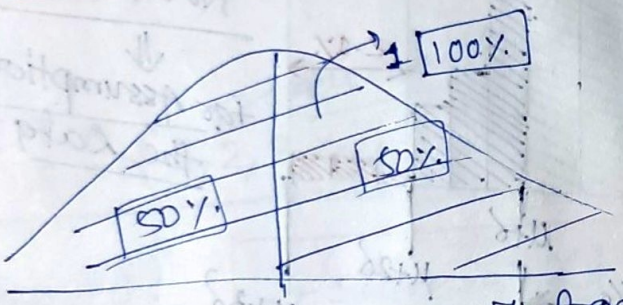
10th Sept '22 | DAY#3 | FSDS Bootcamp | Statistics

Agenda :

1. Normal Distribution
2. Standard Normal Distribution
3. Z-score.
4. Standardization & Normalisation.

#1 | Gaussian / Normal Distribution :-

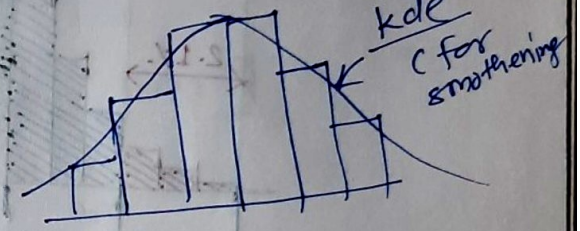


Symmetrical Distribution

(Age, weight, Height)

↑ by Domain Expertise

(Eg. - Doctors, Nutritionalist - etc.)



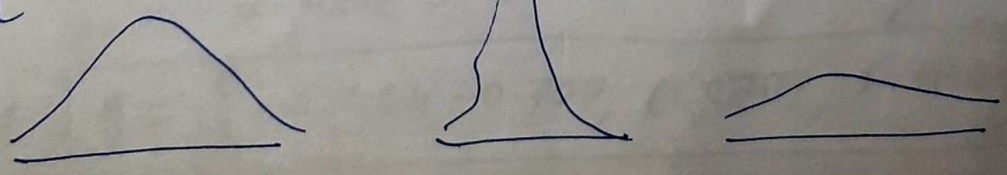
+ IRIS DATASET

It may be the 1st dataset where machine Algorithms have applied.

petal length, Sepal length, petal width, sepal width.

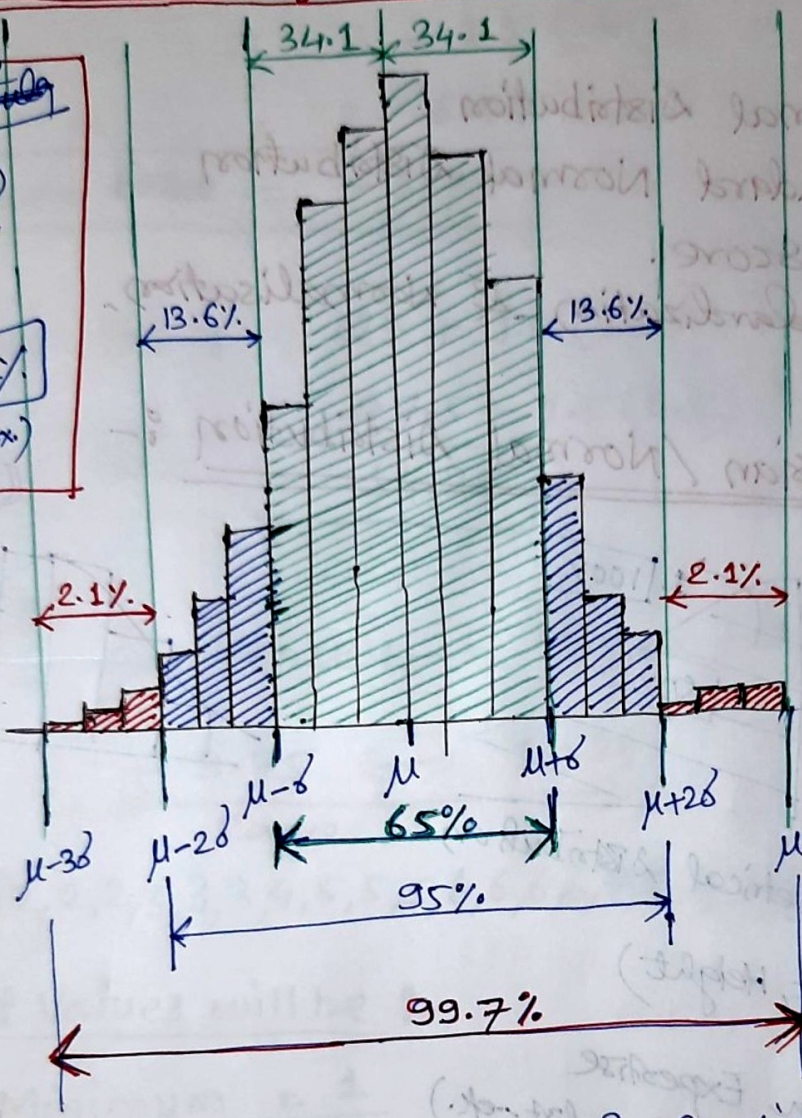
(they all follow ⇒ Gaussian Distribution)

