STATISTIC FOR DATA SCIENCE [DAY 3]

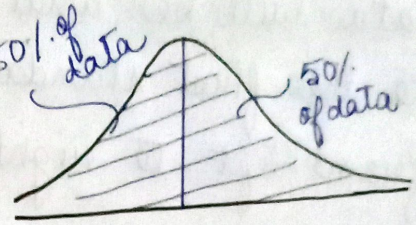
(1) NORMAL DISTRIBUTION

(2) STANDARD NORMAL DISTRIBUTION

(3) Z - SCORE

(1) Gaussian / NORMAL DISTRIBUTION

* Age, Weight, Height follows Normal Distribution



Both side is symmetrical

* Normal Distribution is a symmetrical bell-shaped curve. The area of curve is 1

(data on left side = data on right side)

IRIS DATASET.

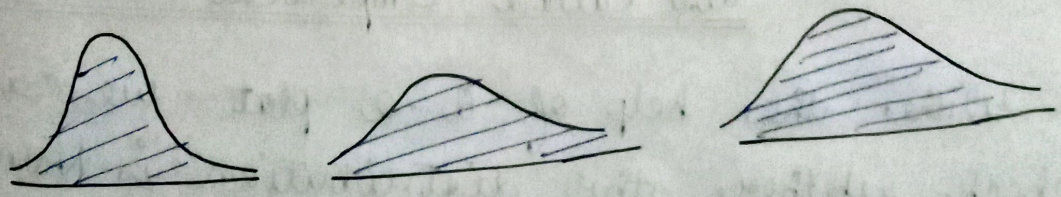
⇓

Petal length, sepal length, petal width, sepal width

⇓

It will follow Gaussian Distribution

⇓



Emperical Rule of Normal Distribution

Because of Normal Distribution we can
m

Within the first standard distribution between
left and right 68% of the entire
data will be available
to the first standard distribution
from left to right.

Within the second standard
distribution to the righ &
the left around 95% of the
entire data falling in
these region.

Within the third stand distribution to the
right & the left around 99.7% of the entire
data falling in this region.

This is called 68-95-99.7% Rule

also called Emperical Rule

* With the help of Q-Q plot we can
check whether the distribution is Normal
or not.

Recapitulate

STANDARD NORMAL DISTRIBUTION

Let a variable $x \in$ Gaussian Distribution (μ, σ)

then we can transform x to

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

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

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

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

for $i = 1$, $\frac{1-3}{1.414} = -1.414$

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

$i = 3$, 0.

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

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

$i = 5$, $\frac{5-3}{1.414} = 1.414$

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

z-score Formula,

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

where, $\frac{\sigma}{\sqrt{n}}$ is called Standard Error

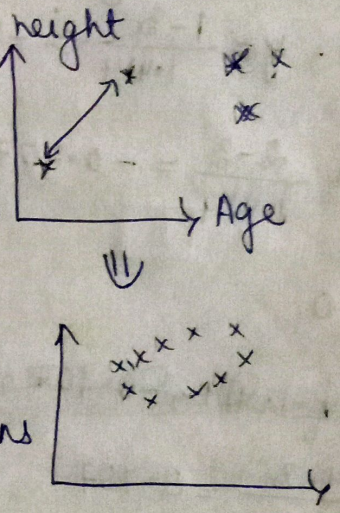
we will take $n = 1$ because we are going to apply z-score to each value of x.

Why ?

Age (yrs)	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

* since units are different values also differ very high so, my mathem -atical calculation Take time ↑↑

* we will try to change the scale and this process is called Standardization

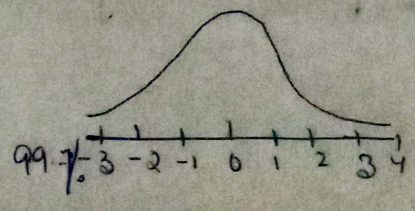


{we will do so to bring the values in the same scale so that our calculations will be easy}

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

For this we will apply $z\text{-score} = \frac{x_i - \mu}{\sigma}$ formula in each table.

