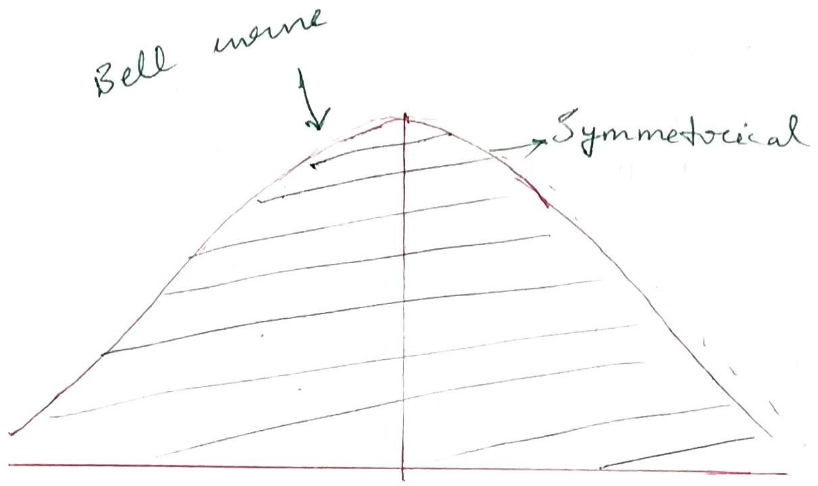
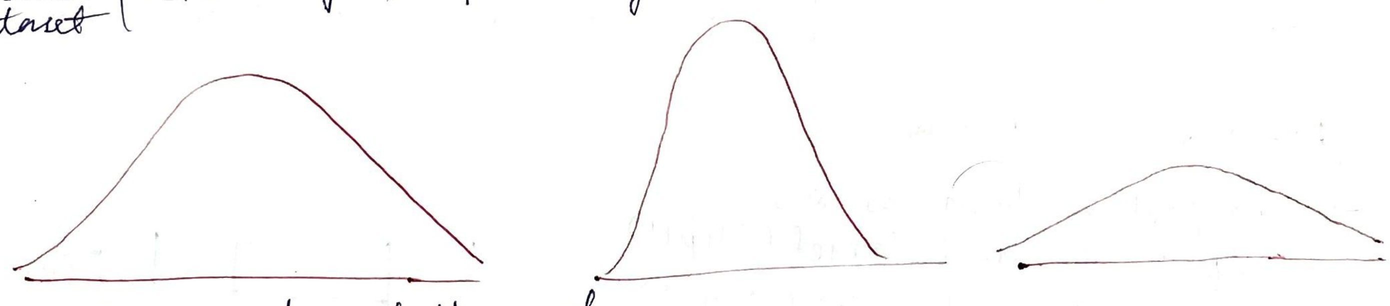
10-SEP-2022

Gaussian / Normal Distribution



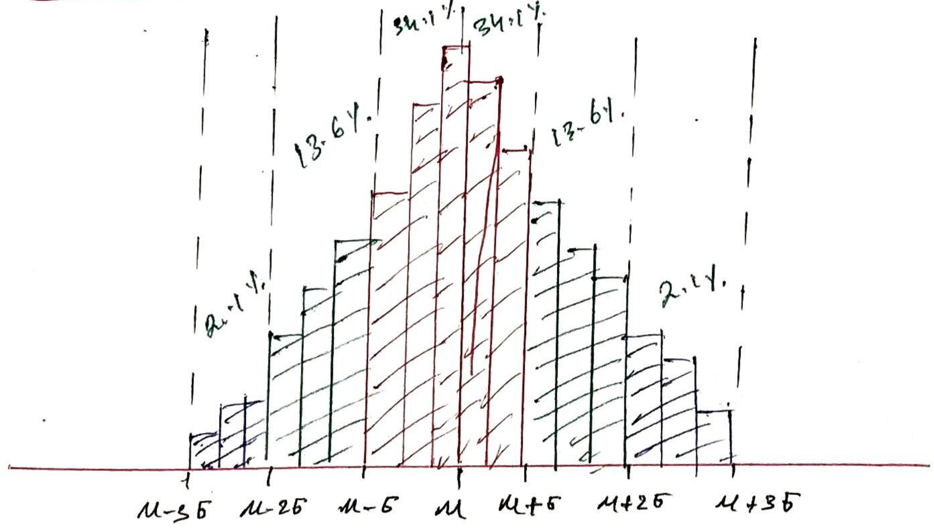
Area under the curve = 1

- By smoothing (using KDE) the histogram, we can obtain this graph.
- Age, weight, height → follows gaussian/normal distribution.
- IRIS dataset → follows gaussian/normal distribution.
  follower dataset (Petal length, Sepal length, Petal width, Sepal width)



- Spread can be different.