Eg:



ⓐ Emperical Rule of Normal Distribution :-

Emperical Rule
(or Six-σ Rule)
⇓
68-95-99.7%
(*Approx.)



Here
$\mu \Rightarrow$ Mean
$\sigma \Rightarrow$ standard deviation.

Gaussian / Normal Distribution
⇓
for Assumption of the Data

Age = { } ⇒ Gaussian Distribution.

ⓑ Q-Q Plot :- To determine, Distribution is Gaussian or Not?

* Standard Normal Distribution :-

#2] Standard Normal Distribution :-

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

$\Downarrow$

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

& #3] z-score :-

Ex: $X = \{1, 2, 3, 4, 5\}$ ← Gaussian/Normal Distribution $[\mu = 3, \sigma = 1.41]$

formula :

$$z\text{-score} = \frac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\frac{\sigma}{\sqrt{n}}$ → Standard Error ⇓ mostly used in Inferential stats.

(∵ But here, $n = 1$, (as sample size is 1) (~~But here~~,

$\therefore$ ~~z-sco~~

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

$\therefore$ for $i = 1$, z-score $= \frac{1-3}{1.41} = -1.414$

$i = 2$, z-score $= \frac{2-3}{1.41} = -0.707$

My, $i = 3$ z-score $= 0$

$i = 4$ z-score $= 0.707$

$i = 5$ z-score $= 1.414$

$\therefore$ Standard Normal Distribution $(\mu = 0, \sigma = 1.41)$

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

P.T.O.

◎ why? SND.

#4 | * standardization {using z-score formula} ($\mu=0, \sigma=1$)

Age (years)	weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175

Here, for Mathematical Calculations, it will take more time.

∴ The standardization is done using z-score ($\mu=0$ & $\sigma=1$)

~~data~~ same scale

∴ To draw Age, weight, Height on same scale, we have to do standardization.

Standardization, generally ranges bet^n $[-3 \leftrightarrow 3]$

* Standardization is done To bring feature in the same scale

#4/ * Normalization → for scaling down the range/value. [Lower Scale 0 ⟷ Higher Scale 1]

① Min Max Scaler [0-1]

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

$x_1 = \frac{1-1}{5-1} = \frac{0}{4} = 0$

$x_2 = \frac{2-1}{5-1} = \frac{1}{4} = 0.25$

$x_3 = \frac{3-1}{5-1} = \frac{2}{4} = 0.5$

$x_4 = \frac{4-1}{5-1} = \frac{3}{4} = 0.75$

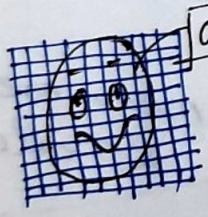
$x_5 = \frac{5-1}{5-1} = \frac{4}{4} = 1$

x	⇒	y
1	→	0
2	→	0.25
3	→	0.5
4	→	0.75
5	→	1

0 ⟷ 1

?? Apply? where ??