Then all the values will be transformed to $[-3 \longleftrightarrow 3]$



In Normalization we give the range we try to normalize the value bet^n lower scale to higher scale

(1) Mean Max Scaler [0-1]

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

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

{0 → 1}

for $x = 1$, $y = \frac{1-1}{5-1} = 0$

$x = 2$, $y = \frac{2-1}{5-1} = \frac{1}{4}$

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

$x = 4$, $y = \frac{4-1}{5-1} = \frac{3}{4}$

$x = 5$, $y = \frac{5-1}{5-1} = \frac{4}{5}$

with the help of Mean Max Scaler we will transform the value in range [0-1]

* we use this in Deep Learning.

* This technique is called Normalization.

Feature Scaling (i) Normalisation
(ii) Standardisation

* In Deep Learning [Use Mean Max Scaler]

* In Machine Learning [Use Standardisation]

When inputs are taking Images [We use Normalisation]

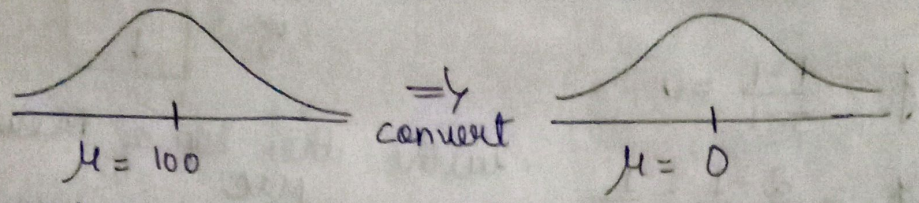
① **Standardization** $\quad z\text{-score} = \frac{x_i - \mu}{\sigma}$

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

⇓ apply z-scalere

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

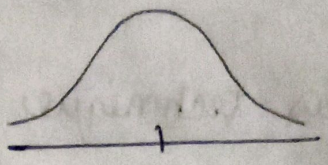
Why do we do this ⟶ Bring the feature in a smale scale.



Advantag :- (i) Calculation will be easy

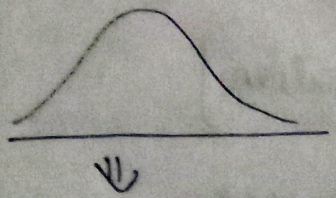
② **Normalization** [0-1]

(i) Min Max Scaler [we are normalising data bet^n a fixed range]



