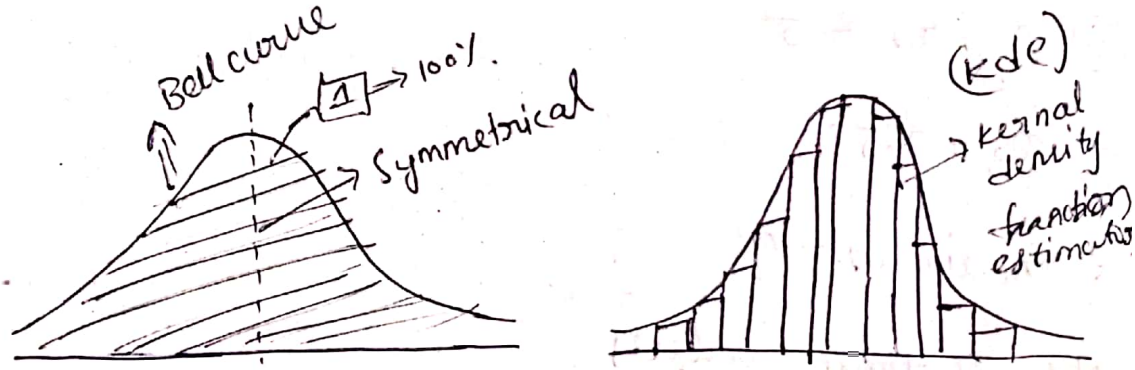
Statistics part-3

Agenda:

* Normal Distribution
* Standard Normal Distribution
* Z-Score
* Standardization and Normalization
* log Normal Distribution

Gaussian / Normal Distribution:



→ the Gaussian distribution is in the bell shaped curve.

→ It is Symmetrical

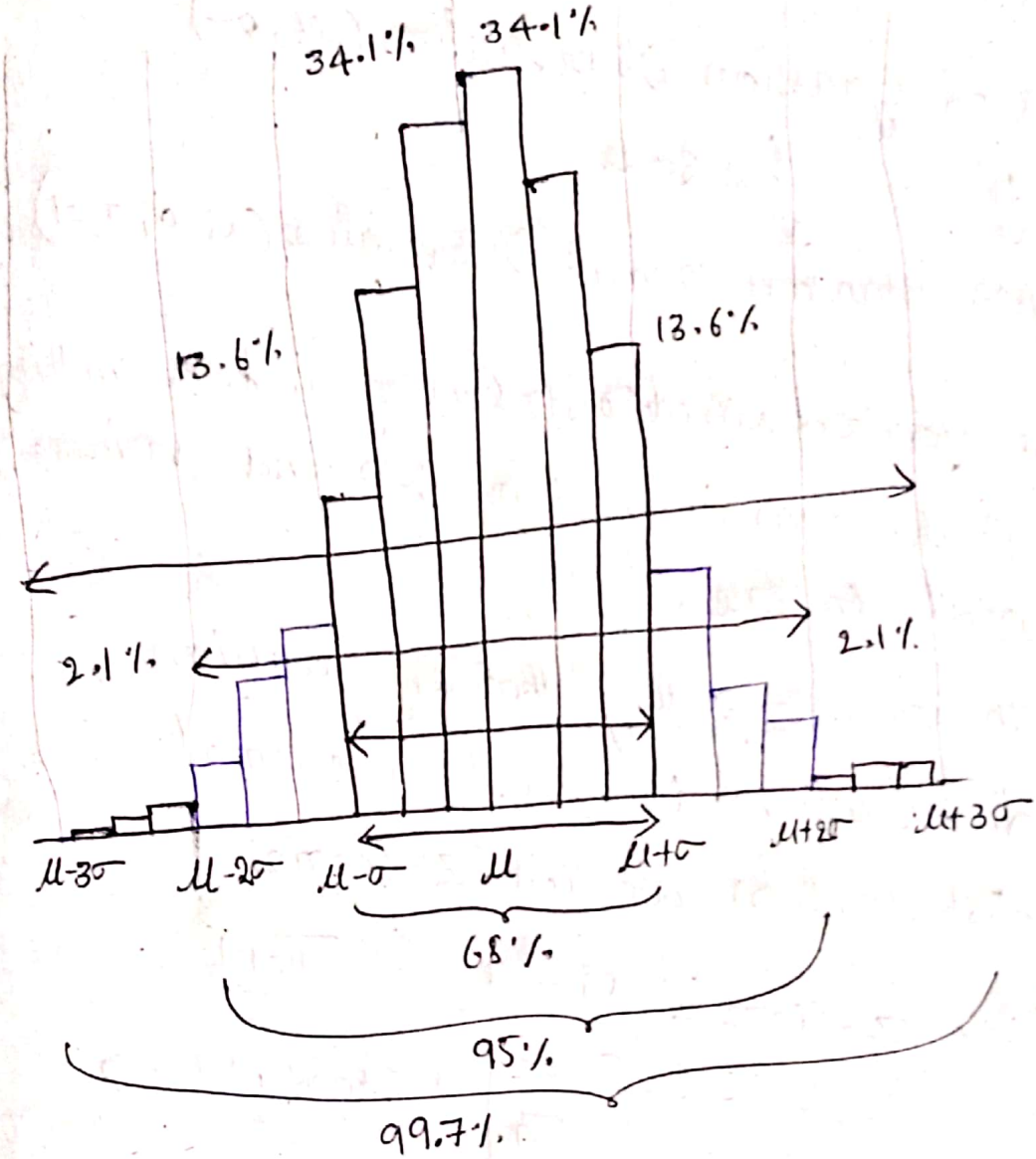
→ the area covered under the curve is 1. that means 100%.