Emperical Rule of Normal Distribution



Assumptions of the data

[68 - 95 - 99.7 % Rule]

① Within 1st S.D of around the mean, there are 68 % of the distribution will be falling, left and right of 1st standard deviation.

② Within 2nd S.D around the mean, there are 95% of the distribution will be falling, left and right of 2nd standard deviation.

③ Within 3rd S.D around the mean there are 99.7% of the distribution will be falling, left and right of 3rd standard deviation.

[Note: With the help of Q-Q plot, we can find whether a distribution is gaussian or not.]

Standard Normal Distribution

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

$\Downarrow$ Transform

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

→ Using Z-score, we can transform gaussian distribution into standard normal distribution.

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

$\mu = 3, \quad \sigma = 1.414$

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

$\left\{ \frac{\sigma}{\sqrt{n}} = \text{Standard Error.} \right\}$

Helpful in inferential Stats

→ Here, $n = 1$, because we apply this formula on each variable, so our sample size will become 1.

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

for $x=1$,

$$z\text{-score} = \frac{1-3}{1.414} = \cancel{4}\ -1.414$$

for $x=2$,

$$z\text{-score} = \frac{2-3}{1.414} = -0.707$$

for $x=3$,

$$z\text{-score} = \frac{3-3}{1.414} = 0$$

for $x=4$,

$$z\text{-score} = \frac{4-3}{1.414} = 0.707$$

for $x=5$,

$$z\text{-score} = \frac{5-3}{1.414} = 1.414$$

New distribution.

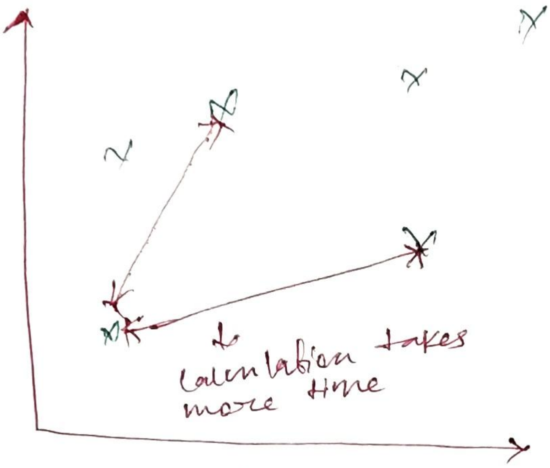
$$y = \{-1.414, -0.707, 0, 0.707, 1.414\}$$

mean $= 0$, standard deviation $= 1$

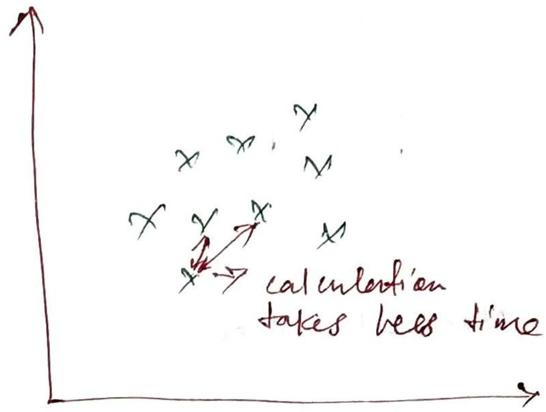
why we need to convert Gaussian distribution to standard normal distribution?

Age (Years)	Weight (Kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	175
34	87	180
28	83	180
29	80	175

→ Units scale is different, so our value will be differ by huge number.
→ Mathematical calculation will take more time.

