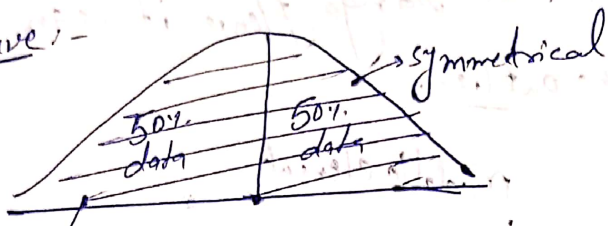
Day 3 Statistics

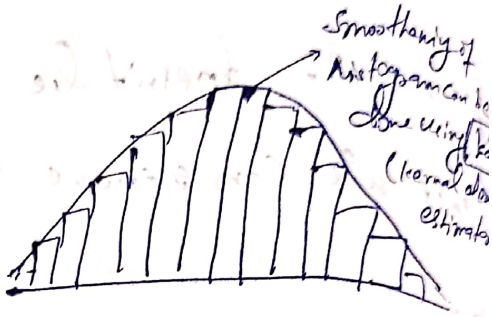
1. Normal Distribution / Gaussian Distribution
2. Standard Normal Distribution
3. Z-Score
4. Standardization and Normalization.

1) Gaussian / Normal Distribution :→

Bell curve :-



→ Area inside this entire Curve is 1 i.e 100%



Most of the data like Age, weight, height they follow this kind of distribution.

↑ Domain Expertise ⇒ Doctors

IRIS DATASET

↓

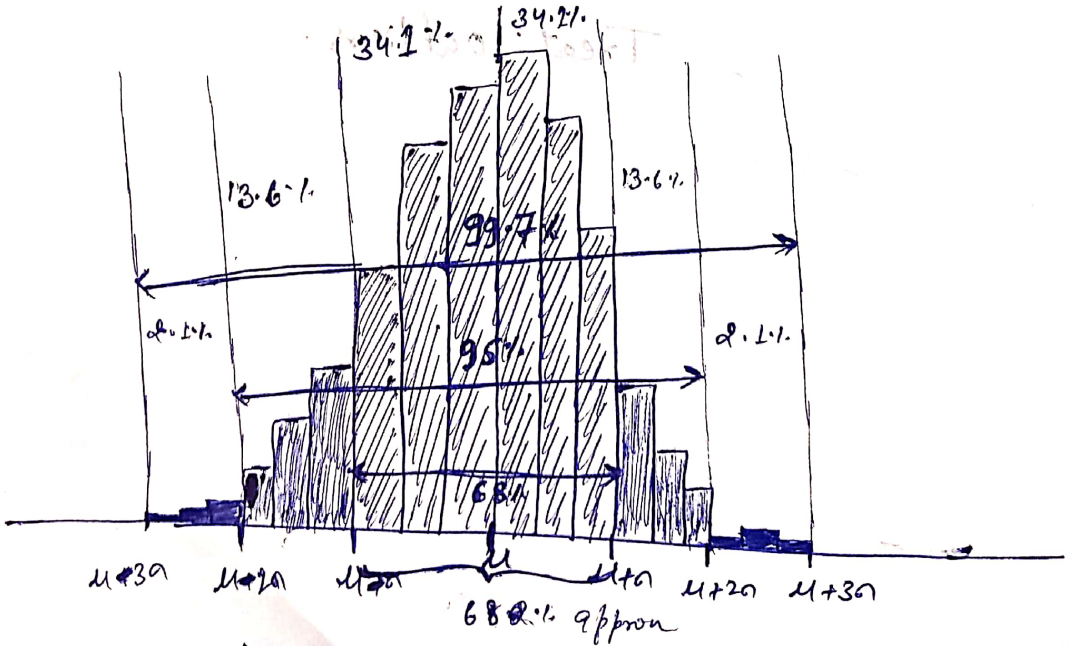
Petal length, sepal length, Petal width, sepal width

↓ follow

Gaussian Dist^r

V.V.I

⇒ Emperical Rule of Normal Distribution :→



~~Empirical formula (68-95-99.7%) :→~~

~~68% of Data~~

Empirical formula (68-95-99.7%) rule :→

68% of Entire data present in $(\mu-\sigma, \mu+\sigma)$ region.

95% of Entire data present in $(\mu-2\sigma, \mu+2\sigma)$ region.