→ why Gaussian / Normal distribution important?

↓

Assumptions of data

Emperical Rule of Normal distribution.



Emperical rule:

68-95-99.7 rule.

Q-Q plot ⇒ Distribution is Gaussion or not!

* ~~The~~ ~~In~~ the Emperical rule
* The sum of one standard deviation to right and one to left is covered 68% of data
* The sum of two standard deviation to right and two standard deviation is covered 95% of data.
* The sum of three standard deviation to right and three standard deviation is covered 99.7% of data.

Standard Normal Distribution:

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

$\Downarrow$ $\Downarrow$ Z-score

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

The standard Normal Distribution is here nothing but mean is equal to zero and variance equal to one.

→ to convert x to y that is Gaussian distribution to standard normal distribution we use Z-score

i.e., $z\text{-score} = \dfrac{x_i - \mu}{\frac{\sigma}{\sqrt{n}}}$ $\quad n=1$

$\frac{\sigma}{\sqrt{n}}$ → Standard Error

$$z\text{-score} = \boxed{\frac{x_i - \mu}{\sigma}} \quad (\because n=1)$$

Ex: $x = \{1, 2, 3, 4, 5\}$ → Gaussian distribution

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

convert to SND

$= \frac{1-3}{1.414} = -1.414$ $\quad y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

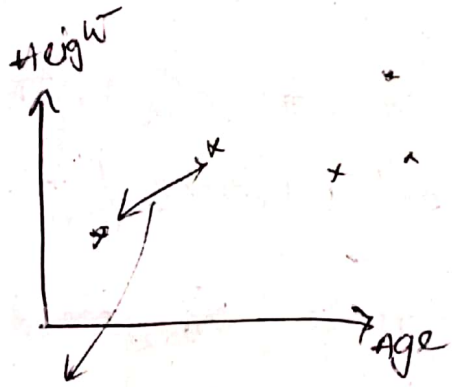
$= \frac{2-3}{1.414} = -0.707$

$= \frac{4-3}{1.414} = \frac{1}{1.414} \approx 0$

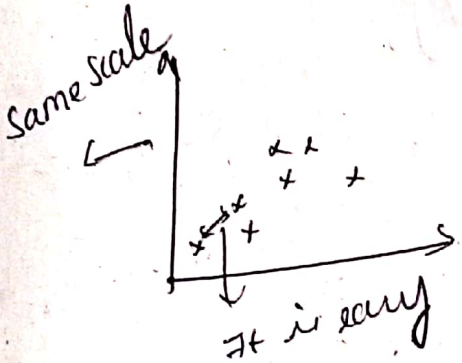
- so on :-

why ?

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



It is difficult to calculate



It is easy

Feature Scaling

1. Standardization

$x \rightarrow$ Normal Distribution (μ, σ)

$\downarrow$ z-score

$y \rightarrow$ SND $(\mu = 0, \sigma = 1)$

why do we do this ?

→ To bring the features in the same scale.

Normalization [0-1]

In standardization we have $\mu = 0$ and $\sigma = 1$ those are pre-fixed.

But in normalization we take the mean and variance [0-1]

Because to scale down

→ Min-max scalar (0-1)

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

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

x	$\Rightarrow$	y
1		0
2		0.25
3		0.5
4		0.75
5		1

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

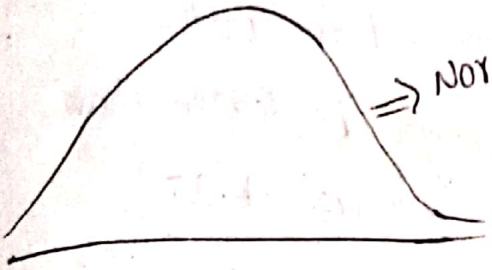
$\frac{1-1}{5-1} = 0$

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

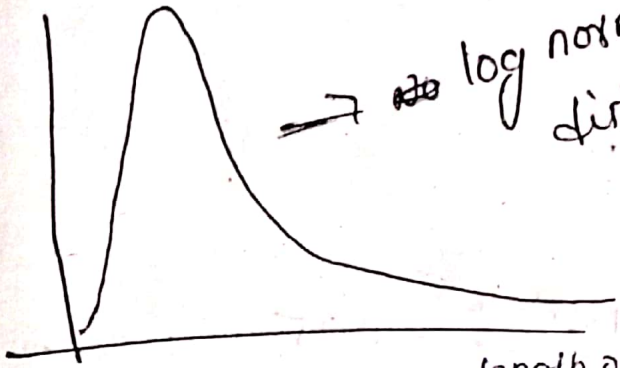
$\frac{2-1}{5-1} = \frac{1}{4}$

* standardization is used in machine learning algorithms.
* Normalization is used in deep learning.

