help of covariance and co-relation

Age	weight
12	40
13	45
15	48
17	60
18	62

Age ↑ Weight ↑

Age ↓ Weight ↓



Can we Quantify the relationship x & y using mathematical equation.

$$\boxed{Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n-1}}$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})^2}{n-1}$$

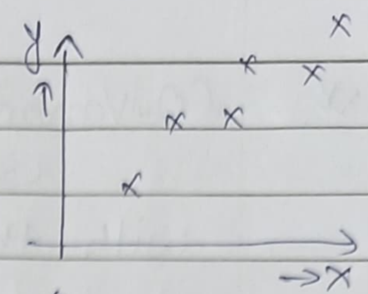
$$\Rightarrow Cov(x,x) = \sigma^2 = Variance(x)$$

$\bar{x} = 15$ $\quad \bar{y} = 51$

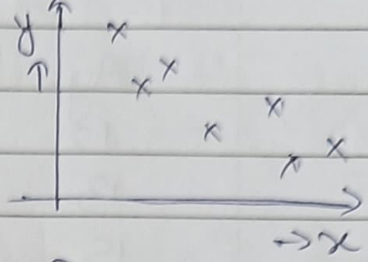
$$= \frac{-3 \times -8 + -2 \times -6 + 0 + 2 \times 9 + 3 \times}{4} = \frac{24 + 12 + 18 + 33}{4} = 21.75$$

$$Cov(x,y) = \frac{(12-15)(43-51) + (13-15)(45-51) + (15-15)(48-51) + (17-15)(60-51) + (18-15)(62-51)}{(5-1)}$$

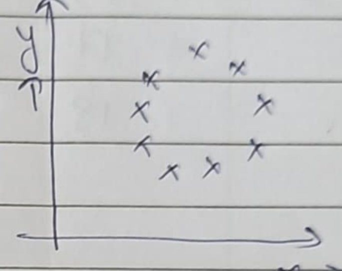
+ve Covariance X↑ y↑
X↓ y↓



-ve Covariance X↑ y↓
X↓ y↑



0 covariance No Relation



Pearson correlation coefficient (-1 to 1)

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

Why pearson correlation?
It is because Covariance of two variable can be any positive value or negative value. That's why we would be able to compare which variable more Correlated to a variable.
But in case of pearson correlation coefficient ~~we~~ Correlation values are always between [-1 1] so we would easily able to correlates the variables that which are more Co-rrelated to one variable.

_/_/____

Covariance

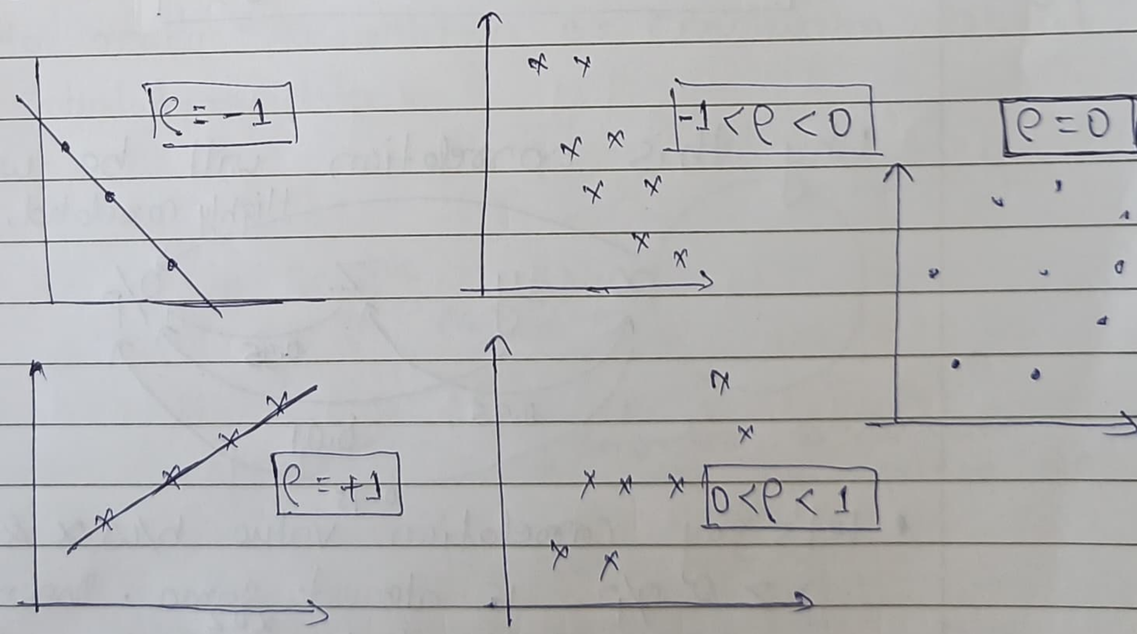
$$\begin{tikzcd} x \arrow[rr, bend left, "+150"] \arrow[r, bend right, "+100"'] & y & z \end{tikzcd}$$