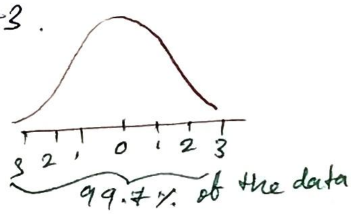


Different scale



within same scale

[Note:
- the entire process of scaling down the unit is called standarization.
- After performing calculation, we can revert back to original scale.
- All the values will be transformed b/w -3 to +3.



~~e.g. calculate for Age~~.

Normalization

→ Here, we scale down the data according to our range.

Min-Max Scaler

→ Transform the value b/w 0 and 1.

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

e.g.

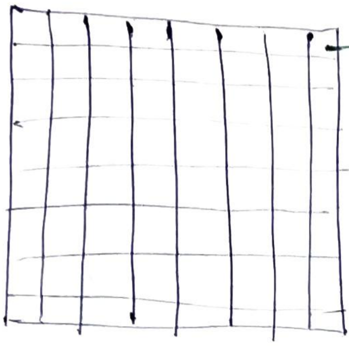
x	y (Normalization)		y' (Standardization)
1	0	⇒ $\frac{1-1}{5-1}$	-1.414
2	0.25	⇒ $\frac{2-1}{5-1}$	-0.707
3	0.5	⇒ $\frac{3-1}{5-1}$	0
4	0.75	⇒ $\frac{4-1}{5-1}$	0.707
5	1	⇒ $\frac{5-1}{5-1}$	1.414

0 to 1 → original distribution will be gone.

Here Gaussian distribution will be intact.

[Note: In images we don't worry about normal distribution.]

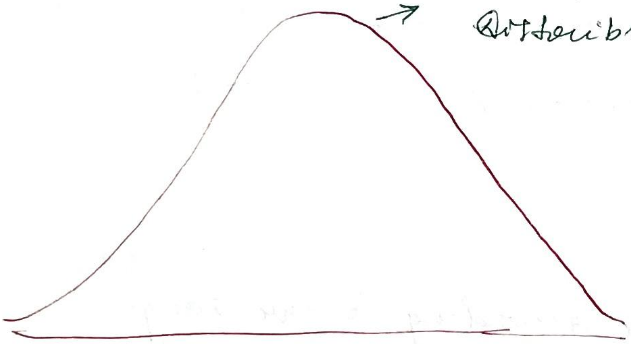
Where do we apply?
→ In deep learning and some part in ML.



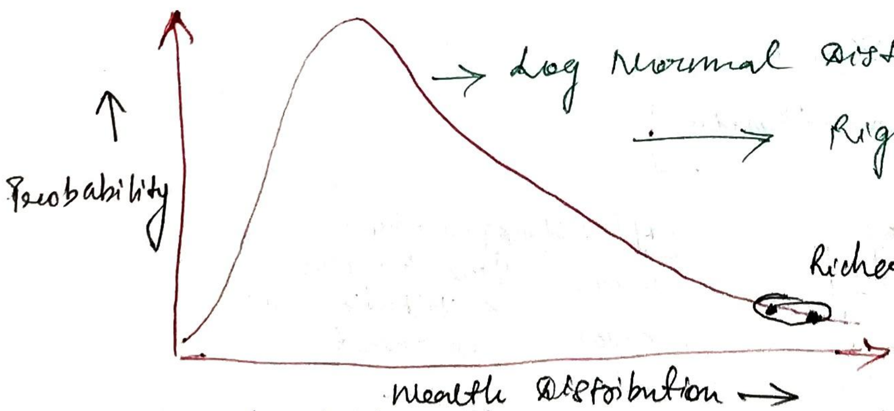
→ Pixel → Range from 0 to 255
↓
We can convert it to 0 to 1.
(Normalization.)

→ Normalization will be mostly used in deep learning i.e CNN.

