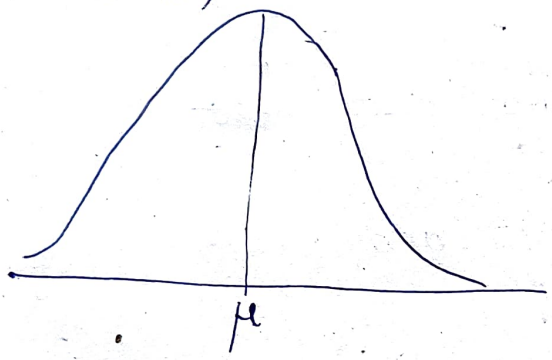
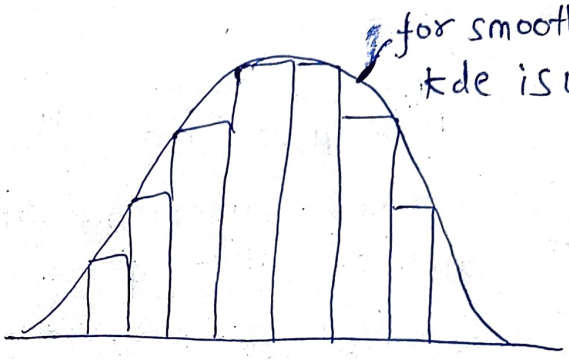
Stats Basics

① Normal distribution
② standard normal distribution
③ Z-score.

A) Normal distribution (Gaussian)





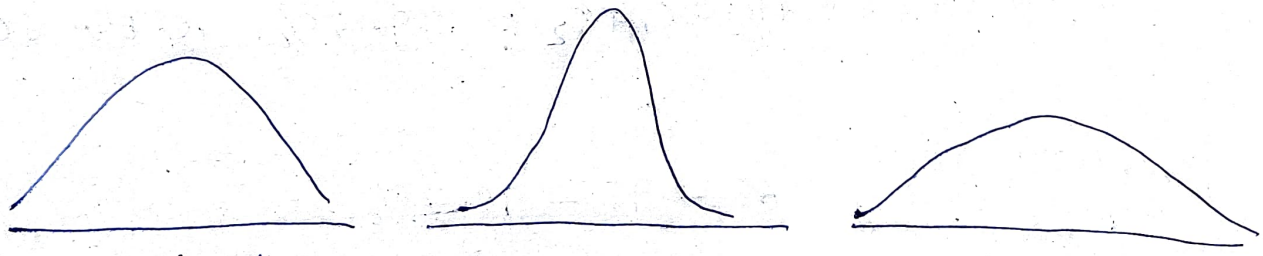
Kde ≡ kernal density Estimator

① symmetrical about mean μ.

② Area under normal distribution curve is 1

③ most features follow normal distribution.

Eg: Iris data set:
petal length, width
sepal length, width.
} follow gaussian distribution

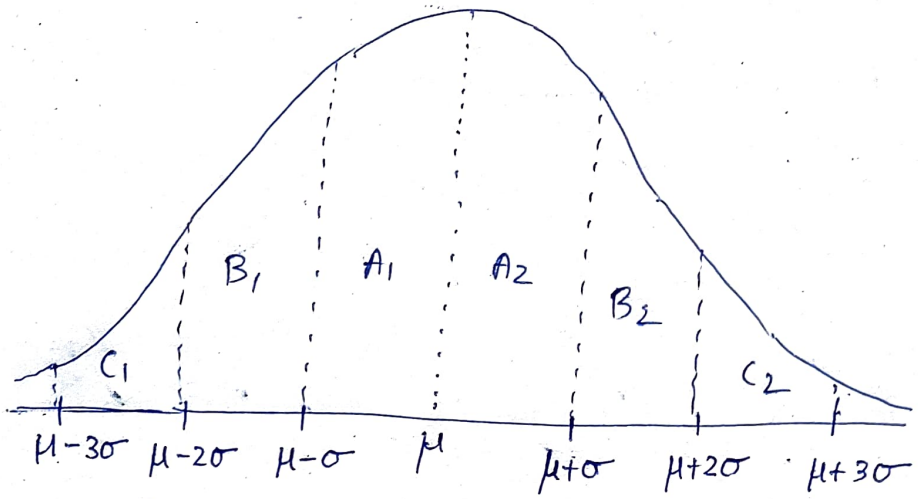


* spread may vary depending on the data set.