→ In deep learning



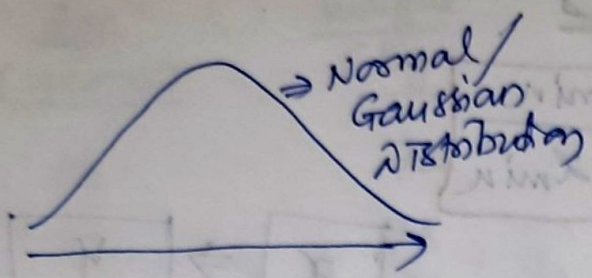
0-255 → 0-1 range. → By Normalisation.

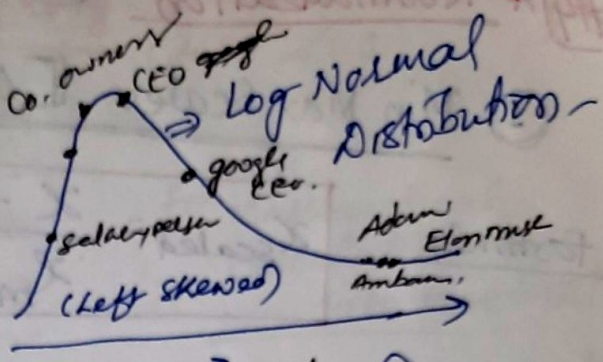
pixels

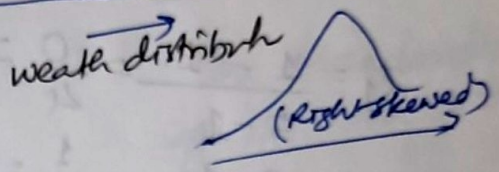
0-255 (range)

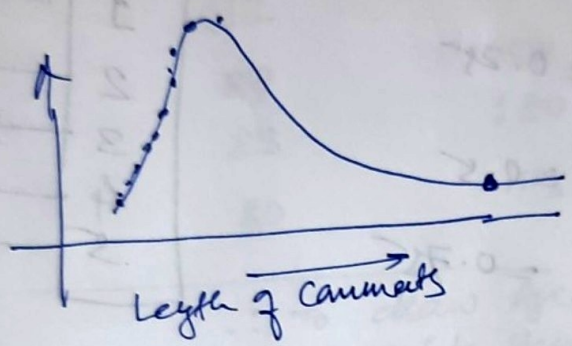
— we use Normalisation & Standardization for Feature Scaling.

Log Normal Distribution :-

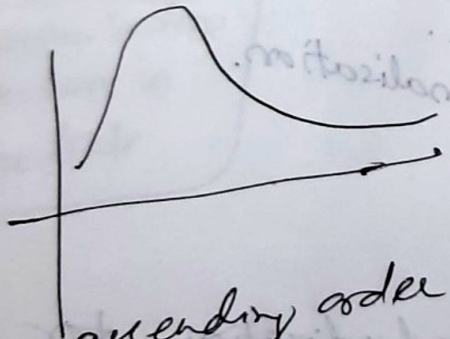


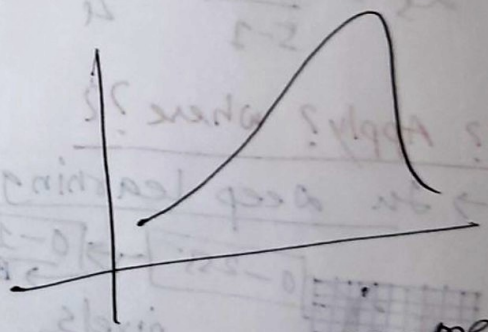






Que: Relⁿ betⁿ mean, median, mode, of two graphs shown below.