⇓ apply Min Max Scaler

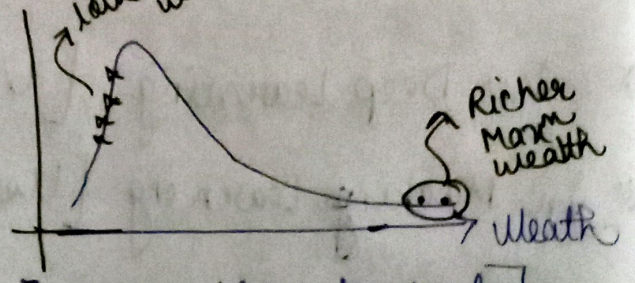
Log Normal Distribution



⇓

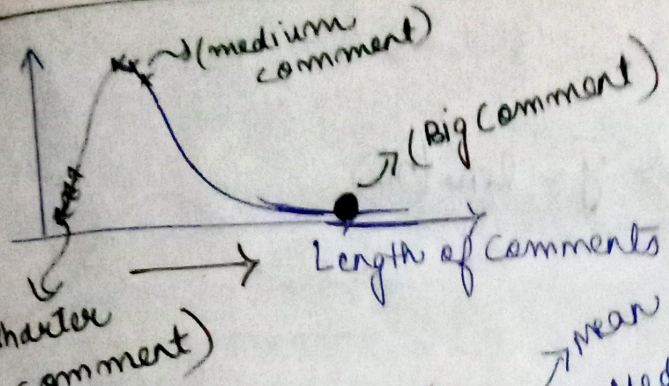
Normal Distribution

⟹

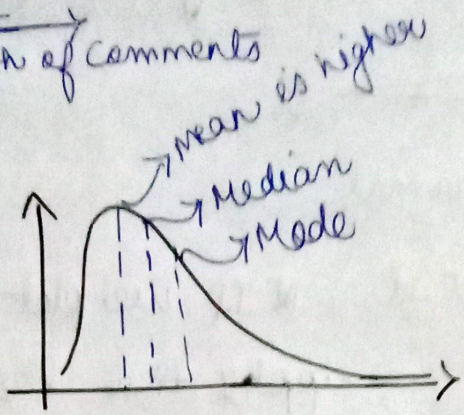


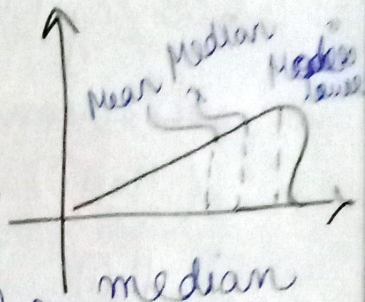
[Right - side skewed]

E.g. (i) Wealth Distribution



Q If we have



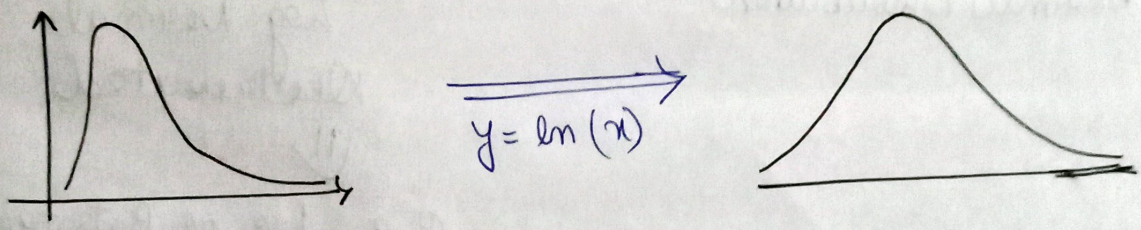


what is the relationship of mean, median mode.

From ascending order give relation of mean, median, mode.

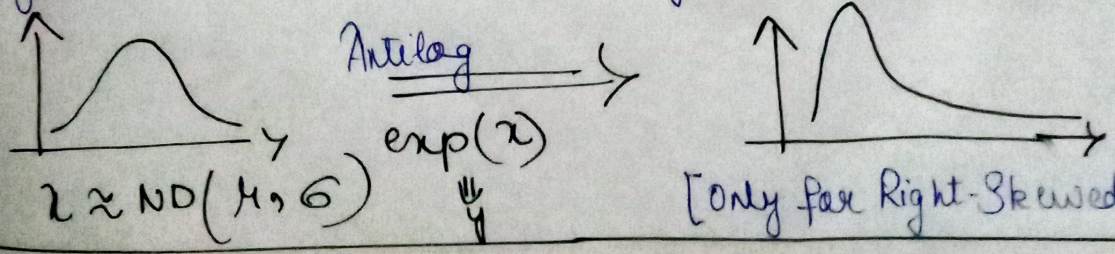
* For right-skewed distribution Mean > Median > Mode

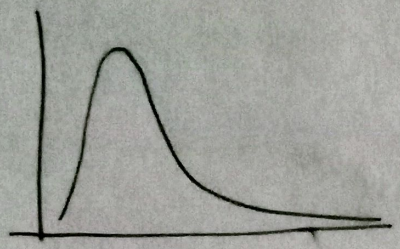
For left-skewed distribution Mean < Median < Mode



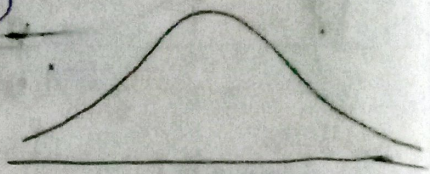
$x \sim$ Log Normal Distribution

* If the variable is log normally distributed then $y = \ln(x)$ has a normal distribution. If Y has a normal distribution then the exponential of Y, $x = \exp(Y)$ has a log-normal distributed





