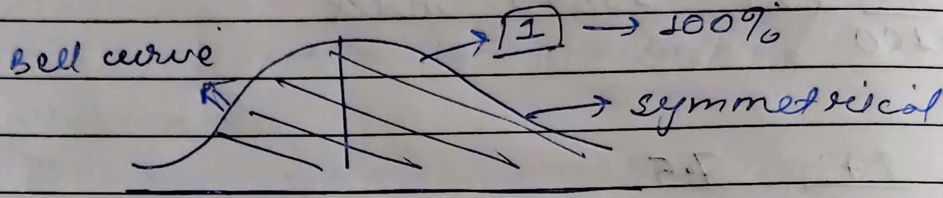
Topics

1. Normal Distribution / Gaussian Dist.
2. Standard Normal Distribution
3. Z score
4.

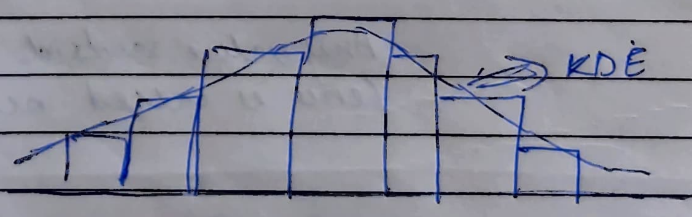
Gaussian / Normal Distribution



As per domain expertise Age, weight, height data were having ~~Gaussian~~ Gaussian distribution

IRIS DATASET

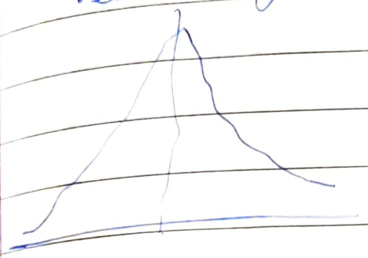
→ when histogram is smoothen using Kernel Density Estimator (KDE) then we get Guassian Distribution

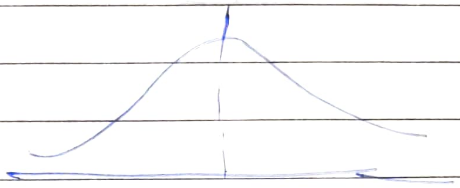


→ IRIS Dataset was first where ML was applied

IRIS Dataset

Petal length Sepal length





Emperical Rule of Normal / Guassian Distribution

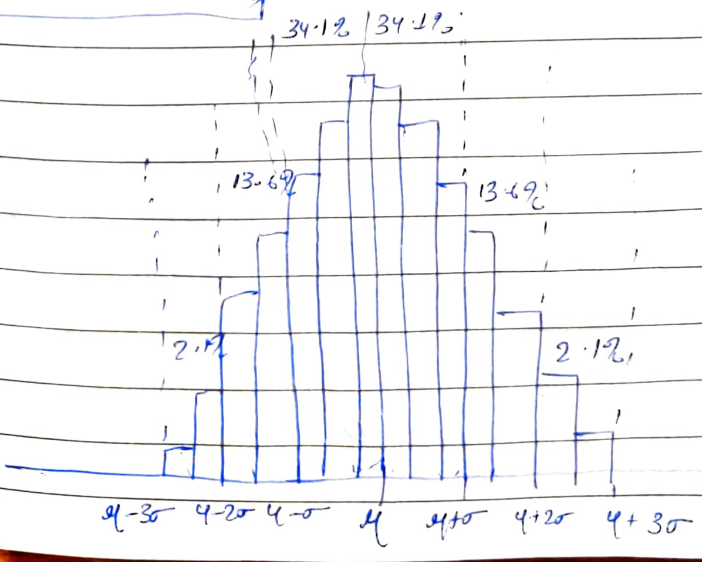
* Assumption of the data

→ Around ~~to~~ 68% of ~~data~~ the entire distribution falls under first standard deviation to the ~~to~~ left and right

→ Around 95% of the entire distribution falls under second standard deviation to the left and right

→ ~~Arou~~ Around 99.7% of the entire distribution falls under three standard deviation to the ~~to~~ left and right

$\boxed{68-95-99.7\%}$ → Emperical ~~formula~~ rule



Page No.:
Date:

→ using Q-Q plot we can find whether distribution is Gaussian or not

Standard Normal distribution (SND)

~~Let~~ x is a random variable with Gaussian Distribution

$$x \approx \text{Gaussian Distribution}(\mu, \sigma)$$
$$\Downarrow$$
$$y \approx SND(\mu = 0, \sigma = 1)$$

→ using Z score formula we ~~get SND~~ can convert Gaussian to SND

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

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

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

$\Rightarrow$ Standard error (will be used in inferential stats)