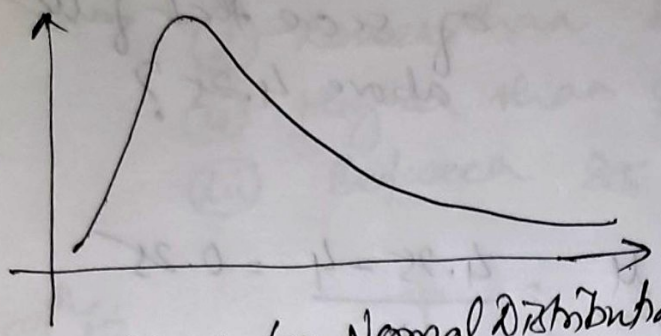




Q. from ascending order give Relⁿship of mean, median, & mode.

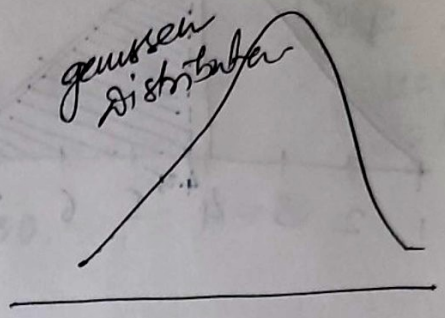
→ If the random variable X is log-normally distributed, then $Y = \ln(X)$ has a normal distribution

(∵ $\ln = \log_e(x)$ :: natural log)

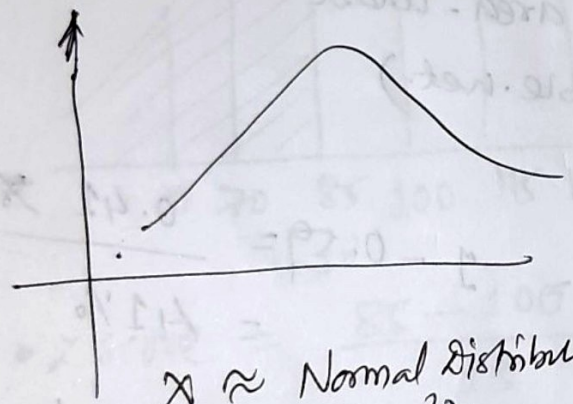


$X \approx$ Log Normal Distribution

$\approx$

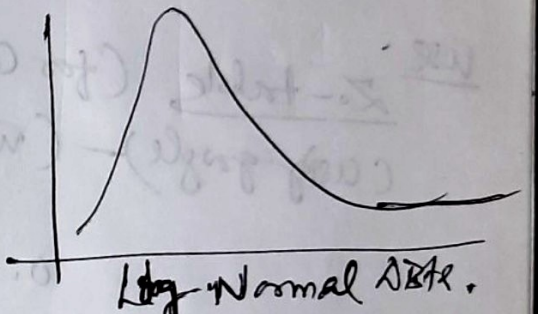


$Y = \ln(X)$



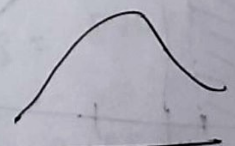
$X \approx$ Normal Distribution (μ, σ)