99.7% of Entire data present in $(\mu-3\sigma, \mu+3\sigma)$ region.

and remainig all will be the outside the $(\mu-3\sigma, \mu+3\sigma)$ region.

and this rule is called (68-95-99.7% rule) or Empirical formula.

Q-Q Plot → How to determine whether the Distr is Gaussian or not? (will do it in Practicle)

2) Standard Normal Distribution :→

$X \sim$ Gaussian Distribution (μ, σ) ← Random var

convert ↓ ← Why converting?

$y \sim$ Standard Normal Distr $(\mu=0, \sigma=1)$ ← how?

↓ simple formula

Z-score

↓

sample scaling or standardisation.

Eg:- $X = \{1, 2, 3, 4, 5\}$

$\mu = 3$

$\sigma = 1.41$

$n=1$ ← Bcz we are going to apply this formula on each & every value.

$$\text{Z-score} = \frac{X_i - \mu}{\sigma/\sqrt{n}}$$ → Standard error (useful while doing Inferential stats)

if $n=1$ then,

$$\text{Z-score} = \frac{X_i - \mu}{\sigma}$$

there are cases where n can be different than '1' we will see that later.

Now Apply z-score on $x = \{1, 2, 3, 4, 5\}$

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

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

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

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

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

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

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

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

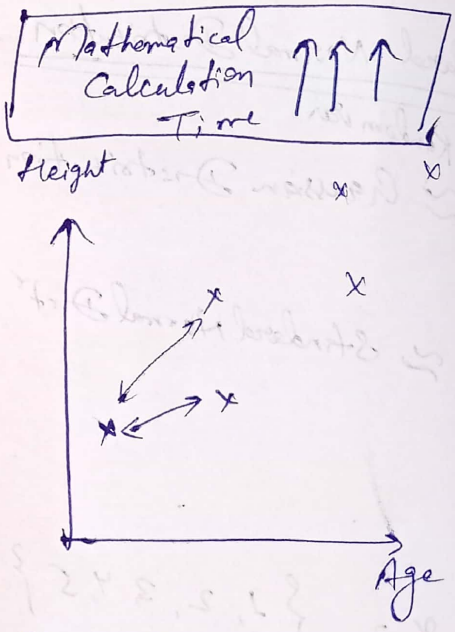
N.V.D

Q. Why are we converting Gaussian Distr to Normal standard Distr?

Sol

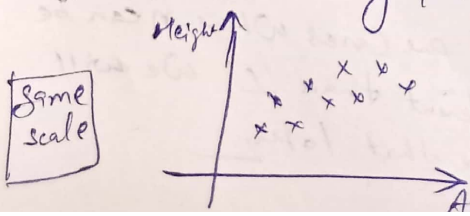
Example :→

(years) Age	(kgs) Weight	(cm) Height
24	72	180
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	178



Data Points Quiet spread it will take time to calculate the distance b/w two or for any mathematical Calculation it will take time.

Hence if we try to bring the data points within the same scale so that it will be nearby populated



So it will have advantage. it will take less time in mathematical calculation. So with this we try to bring Data

Hence by applying Z-score formula we bring the data in same scale i.e will be ranging b/w -3 to +3 (most of the values) & $\mu = 0$ & $\sigma = 1$

Bcz of this calculation will be faster.

and we are not loosing any data.

$\mu = 0, \sigma = 1$ → N.V.D

And the entire Process is called **Standardization**.

After training the model we can revert the scale back to get the original data. for that again we will have to apply the same formula

Standardization :→ W.r.t every feature we apply the Z-score to scale down the value with $\mu = 0, \sigma = 1$.

Age = {24, 26, 32, 33, 34, 28, 29}

$\mu = 29.42$

$$\sigma^2 = \frac{(24-29.42)^2 + (26-29.42)^2 + (32-29.42)^2 + (33-29.42)^2 + (34-29.42)^2 + (28-29.42)^2 + (29-29.42)^2}{7}$$

$$= \frac{29.37 + 11.69 + 6.65 + 12.81 + 20.97 + 2.016 + 0.176}{7}$$

$$= 11.9545$$

$$\sigma = \sqrt{11.9545} = 3.45$$

$\sigma = 3.45$

$X_{(Age)} = \{24, 26, 28, 29, 32, 33, 34\}$

$$Z\text{Score} = \frac{24-29.42}{3.45}, \frac{26-29.42}{3.45}, \frac{28-29.42}{3.45}, \frac{29-29.42}{3.45}, \frac{32-29.42}{3.45}, \frac{33-29.42}{3.45}, \frac{34-29.42}{3.45}$$

$$Y_{(Age)} = \{-1.57, -0.99, -0.41, -0.12, 0.74, 1.03, 1.32\}$$

$\mu = 0 \quad \sigma = 1$

Normalization :→

Convert the value b/w any particular range you give that could be [0 → 1]

