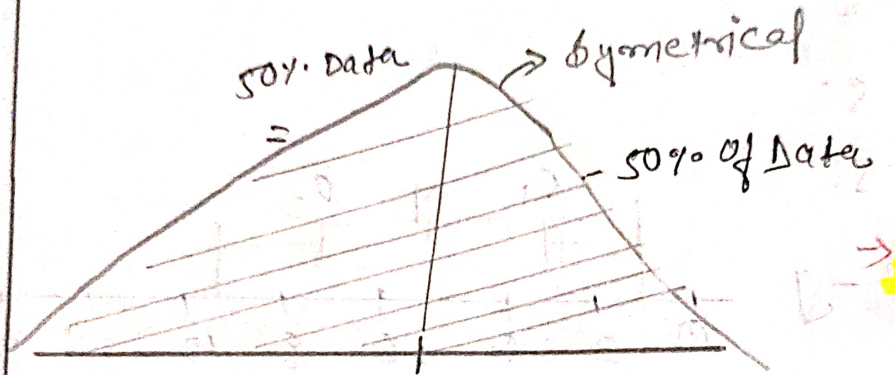
DATA SCEINCE
DAY-3
Statistics

Agenda

1. Normal Distribution
2. Standard Normal Distribution
3. Z-Score
4. Log Normal Distribution
5. Standardization And Normali[zation]

* Gaussian/Normal Distribution

Data on left = Data on Right



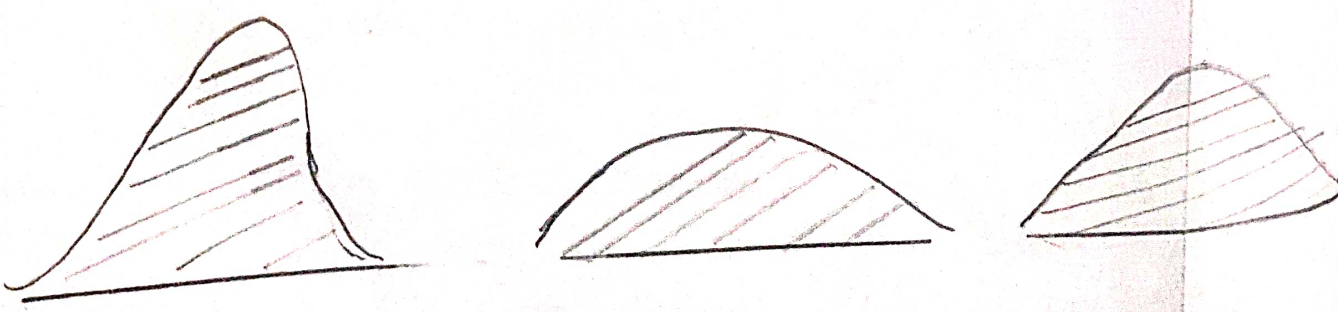
* KDE
→ Kernal Density estimator

* Normal Distribution is a Symmetrical bell-Shaped curve. The Area of curve is $\boxed{1}$
* Age, weight, Height follows Normal/Gaussian Distribution.

IRIS DATASET

⇓

Petal length, Sepal length, Petal width, Sepal width } It follows Gaussian Distribution



Emperical Rule of Normal Distribution

In this

⇒ within the first standard deviation between left & right 68% of the entire data will be available

⇒ Within the second standard deviation to right & left is around 95% of the entire Data falling in these region

⇒ Within the third standard deviation to right & left around 99.7% of the entire Data falling in this region.

This is called 68-95-99.7% Rule

are called Emperical Rule

* with the help of Q-Q Plot we can check whether the Distribution is Normal or not.