Log Normal Distribution



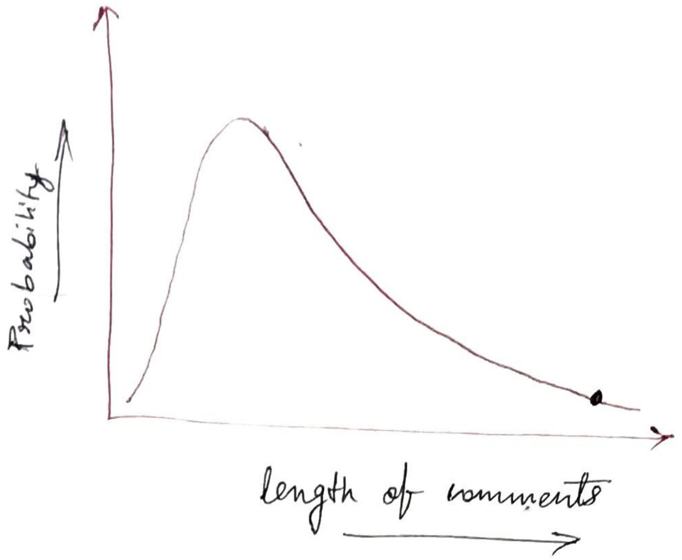
→ Normal / Gaussian Distribution



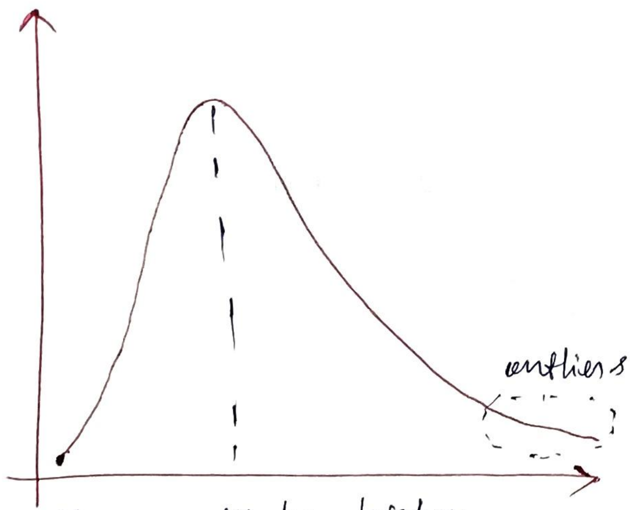
→ Log Normal Distribution
→ Right skewed

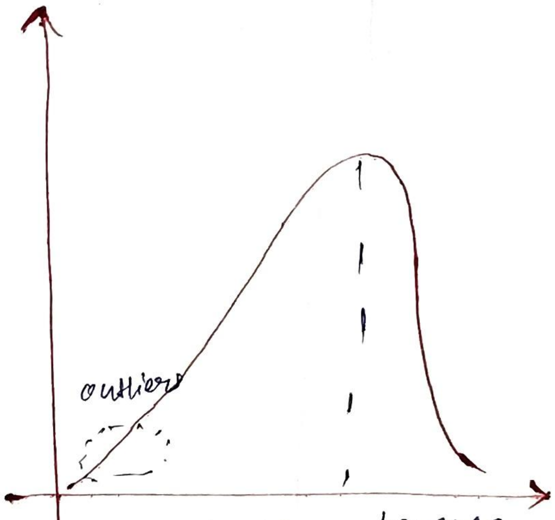
e.g. ① Wealth distribution

e.g length of comments



Que what is the relationship b/w mean, mode and median b/w these two graphs?





Ans Mean will be higher because outlier is present at the higher side and outliers impacts mean.

$\boxed{\text{Mode} < \text{Median} < \text{Mean}}$

Mean will be lower because outlier is present at lower size and outlier impacts mean.

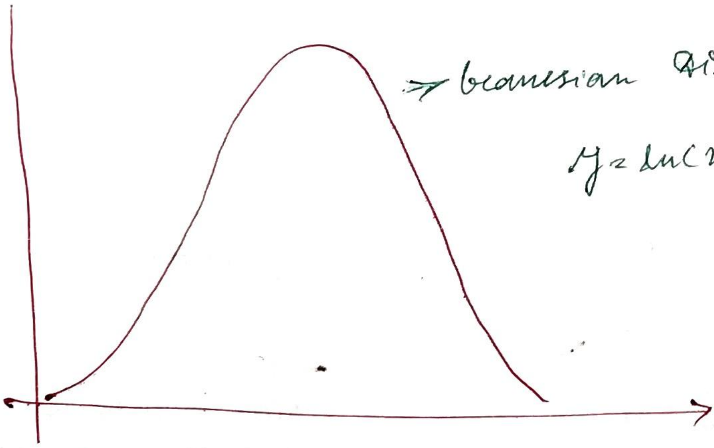
$\boxed{\text{Mean} < \text{Median} < \text{Mode}}$

→ log-normal distribution (Right Skewed Data)

$X \approx$ Log Normal Distribution

then

$Y = \ln(X)$ has a normal distribution.