Eg: Age, weight, height also follow normal distribution.

↳ This is coming from domain experts who have collected and analysied data sets.

* Emperical rule of normal distribution



for normal distribution.

① within first σ to left and right 68% of total data lies.

ie $A_1+A_2 \Rightarrow 68\%$. ie blw 2σ

② within second σ to left and right 95% of total data lies.

ie. $A_1+A_2+B_1+B_2 = 95\%$. ie blw 4σ

③ within third σ to left and right 99.7% of total data lies.

ie $A_1+A_2+B_1+B_2+C_1+C_2 = 99.7\%$. ie blw 6σ

∴ Emperical formula $= 68-95-99.7\%$

* Q-Q plot is used to determine whether a distribution is Gaussian or not.

* <u>standard Normal Distribution</u> (SND)

- say Random variable

  x belongs to Gaussian Distribution with mean (μ) std deviation (σ).

→ It can be converted to y with $\mu=0$, $\sigma=1$ using z score

x = Gaussian distribution (μ, σ)

⇓ using ← z score

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

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

→ Here $n=1$ since this will be applied to all elements of data set.

n ≈ sample size.

Eg: {1, 2, 3, 4, 5}

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

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

SND = {-1.42, -0.71, 0, 0.71, 1.42}.

$\frac{1-3}{1.41} = -1.42$, $\frac{2-3}{1.41} = -0.71$, $\frac{3-3}{1.41} = 0$,

$\frac{4-3}{1.41} = 0.71$, $\frac{5-3}{1.41} = 1.42$

* why Gaussian distbⁿ is converted to SND ??

AGe (years)	weight (kgs)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175
$\mu=0, \sigma=1$	$\mu=0\ \sigma=1$	

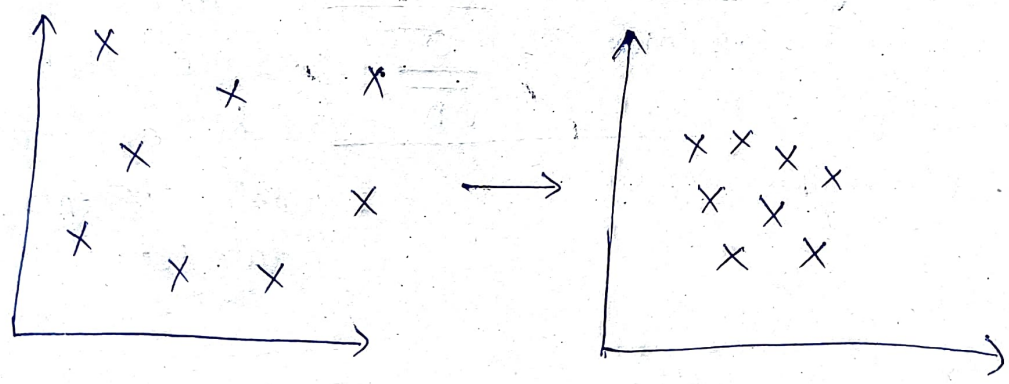
machine learning

maths eqⁿ
⇓ Applied
Algorithm ⇒ mathematical model

Since All data have diff units then the calculation time will be high.

So using std Normal distribution, Applying Z score to Age, Height, and weight will bring all the Scale of data to be same.

ie $\mu=0$, $\sigma=1$ for Age, Height, weight



Bring values to Same Scale

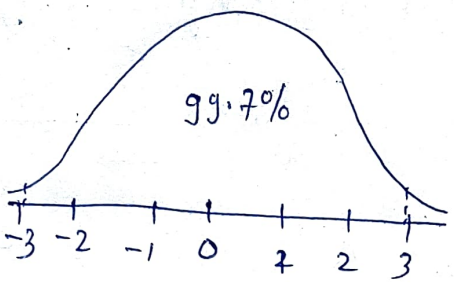
ie Scaling down the data ⇒ Standardisation.

Age ⇒ $\mu = 29.43$

* Standardization { by applying Z score }

$\mu = 0$, $\sigma = 1$.