⇒ Z score tells us how many standard deviation(σ) away from the mean (μ)

here $\boxed{n = 1}$, since $n = 1$ will be applied to all value of dataset x

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

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

$\mu = 3$

$\sigma = 1.414$

$= \frac{1-3}{1.414} = \cancel{-1.14}\ -1.414$

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

$= \frac{3-3}{1.414} = 0$

$= \frac{4-3}{1.414} = 0.707$

$= \frac{5-3}{1.414} = +1.414$

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

mean (y) has become 0

Q) Why we are converting Gaussian to SND

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

This above dataset has different units and data is having high difference values

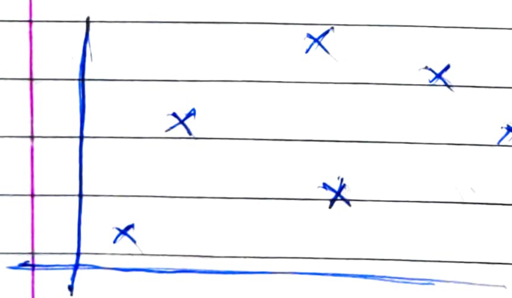
24 → 72 → 150

72 150

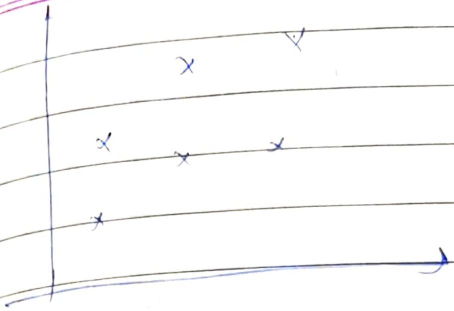
⋮ ⋮

80 175

Initially data is



By using standardization, we are trying to bring the data closer



imp

standardization is used to achieve this scaling down of values, where

$\mu = 0$

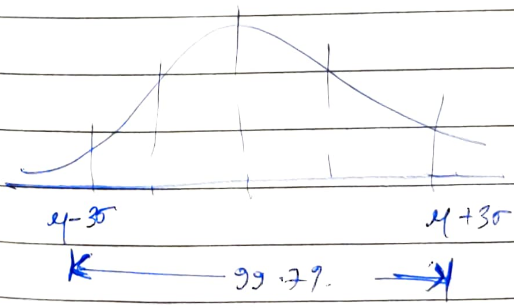
$\sigma = 1$

using Z score formula we can achieve this

standardization {Z-score}

$\boxed{\mu = 0, \sigma = 1}$

[$-3 \longleftrightarrow +3$] values are converted in this range automatically



→ But in Normalization this scale of min and max is defined by user

Normalization [Lower scale ⟷ Higher scale]

① Min Max Scalar [0 - 1]

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

using above formula

x ⇒	y
1	0
2	0.25
3	0.5
4	0.75
5	1

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

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

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

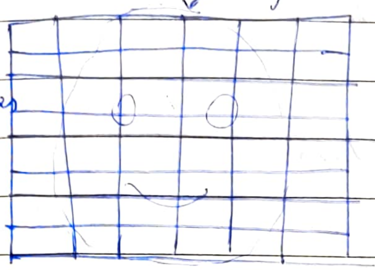
$\Rightarrow \frac{4-1}{5-1} = \frac{3}{4} = 0.75$

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

Application of Normalization

- Deep learning
- where ~~pix~~ pixels are involved
- where Gaussian distribution is not required

Image



In an image the ~~piex~~ ~~pix~~ pixels will range between (0 - 255)

It can be scaled down to 0 - 1 using Normalization

Imp

- Feature scaling
  - → Standardization (In ML)
  - → Normalization (In DL)

{In most cases.}

RECAP

Standardization

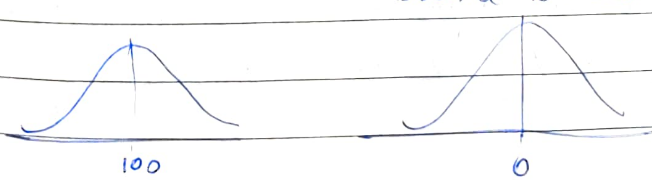
$x \rightarrow$ Normal distribution (μ, σ)

$\Downarrow$ Z score

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

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

Why we are doing → Bring the features in the same scale



B.