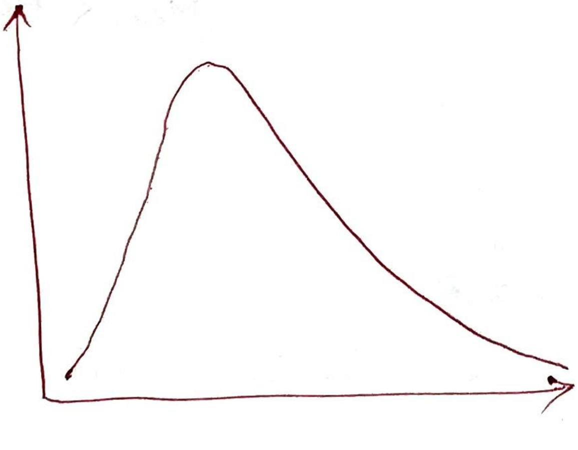


→ Gaussian distribution

$Y = \ln(X)$

e.g.: (a) No. of Runs by batsman

(b) Marks by students in a classroom

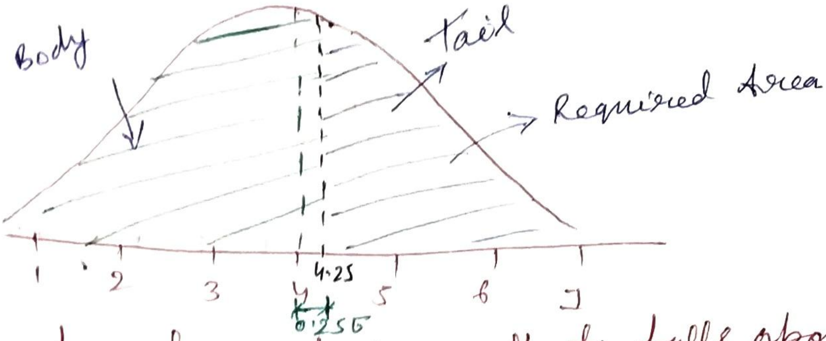


Que.

$x = \{1, 2, 3, 4, 5, 6, 7\}$

Let's $\mu = 4$

assume $\sigma = 1$



What is the percentage of score that falls above 4.25?

Ans

Area of entire curve = 1

$$Z\text{-score} = \frac{4.25 - 4}{1} = \boxed{0.25}$$ S.D towards right from the mean.

Z-table (Area under the curve)

- Negative Z-score table
- Positive Z-score table

for Z-score = 0.25

Area = 0.5987 (Green colour shaded area)

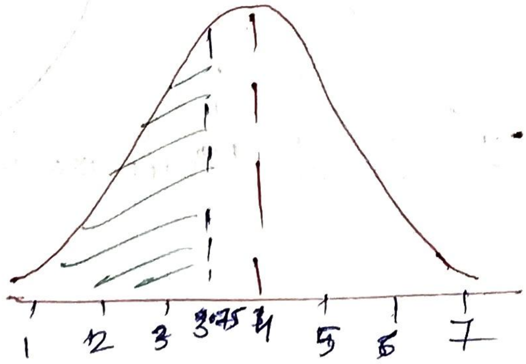
Required Area = 1 - 0.5987 = 0.4013 = 40.13%

Que what is the percentage of score that falls below 3.75?

Ans Area of entire curve = 1

$$Z\text{-score} = \frac{3.75 - 4}{1} = \boxed{-0.25}$$ S.D towards left from the mean.