As we see covariance b/w x & z is greater than covariance b/w x & y even the we would not be able to comment if x & z is more correlated than x & y or not. Because covariance b/w x & z and x & y might not be in same scale.

$$\begin{tikzcd} x \arrow[r, bend left, "+0.5"] \arrow[rr, bend right, "+0.7"'] & y & z \end{tikzcd}$$

Pearson correlation coefficient

Here we can visualize which variables are more correlated to each other because pearson correlation coefficient is always is same scale [-1 1]



Spearman Rank Correlation:-

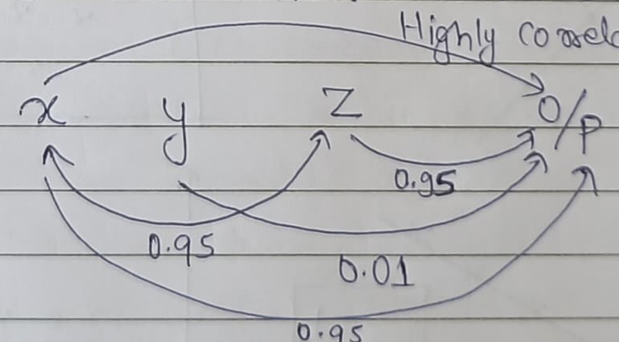
Pearson correlation coefficient holds good only for linear data.

Spearman Rank correlation holds good for Non-linear data.

x	y	R(x)	R(y)
10	4	4	1
8	6	3	2
7	8	2	3
6	10	1	4

$$\boxed{r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}}$$

Why this correlation will be used?



- let's say correlation value b/w x & z and z & o/p is almost same. Doesn't it mean that both feature x & z are same? yes we can remove one feature because both are same features.

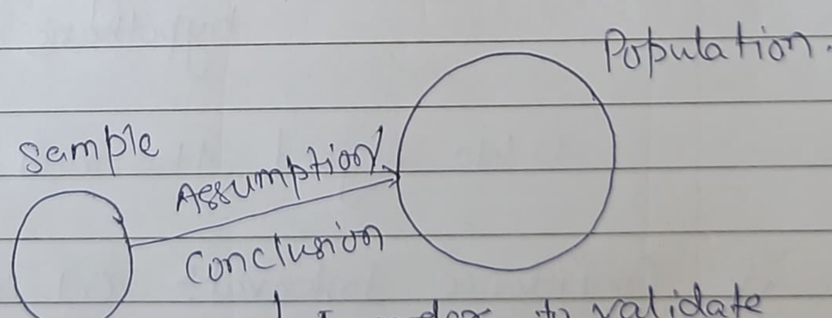
- let's assume x is highly correlated with o/p and y is very less correlated with o/p then we can neglect the feature y.

- correlation coefficients helps us in feature engineering.

Inferential statistics :-

Hypothesis Testing :-

we have population data and sample data. Using sample data we are going to make assumption or conclusion about population data.



↓ In order to validate the assumption we do

Hypothesis testing.