Standard Normal Distribution

Lets take variable x = Guassian Distribution (μ, σ)

then we transform x to

⇓

$$y = SND\,[\mu = 0, \sigma = 1]$$

Z-Score Formula

$$Z = \frac{x_i - \mu}{\sigma / \sqrt{n}}$$, where, $\frac{\sigma}{\sqrt{n}}$ is called Standard Error

we will take $n = 1$ because we are going to apply z-score formula to each value of x

eg:-

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

$\mu = 3 \quad \sigma = 1.41$

then, $Z = \frac{x_i - \mu}{\sigma}$

For $i = 1$ $\quad \frac{1-3}{1.414} = -1.414$

For $i = 2$ $\quad \frac{2-3}{1.414} = -0.707$

For $i = 3$ $\quad \frac{3-3}{1.414} = 0$

For $i = 4$ $\quad \frac{4-3}{1.414} = 0.707$

For $i = 5$ $\quad \frac{5-3}{1.414} = 1.414$

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

Why? we do SND

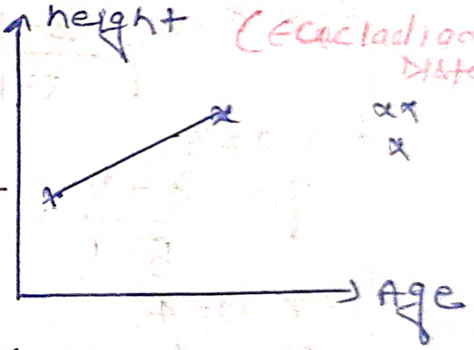
Age (years)	Weight (Kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
28	80	175

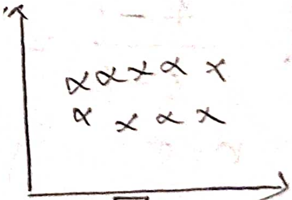
* Since units are different
* Value also differ very high

So, Mathematical Calculation take time ↑↑

* We will try to scale the ~~ar~~ entire dataset in same & this Process is called Standaziation.

[we will do so to bring the value in same scale so, that our calculation will be easy.]



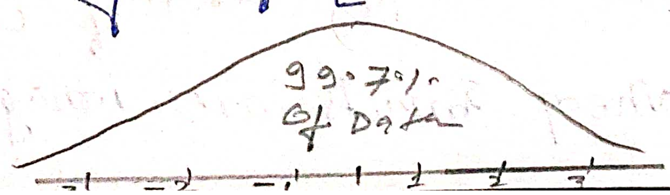


* $\boxed{\text{Feature Scaling}}$

Standardization $[\mu = 0, \sigma = 1]$

For this we will apply in each table. Z-score $= \dfrac{x_i - \mu}{\sigma}$

then all the value will be transformed to or in range of $[-3 \text{ to } 3]$



Normalization

In normalisation we give the range we try to normalise the value into lower scale & higher scale. → 0 to 1 or 0 1

① Min Max Scaler [0-1] $x = [1, 2, 3, 4, 5]$

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

x	y
1	0
2	0.25
3	0.50
4	0.75
5	1

for $x = 1$

$$y = \frac{1-1}{5-1} = 0$$

for $x = 2$

$$y = \frac{2-1}{5-1} = 1/4 = 0.25$$

for $x = 3$

$$y = \frac{3-1}{5-1} = 2/4 = 0.50$$

for $x = 4$

$$y = \frac{4-1}{5-1} = 3/4 = 0.75$$

for $x = 5$

$$y = \frac{5-1}{5-1} = 4/4 = 1$$

With the help of min-max scaler we will transform the value in rang [0 to 1]

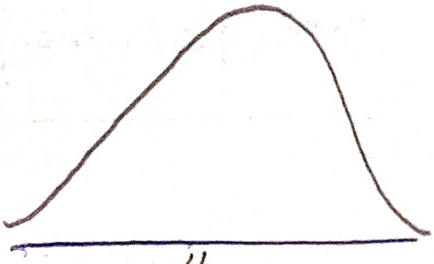
* We use this in Deep learning

This technique is called Normalization.

* Feature Scaling → ① Normalization
  ② Standardization.

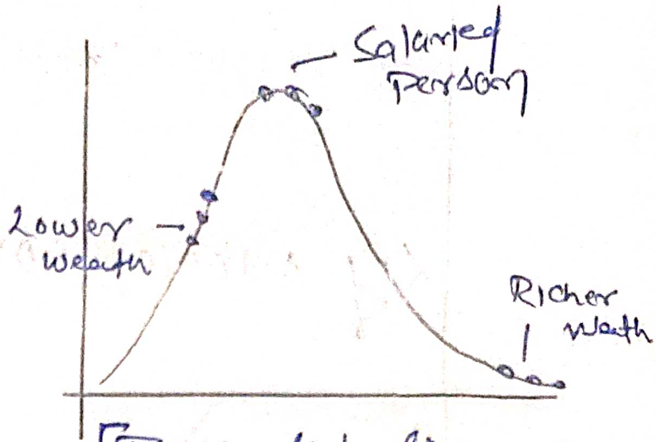
* In Deep learning [use Min Max Scaler]
* IN Machine Learning [use Standard Scaler]
* When Inputs are image use Min Max Scaler.