∴ most values will lie b/w [−3 to 3] range



* Normalization [lower scale ↔ higher scale]

The range is defined by user.

ie [−1, 1]; [−2, 2]; [−3, 3], etc.
[0−1] [0−2]

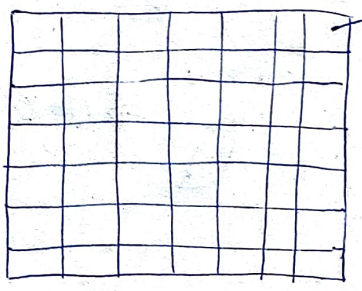
Eg: min-max Scalar. [0 − 1]

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

⇓

This is applied in deep learning and some machine learning problems.

	x	→	y
x_{min}	1		0
	2		0.25
	3		0.5
	4		0.75
x_{max}.	5		1



↑ B/w image

Pixel ranging from (0 to 255) can be converted to 0 − 1 ie Normalization.

R → 0 to 255

G → 0 to 255

B → 0 to 255

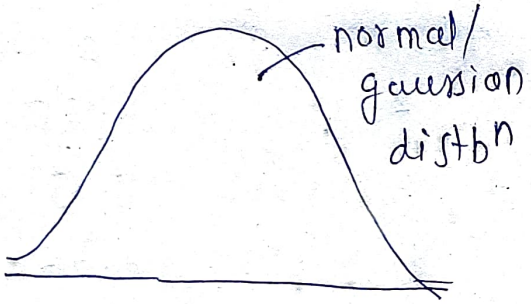
* Standardization } Feature Scaling
* Normalization } Techniques.

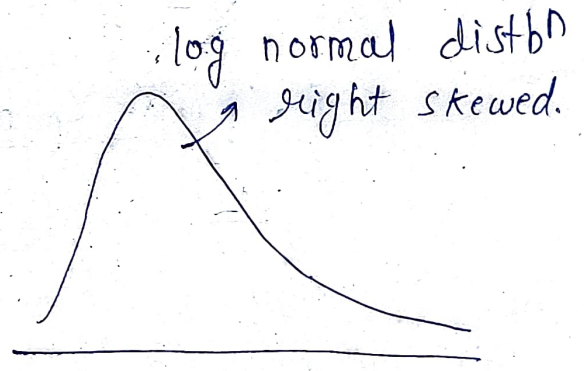
↳ m/c learning → deep learning.

Eg:

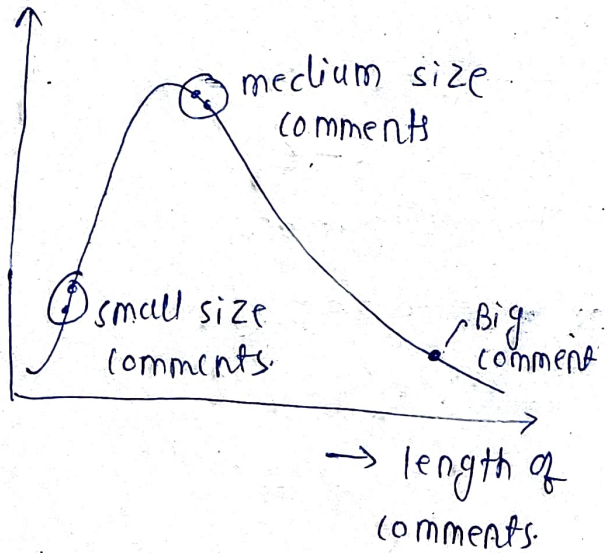
normal distbn x	→	normalization y	→	standardization y' (SND).
1		0		-1.42
2		0.25		-0.72
3		0.5		0
4		0.75		0.72
5		1.		1.42