Area = 0.4013 (Green colour shaded Area)



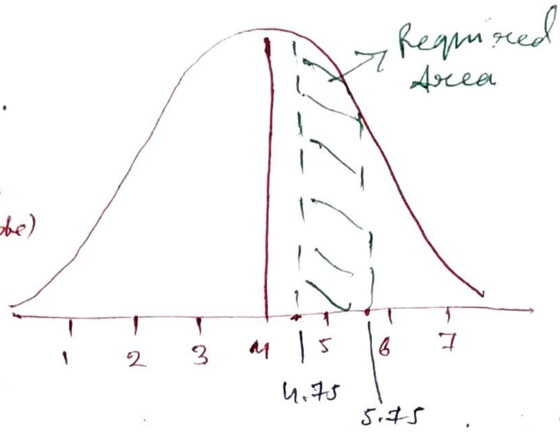
Que what is the percentage of score that falls b/w 4.75 and 5.75.?

Ans

for 4.75,

Z-score = $\frac{4.75-4}{1} = 0.75$

Area = 0.77337 (till 4.75) (from Z-table)



for 5.75,

Z-score = $\frac{5.75-4}{1} = 1.75$

Area = 0.95994 (till 5.75) (from Z-table)

Required Area = 0.95994 − 0.77337
= 0.18657 ≈ 18.6%

Que In ~~Indina~~ India the average IQ is 100 with a standard deviation of 15. What is the % of population would you expect to have an IQ

(a) Lower than 85.

(b) Higher than 85

(c) Between 85 and 100.

Ans (a)

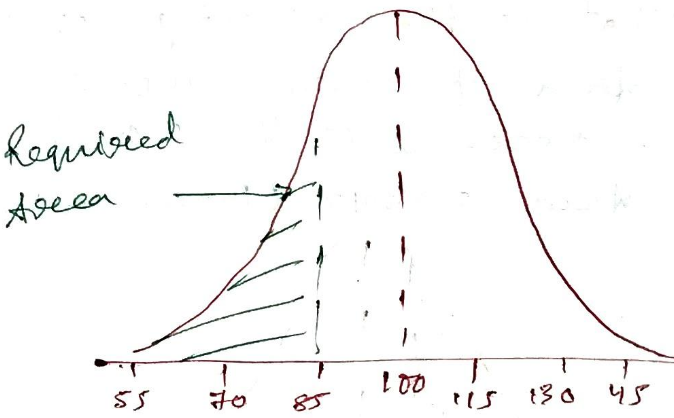
$\mu = 100$
$\sigma = 15$

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

= −1

Area = 0.15866

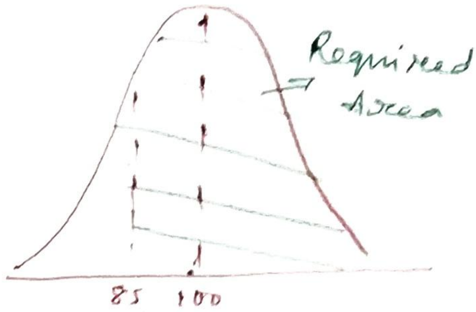
≈ 15.8%



(b) $\mu = 100$

$\sigma = 15$

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



Area = 0.15866

= 15.86%

Required Area = 100 − 15.86

= 84.14%

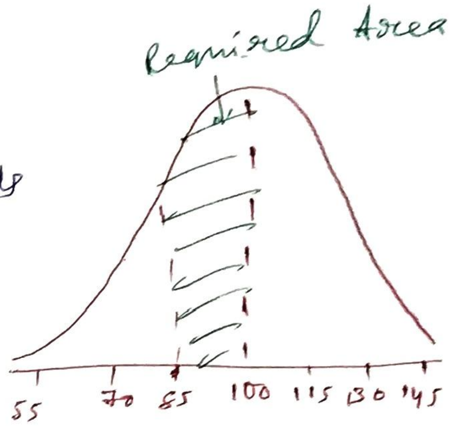
(c) $\mu = 100$

$\sigma = 15$

for $x = 85$,

$$\text{Z-Score} = \frac{85-100}{15} = -1$$ S.D towards left from mean

Area = 0.15866 till 85



for $x = 100$,

$$\text{Z-Score} = \frac{100-100}{15} = 0$$

Area = 0.5 till 100

Required Area = 0.5 − 0.15866

= 0.34134

= 34.13%