Log Normal Distribution



Normal Distribution

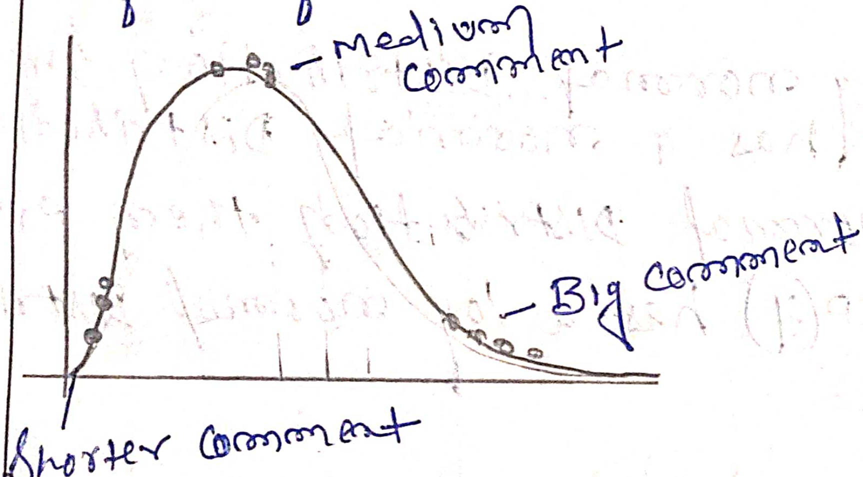
⇒



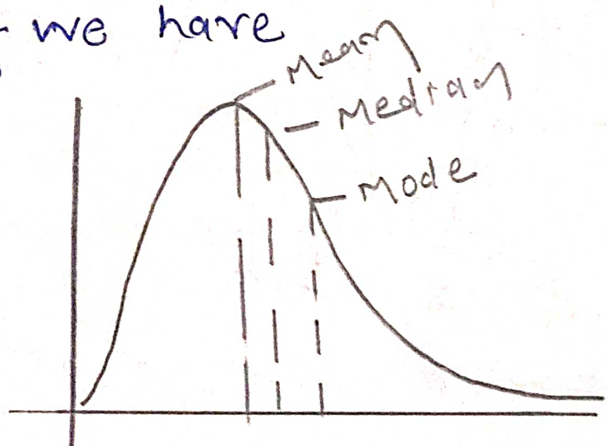
[Right-side Skewed Wealth Distribution

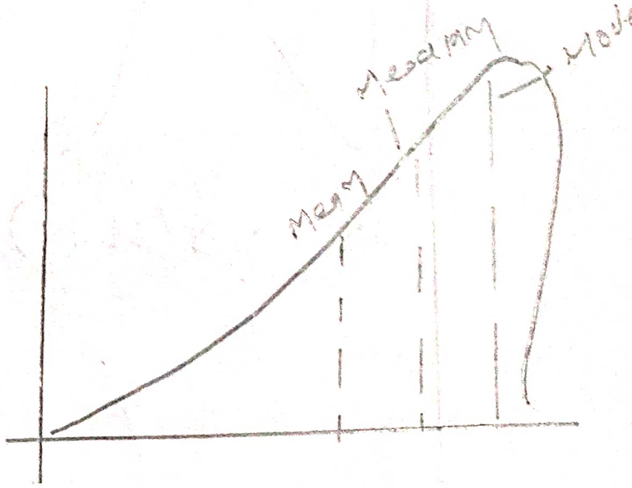
or

Length of Comments on videos



Q. if we have

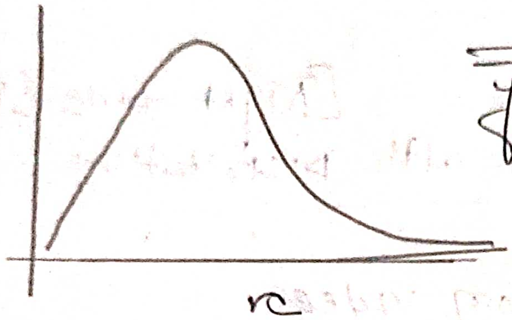




Q. What is the Relationship of Mean, median, Mode?

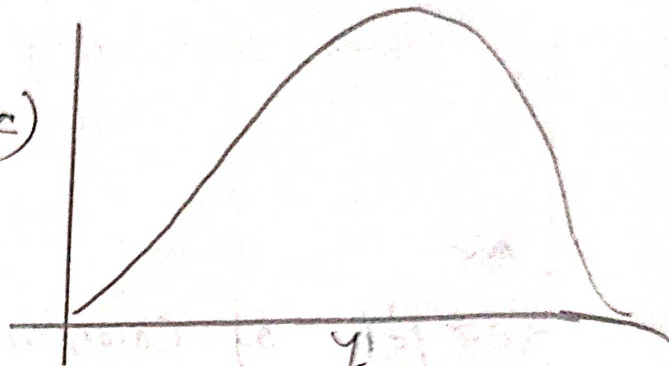
* For Right Skewed Distribution Mean > Median > Mode
For Left Skewed Distribution Mean < Median < Mode

Log Normal Distribution