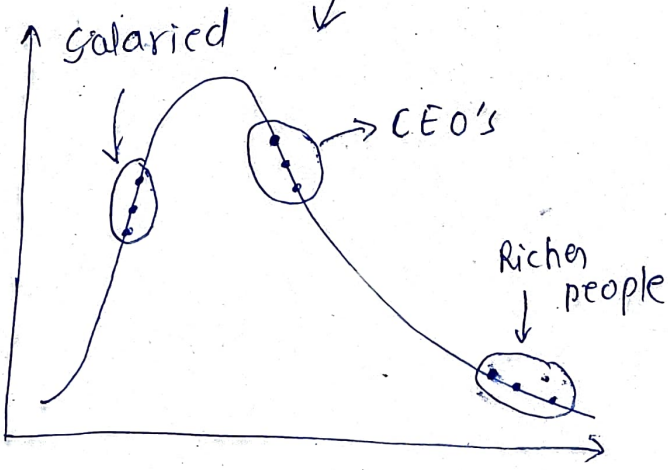
* Log Normal distribution



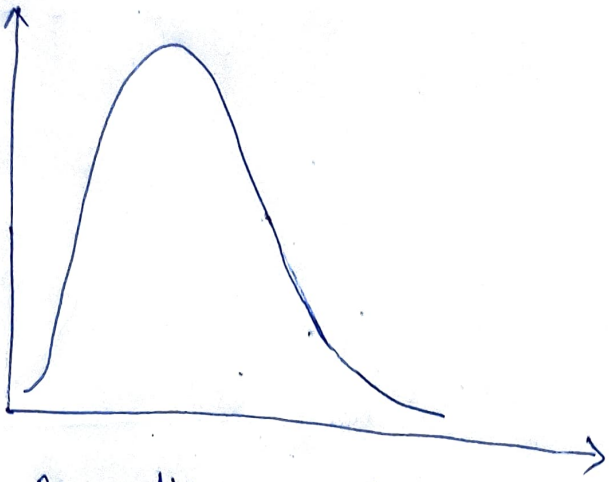


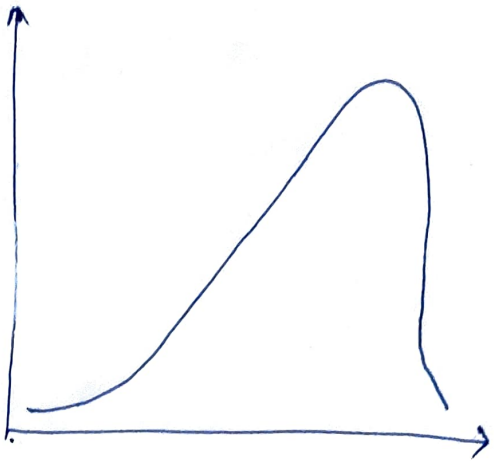
Eg: wealth distribution



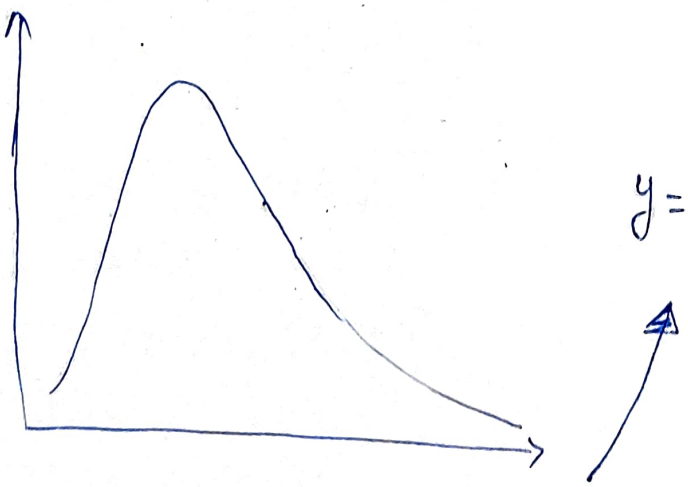


Que: what is relationship of mean, mode median for below graphs.

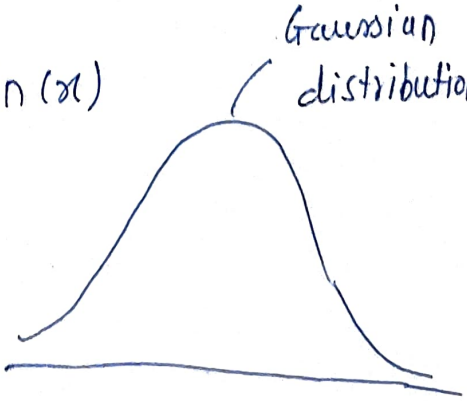




* In Ascending order give relationship of mean, median and mode



$y = \ln(x)$

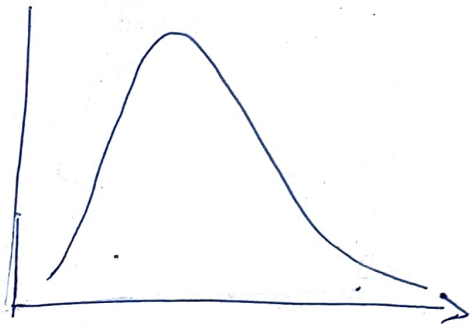


x = log normal distribution

Gaussian distribution $\Rightarrow$ $y = e^x$

$x \sim ND(\mu, \sigma)$.