Normalization

→ Here we are not doing $\mu = 0$, & $\sigma = 1$

→ we are using Min Max scalar to bring it in a range

x	y (Normalization)	y'	y' (Standardization)
~~0~~	~~0~~	~~-1.414~~	
1	0	-0.707	-1.414
2	0.25	0	-0.707
3	0.5	+0.707	0
4	0.75	-	+0.707
5	1		+1.414

Here GD is not maintained

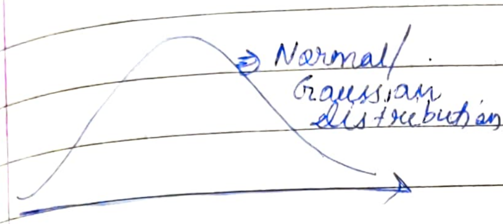
Here Gaussian Distribution is maintained

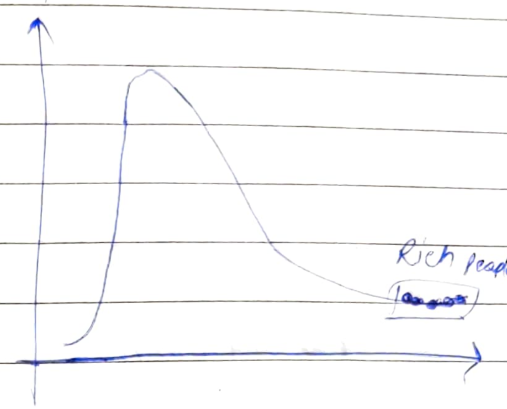
→ Standardization is ideal for data that gets a normal / Gaussian distribution

→ It's superior when handling data with outliers

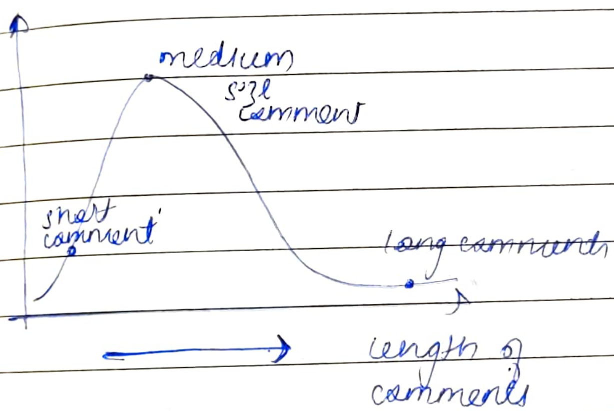
→ Normalization is safer alternative when we are unsure about the distribution of data

Log Normal Distribution





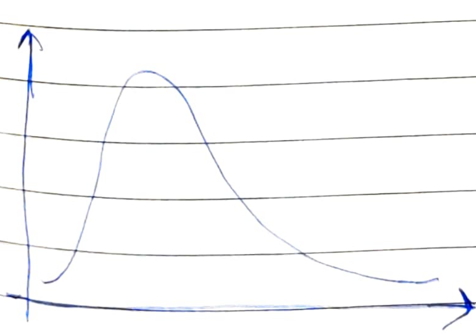
Wealth
Distribution

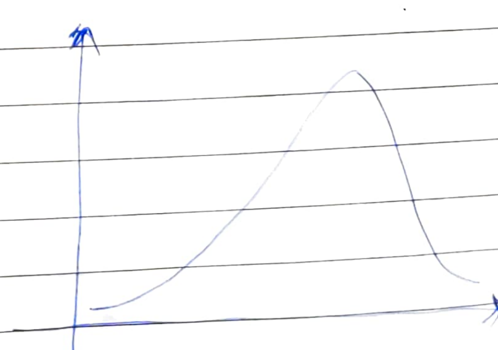


Assignment

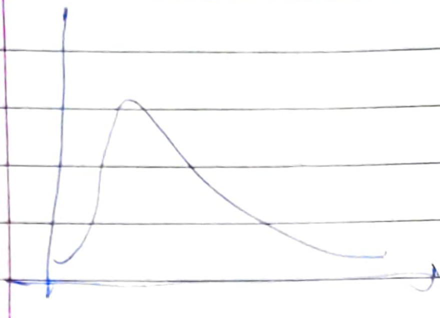
Assignment

From ascending order give the relation of mean, median and mode?