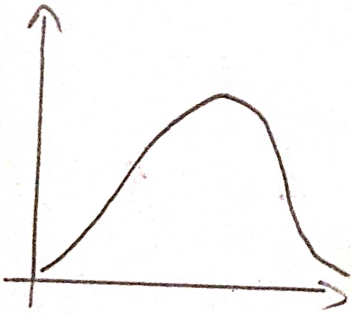
x

$\overrightarrow{y = \ln(x)}$



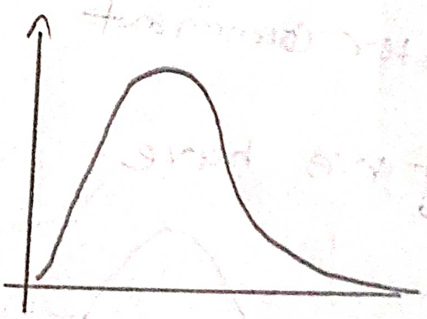
y

* if the log normal distribution then $y = \ln(x)$ has a normal distribution. if y has a normal distribution then exponent of y, x exp(y) has a log normal distribution



$x = ND(\mu, \sigma)$

$\xrightarrow[\text{exp}(x)]{\text{Antilog}}$



[only for Right Skewed Dist

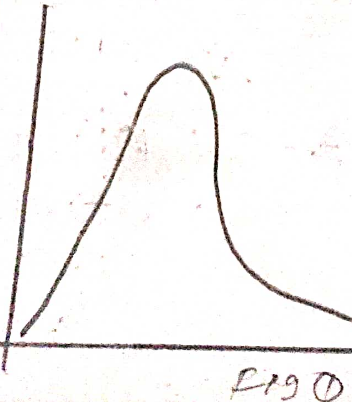


Fig ①

$\boxed{\Rightarrow y = \ln(x)}$

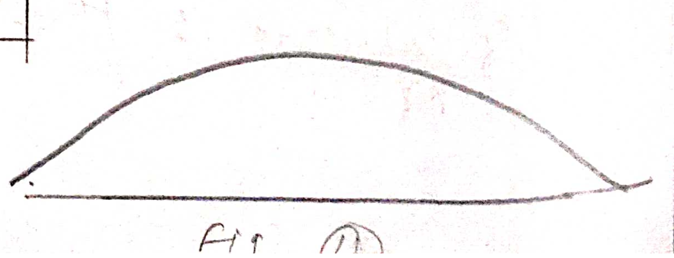
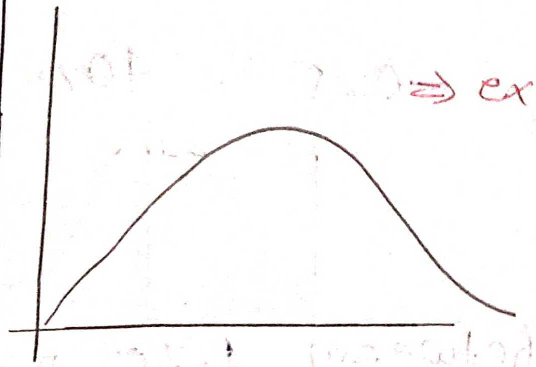