↳ log normal distribution



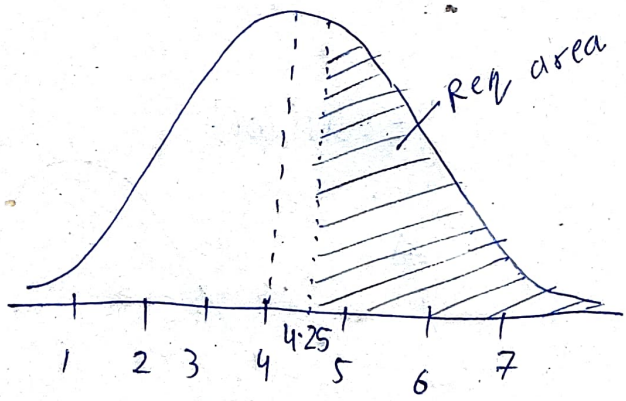
① How to check whether a random variable is

a) log normal distributed.

$\Rightarrow$ $x \approx$ log normal distributed (μ, σ) $\Rightarrow$ if $y = \ln(x)$ is normal distb^n then x is log normal distributed

Qu: ①

$x = \{1, 2, 3, 4, 5, 6, 7\}$ $\left.\begin{matrix} \mu = 4 \\ \sigma = 1 \end{matrix}\right\}$ Assumptions



what is the % of score that falls above 4.25 ???

Area under whole curve = 1.

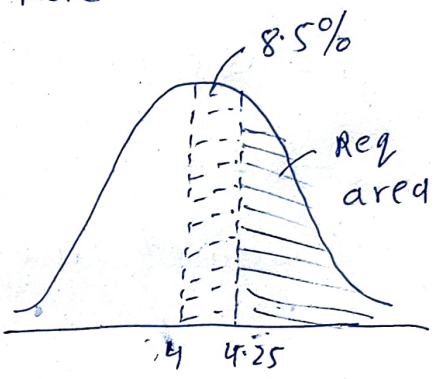
$$Z\text{score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

① Approx calculn

$2\sigma \rightarrow 68\%$ data .... 68-95-99.7 % rule

$0.25\sigma \rightarrow x\%$ data

$$x = \frac{0.25 \times 68}{2} = 8.5\%$$

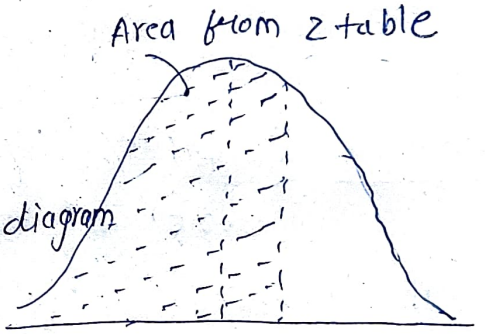


∴ required area will be

$= 50 - 8.5$

$= 41.5\%$

② Exact calculations (z table).



Note: There are different z tables depending on the distribution diagram that they represent.

So it is important to first check the z table diagram that is with the z table for calculation purpose.

From z table $\Rightarrow 0.2 + 0.05$

∴ Area from z table $= 0.59871$

The required area will be $= 1 - 0.59871$

$= 0.40129$

$\approx 40.129\%$ data that falls above 4.25σ.

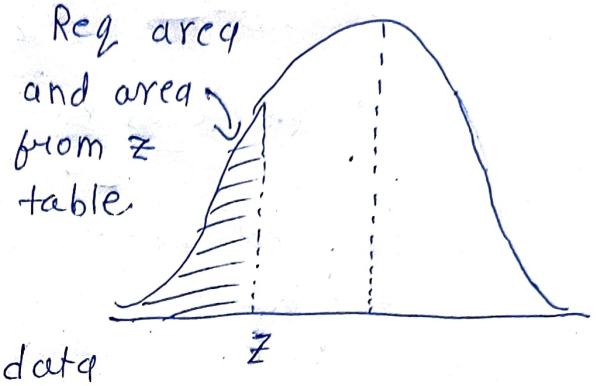
Que: 2 what % of score that falls below 3.75 ? $\mu = 4$, $\sigma = 1$

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

from Z score table (Ztable.net)

$-0.2 + -0.05$

$\therefore$ required area = 0.40129



$\approx$ 40.129 % data that falls below 3.75 σ or -0.25 Z score.