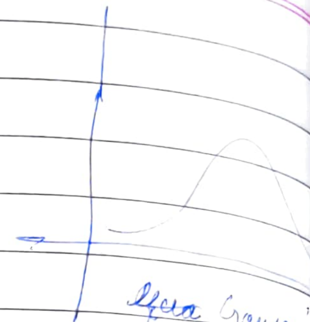




$\ln(x) = \log_e(x)$ (natural log)
$\log(x) \Rightarrow \log_{10}(x)$

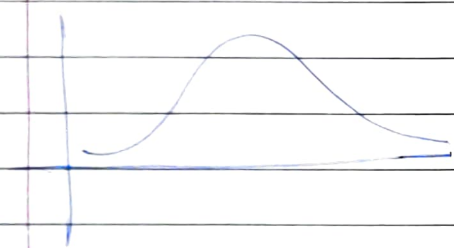


$\Rightarrow$ ~~$\log(x)$~~ $y = \ln(x)$ $\rightarrow$

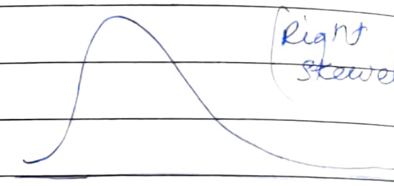


Gaussian distribution

$x \approx$ log normal distribution



$\Rightarrow \exp(x)$
$\Downarrow$
y



(Right skewed)

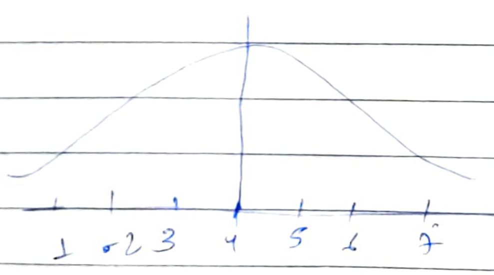
$x \approx$ Normal Distribution

$y \approx$ log normal distribution

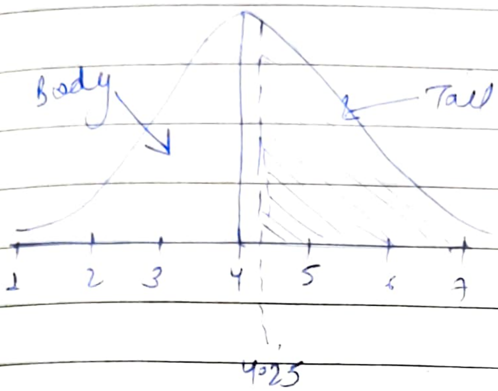
→ only right skewed is log normally distributed

Q

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

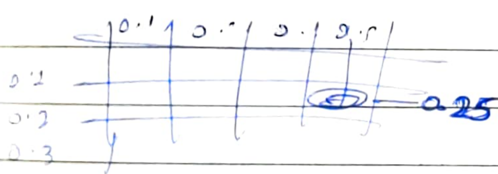


What is the ~~perc~~ percentage of score that fall above 4.25?



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

Using Z table we can get the area



You can get values of Z score from google search

0.59 ⇒ 59% -0.25

1 - 0.59% = 0.41%

For 3.75

Z score = $\frac{3.75 - 4}{1} = -0.25$

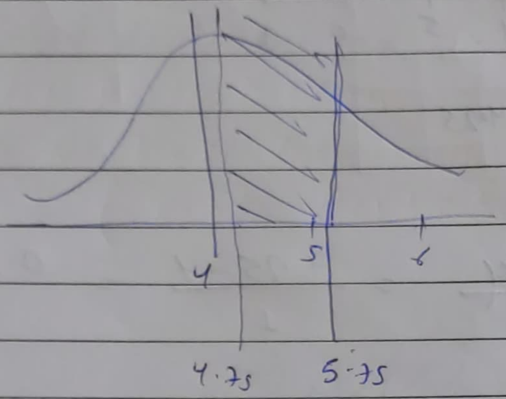
using Z table .4052

∴ % age = 40%

4.75 & ~~4.5757~~ 5.75

$\frac{4-4.75}{1} = -0.75$ $\quad .22663$

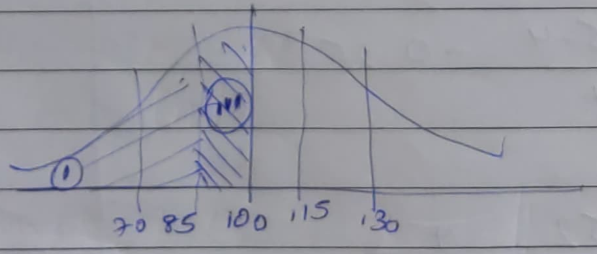
$\frac{4\cancel{0}-5.75}{1} = -1.75$ $\quad 0.4006$



$40 - 22 = 18\%$

(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 an IQ

(i) Lower than 85
(ii) Higher than 85
(iii) Between 85 and 100



(i) $Z\ score = \frac{85-100}{15} = -1$ $\quad = .1587$
$= 15.87$ (Ans)

Page No.:
Date: 11/10/22

$1 - 15.87$

$= 84.13$

$0.5 - 0.1587$

$= 0.3413$