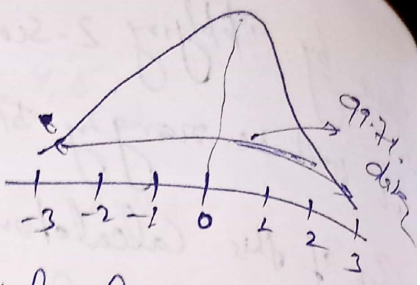
[0,1] or [0,5], [0,4], [-1 1]

Standardization

[μ: 0, σ: 1]

[-3 ⟷ 3]

↳ 99.7% of data lies b/w -3 to 3 range.



Normalization :→ [Lower Scale ⟷ Higher Scale]

← Range

① Min Max Scaler [0,1]

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

Converting → apply in CNN,

x	y	
1	0	$\frac{(1-1)}{5-1} = 0$
2	0.25	$\frac{(2-1)}{5-1} = 1/4 = 0.25$
3	0.5	$\frac{(3-1)}{4} = 2/4 = 0.5$
4	0.75	$(4-1)/4 = 3/4 = 0.75$
5	1	$(5-1)/4 = 4/4 = 1$

got converted in the range of [0,1].

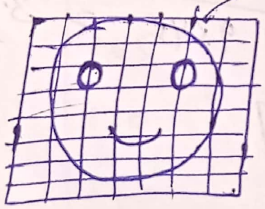
Q. Where do we apply this?

Ans. In deep learning & some of the machine learning techniques.

In Deep learning = Images pixels range (0, 255)

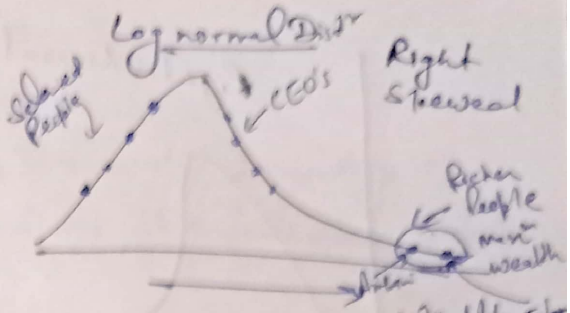
↓ Convert (Scale down)

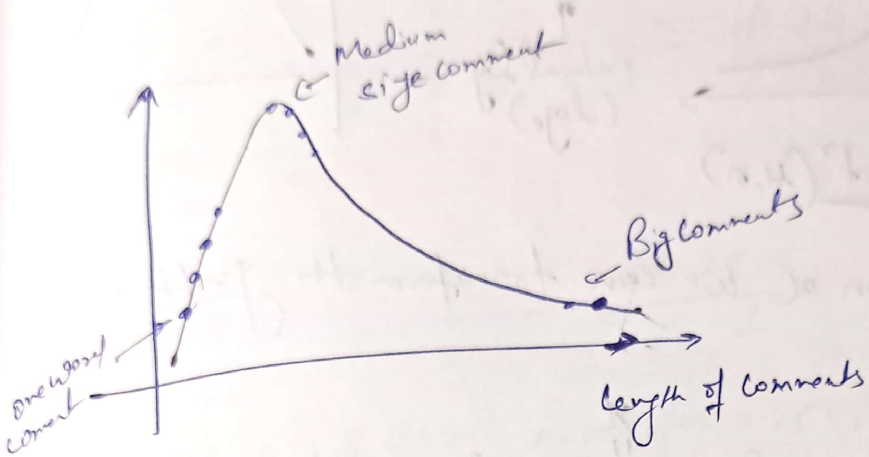
(0,1) ← Normalization



Feature Scaling ← Standardization
Normalization

1) Log Normal Distribution :→





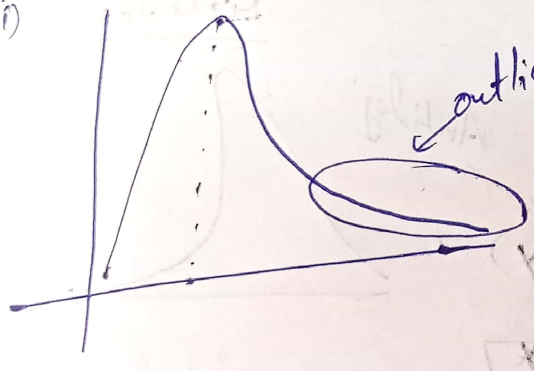
Assignment

Q) What is the relationship of Mean, Median & Mode in this two graphs:

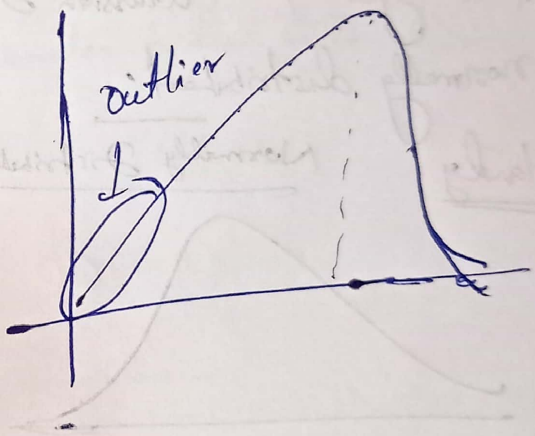
Left skewed

Right skewed

(i)



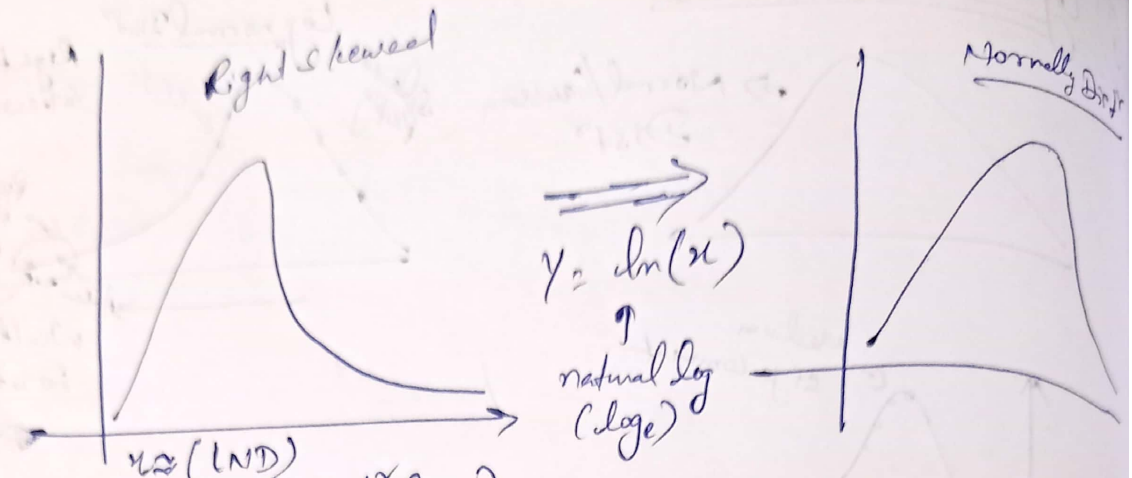
(ii)



if outliers is biggon
no. it affects the mean
but doesn't affect median.

In (i) outliers are +ve bigger
hence mean will be
bigger. but in case of (ii) outliers are
smaller nos. Hence mean will be small.

mean(i) > mean(ii)



$x \simeq$ log-normally Distr^n (μ, σ)

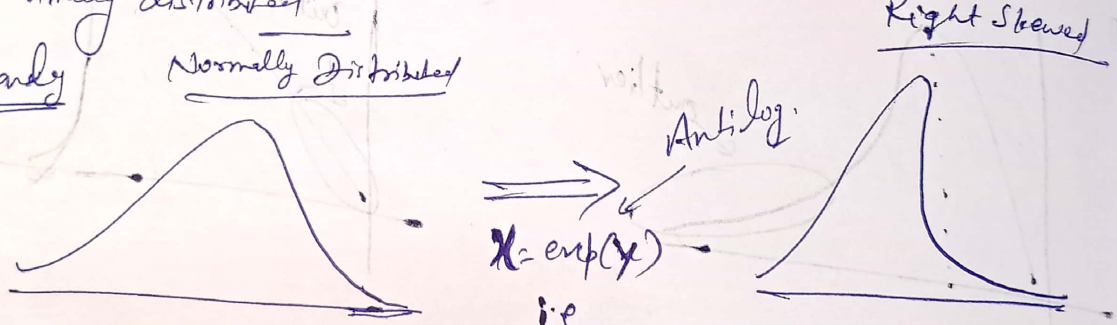
If we perform log on x we can transform the graph in Gaussian Distr^n

x
=

⇒ If the random variable x is log normally distributed, then $y = \ln(x)$ has a normal distribution

⇒ If we get y as Gaussian Distr^n then we can say that x is log-normally distributed.