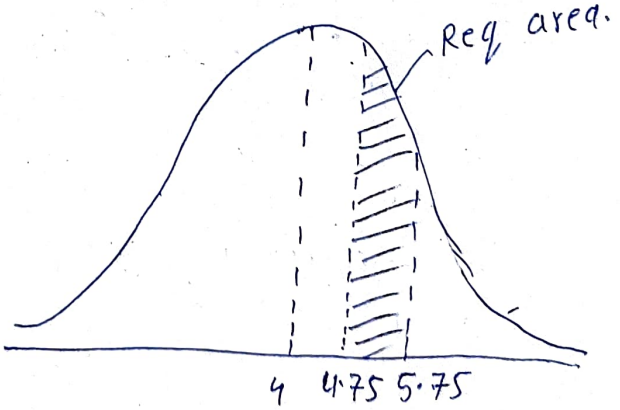
Que: 3 what % of score that falls b/w 4.75 to 5.75 for $\mu = 4$ and $\sigma = 1$?

for 4.75

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

for 5.75

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



* Area upto 4.75 σ. (0.75 Z score).

$0.7 + 0.05$

$\therefore A_1 = 0.77337$

* Area upto 5.75 σ (1.75 Z score).

$1.7 + 0.05$

$\therefore A_2 = 0.95994$

∴ The required area is

$$\text{Area} = A_2 - A_1$$
$$= 0.95994 - 0.77337$$
$$= 0.18657$$
$$\approx 18.657\%$$ of data that falls blw 4.75 and 5.75.

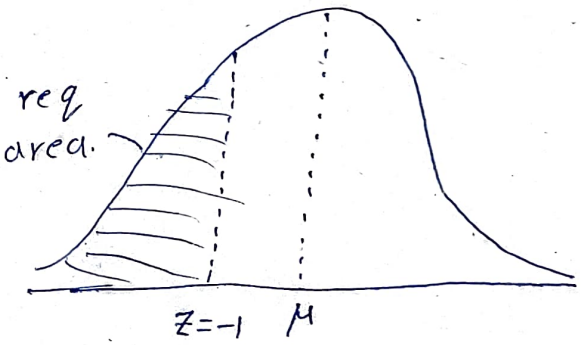
Que: 4 In india Average IQ is 100 with standard deviation 15. what % of population would expect to have an IQ.

1) lower than 85
2) higher than 85
3) Between 85 and 100.

① for 85 (lower than)

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

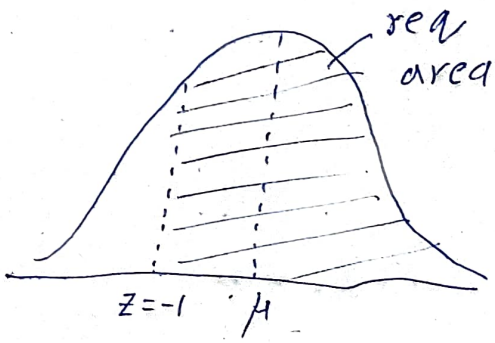
(-1)



∴ required area
$$= 0.15866$$
$$\approx 15.87\%$$

② for 85 (more than).

$$Z\text{score} = -1$$
∴ required area $= 1 - 0.15866$
$$= 0.84134$$
$$\approx 84.134\%$$



③ 85 to 100

Zscore = −1 85

for 100

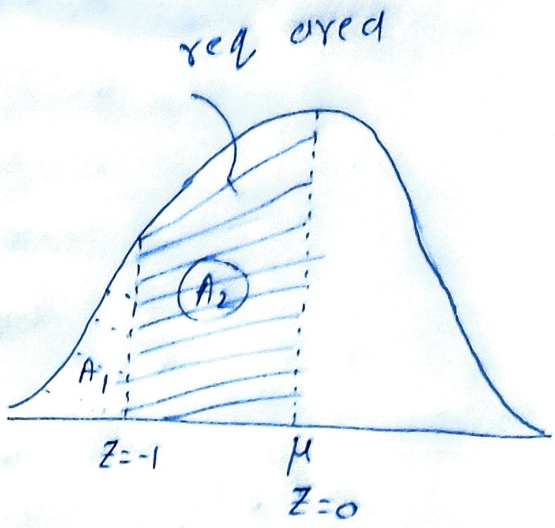
Z score = $\frac{100-100}{15}$ = 0.

for z score = −1

A_1 = 0.15866.

for z score = 0.

A_2 = 0.5



∴ required ared

= $A_2 - A_1$

= 0.5 − 0.15866

= 0.34134

≈ 34.134%.