Log Normal Distribution

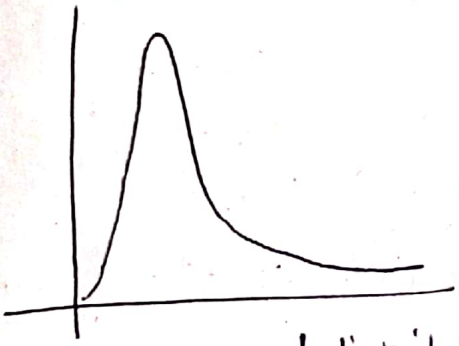


⇒ Normal / Gaussian Distribution

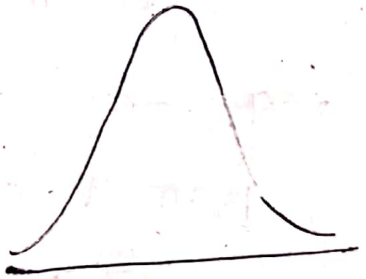


⇒ log normal distribution

* To convert log normal Distribution to Gaussian

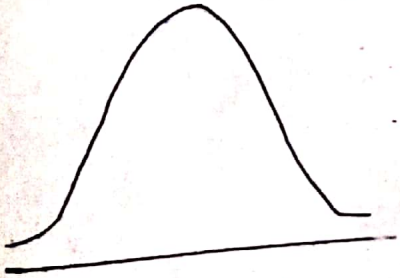


⇒ $y = \ln(x)$



$x \approx$ log normal distribution

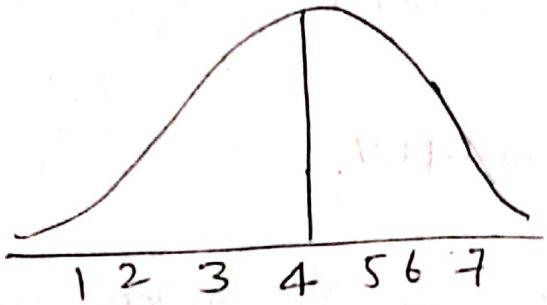
* To convert Gaussian to log normal Distribution



⇒ $x = \exp(y)$

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

⇒ what is the percentage of score that falls above 4.25?



$$z\text{-score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

z-table (Area under the curve)

↓

See in google for z-table positive and negative values.

According to that values we can find the percentage of area covered under the curve.

for $0.25 \Rightarrow 0.598 \Rightarrow 0.598 \times 100\% = 59.8\%$

⇒ what is the percentage of score that fall below 3.75

$$z\text{-score} = \frac{x_i - \mu}{\sigma} = \frac{3.75 - 4}{1} = -0.25$$

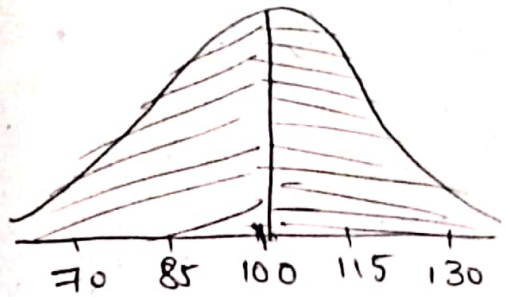
$\Rightarrow -0.25 \Rightarrow 0.40 \Rightarrow 0.40 \times 100\% = 40\%$

* 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

(i) lower than 85

(ii) Higher than 85

(iii) Between 85 and 100



(i) lower than 85

$\mu = 100$

$\sigma = 15$

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

$= \frac{85 - 100}{15} = -1$

$-1 \Rightarrow$ z-~~pos~~ Negative score is 0.15 $\Rightarrow$ 0.15 $\times$ 100%.

$= 15\%$

(ii) ~~lower~~ Higher than 85

z-score $= +1 \rightarrow$ z-positive score for +1 is 0.84

$\Rightarrow 0.84 \times 100\% = 84\%$.

(iii) Between 85 and 100 :

~~first find lower~~ Subtract ~~higher~~ Smaller from higher after finding z-scores. $\rightarrow$ 84% covered higher than 85

$\Rightarrow$ 50% covered lower than 100

84% - 50% $\Rightarrow$ 34% between 85 & 100