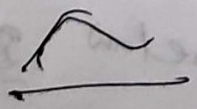
$\approx$

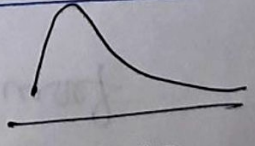


$\exp(X) \Rightarrow Y$

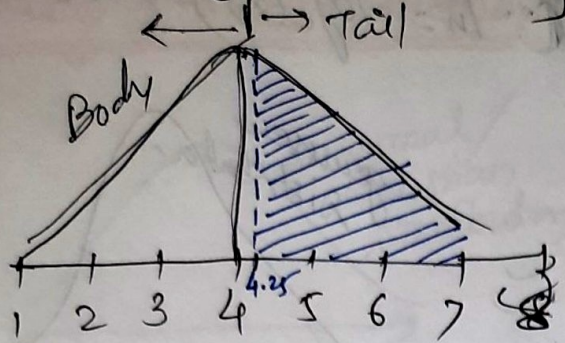
(Antilog)

$X \approx LND \Rightarrow \log_e(X) \Rightarrow$ 



$\Rightarrow \exp(X) \Rightarrow$ 

Que: $X = \{1,2,3,4,5,6,7\}$ $\mu = 4$, $\sigma = 1$ } assume



Que: what is the percentage of score that falls above 4.25?

Sol^n:- Z-score: $\frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$

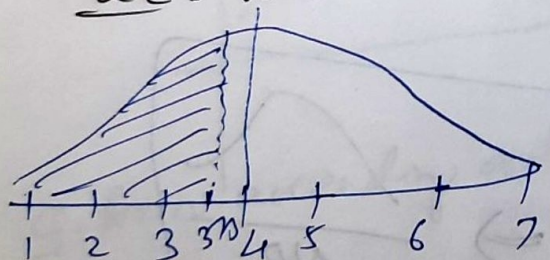
use: Z-table (for calculating area-under the curve) - (using google) - (www.ztable.net)

0.59

Area under curve = $1 - 0.59 = 0.41$

$= 41\%$

Q2: what is the %ge of score that falls below 3.75?



Z-score $= \frac{3.75 - 4}{1}$

$= -0.25$

from z-table $-0.25 = 0.40$

∴ 40% Area falls below 3.75.

Q3 - Bet 4.75 & 5.75 - Assignment

Assignment

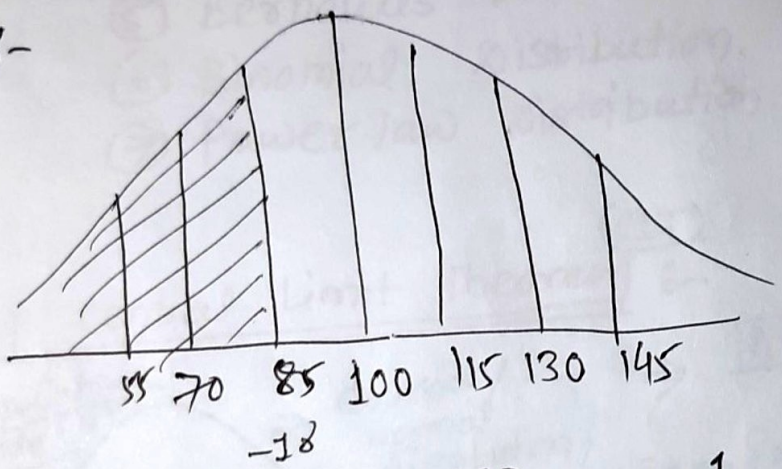
Que ① In India the average IQ is 100 with a standard deviation of 15. what is the percentage of population would you expect to have an IQ

Answers.

(i) Lower than 85 ⇒ 0.1587
(ii) Higher than 85 ⇒ 0.8413
(iii) Between 85 and 100, ⇒ 0.3413

$\therefore \mu = 100$
$\sigma = 15$

Sol^n:-



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

$\therefore$ using Z-table $\boxed{-1 = 0.1587.}$

$\therefore$ (i) Lower than 85 = 0.1587 = 15.87%

(ii) Higher than 85 = 1 - 0.1587
= 0.8413 = 84.13 %
$\because$ 100 is mean also.

(iii) Bet^h 85 & 100 = 0.5 - 0.1587
= 0.3413 = 34.13%