Fig ②

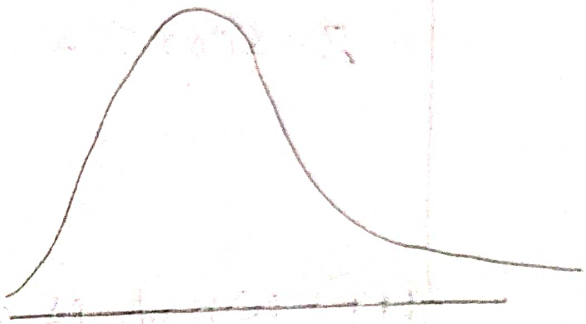
$x = \log$ Normal Distributed (μ, σ) – Fig (1)

* if we plot fig (1) & apply the log and as a output if we get Normal Distribution

then we can say fig (1) is Log Normal Distribution.



⇒ exp(x) ⇒



⇓ Log Normal Distributed

eg:- No. of Batsman scored high, marks in classroom.

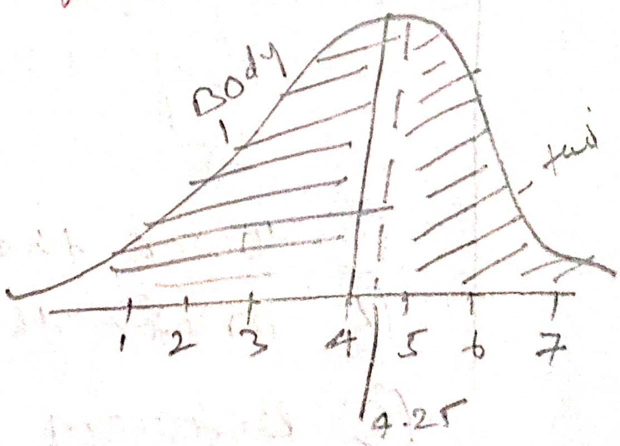
Problem Statement

Q. Lets a variable $x = \{1, 2, 3, 4, 5, 6, 7\}$ with mean $(\mu = 4)$ & standard deviation $\sigma = 1$

Q. What is the Percentage of score that fall above 4.25?

⇒ $Z\text{-Score} = \dfrac{x_i - \mu}{\sigma}$

$= \dfrac{4.25 - 4}{1} = 0.25$



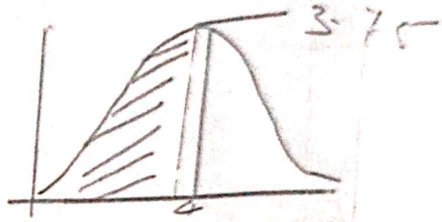
(i) Z-table [Area under the curve]

* The Percentage the curve from 1 to 4.25 is 59%.

* Percentage of scale falls above 4.25 is 41%.

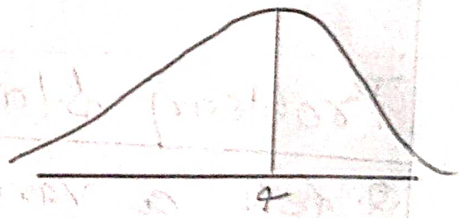
(ii) What is the Percentage of Area falls below 3.75?

$$Z\text{-score} = \frac{3.75-4}{1} = -0.25 \simeq 40\%$$



(iii) What is the Area between 4.75 & 5.75

$$Z_1 = \frac{4.75-4}{1} = 0.75 \qquad Z_2 = \frac{5.75-4}{1} = 1.75$$



Assignment

In India the average IQ is 100 with standard deviation 15. What is the percentage of population would you expect to have an IQ

(a) lower than 85 (Ans – 0.1587)

(b) greater than 85 (Ans – 0.8413)

(c) Between 85 & 100 (Ans – 0.3413)