* Similarly



$y \simeq$ normally distributed

ND (μ, σ)

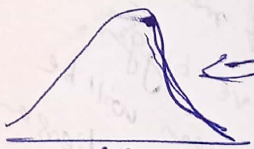
how to know it is normally distr.

In short

$x \simeq$ LND

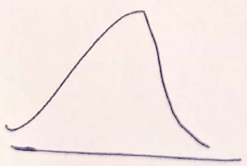
Eg: No. of runs

⇒ $y = \log_e(x)$ or $y = \ln(x)$ ⇒



QQ Plot

(Gaussian Distr^n) (GD)



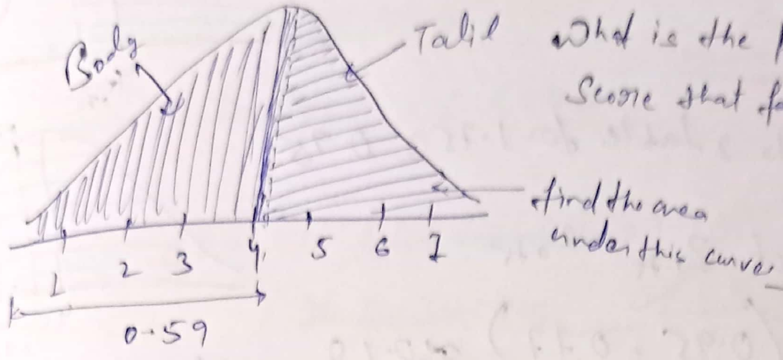
$x =$ (Gaussian Distr^n) (GD)

⇒ $y = \exp(x)$

$y = e^x$

(Log Normal Distr^n) (LND)

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



What is the percentage of Score that falls above 4.25?

Sol^n

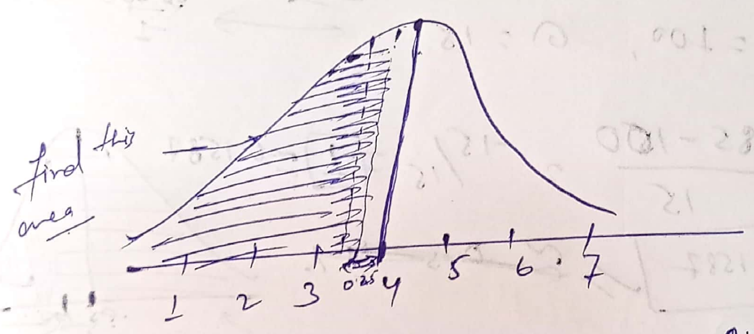
$$z\text{-Score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = \boxed{0.25}$$

now look into Z table look for 0.25 i.e 0.59

Hence area ~~under~~ above 4.25 = ~~1.59~~ 1 - 0.59 = 0.41

= 41% Ans

Q. Area under curve that falls ~~under~~ below 3.75?



Sol^n

$$z\text{-Score} = \frac{3.75 - 4}{1} = \boxed{-0.25}$$

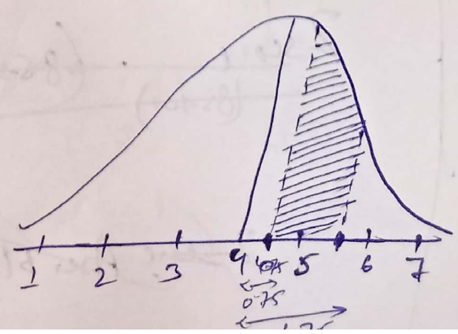
now look into z-table & look for -ve table for -0.25 i.e 0.40