$\Rightarrow y = \ln(x)$



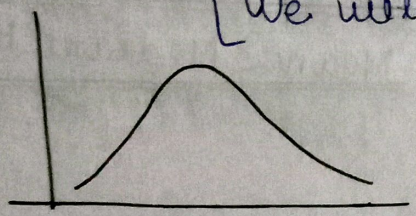
$x \approx$ Log Normal Distributed (μ, σ)

fig (i)

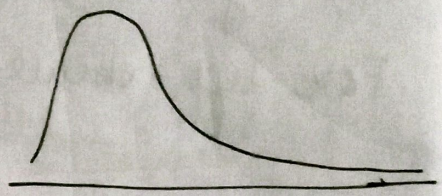
fig (ii)

* If we plot the fig (i) and apply the log and as a output if we get Normal Distribution then we can say fig (i) is Log Normal Distributed.

[We will do by using QQ plot]



$\Rightarrow \exp(x) \Rightarrow$



Normal Distribution

Log Normal Distributed

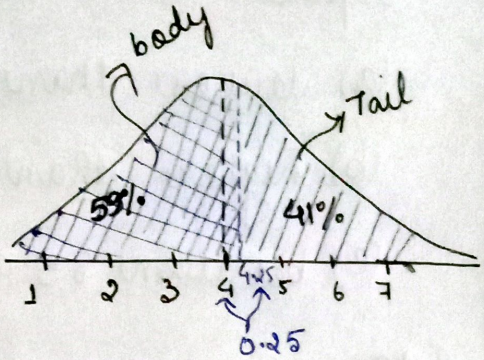
(ii)

e.g: No. of Batsman scored Right,
Marks of student in classroom,

Problem Statement

Q Let a variable $x = \{1, 2, 3, 4, 5, 6, 7\}$ with mean $\mu = 4$ and standards deviation $\sigma = 1$

(i) what is the percentage of score that falls above 4.25



Sol^n $Z\text{-score} = \frac{x_i - \mu}{\sigma}$

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

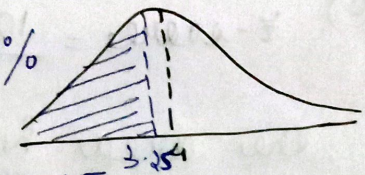
(i) Z-table [Area under the curve]

∴ The percentage under the curve from 1 to 4.25 is 59%

∴ the percentage of scale falls above 4.25 is 41%

(ii) what is the percentage of area falls below 3.75

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



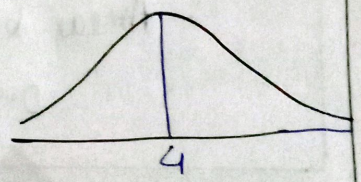
(iii) what is the area bet^n 4.75 & 5.75

$Z_1 = \frac{4.75 - 4}{1}$

$= 0.75$

$Z_2 = \frac{5.75 - 4}{1}$

$= 1.75$



(iii) 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

(a) lower than 85 (Ans-0.1587)

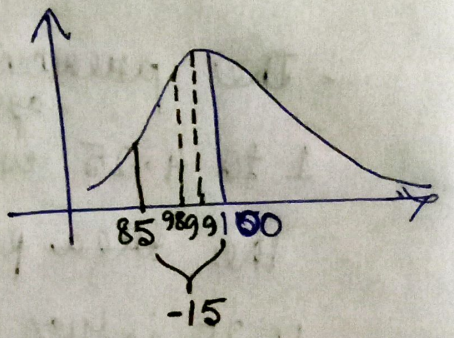
(b) higher than 85 (Ans:-0.8413)

(c) between 85 and 100 (Ans:- 0.3413)

Sol:- Given, $\mu = 100$, $\sigma = 15$

(a) Z-score $= \frac{x_i - \mu}{\sigma} = \frac{85 - 100}{15} = -1.0$

The area below than 85 is .15866



(b) Z-score $= \frac{85-100}{15} = -1$

The area higher than 85 is 0.844134

(c) Z-score $= \frac{100-100}{15} = 0$

The area higher than 100 is .50000

Area between 85 and 100 is

$0.844134 - 0.50000 = 0.344134$