≈ 40% Ans

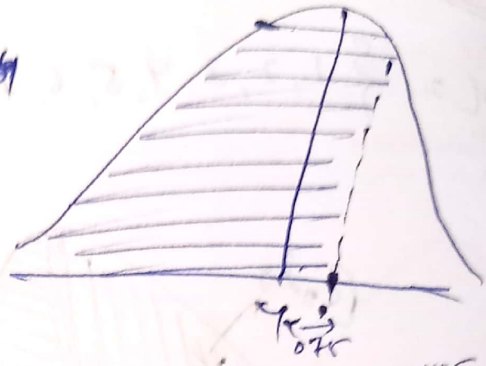
Q. Find the area of curve b/w 4.75 & 5.75

$$z\text{-Score} = \frac{4.75 - 4}{1} = 0.75$$

$$z\text{-Score} = \frac{5.75 - 4}{1} = 1.75$$



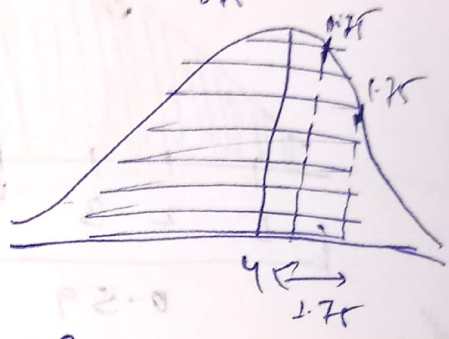
Now look into z table for $0.75 \approx 0.7734$



Now look into z table for $1.75 \approx 0.95$

Now Subtract $Z_{1.75} - Z_{0.75}$

$$(0.95 - 0.77) \approx 0.18 = 18\% \text{ Ans}$$

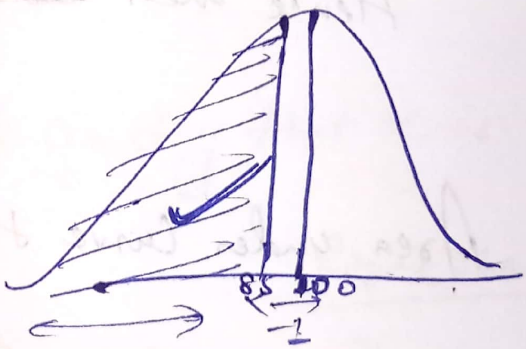


Q) In India the Avg IQ is 100 with a S.D of 15 what is the percentage of population Could you expect to have an IQ

i) lower than 85 = 0.1587

ii) higher than 85 = 0.8413

iii) Between 85 & 100 = 0.3413

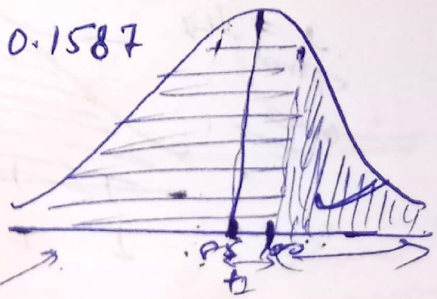


Sol

$\mu = 100, \quad \sigma = 15$

$$Z_{score} = \frac{85 - 100}{15} = -15/15 = -1 \approx 0.1587$$

Area below 85 $\approx$ 0.1587

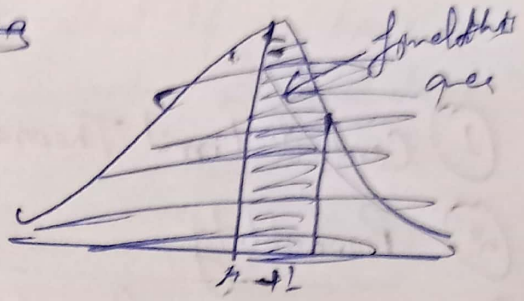


Higher than 85 = $(1 - 0.1587) = 0.8413$

Area above

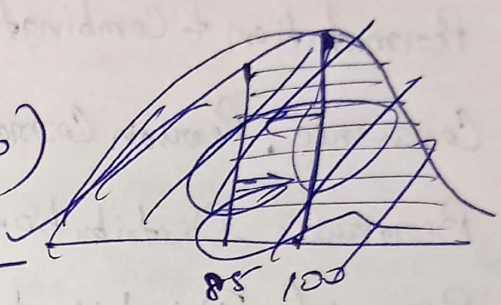
Blw 85 & 100

Area ~~higher~~ lower than 85 = ~~0.8413~~ 0.1587



$Z_{score} = \frac{100 - 100}{15} = 0$

Area below 100 = 0.5000



Now Area blw 85 & 100 = (Area below 100 − Area below 85)

$(0.5000 - 0.1587) \approx 0.3413$ Ans

85 < scores < 100