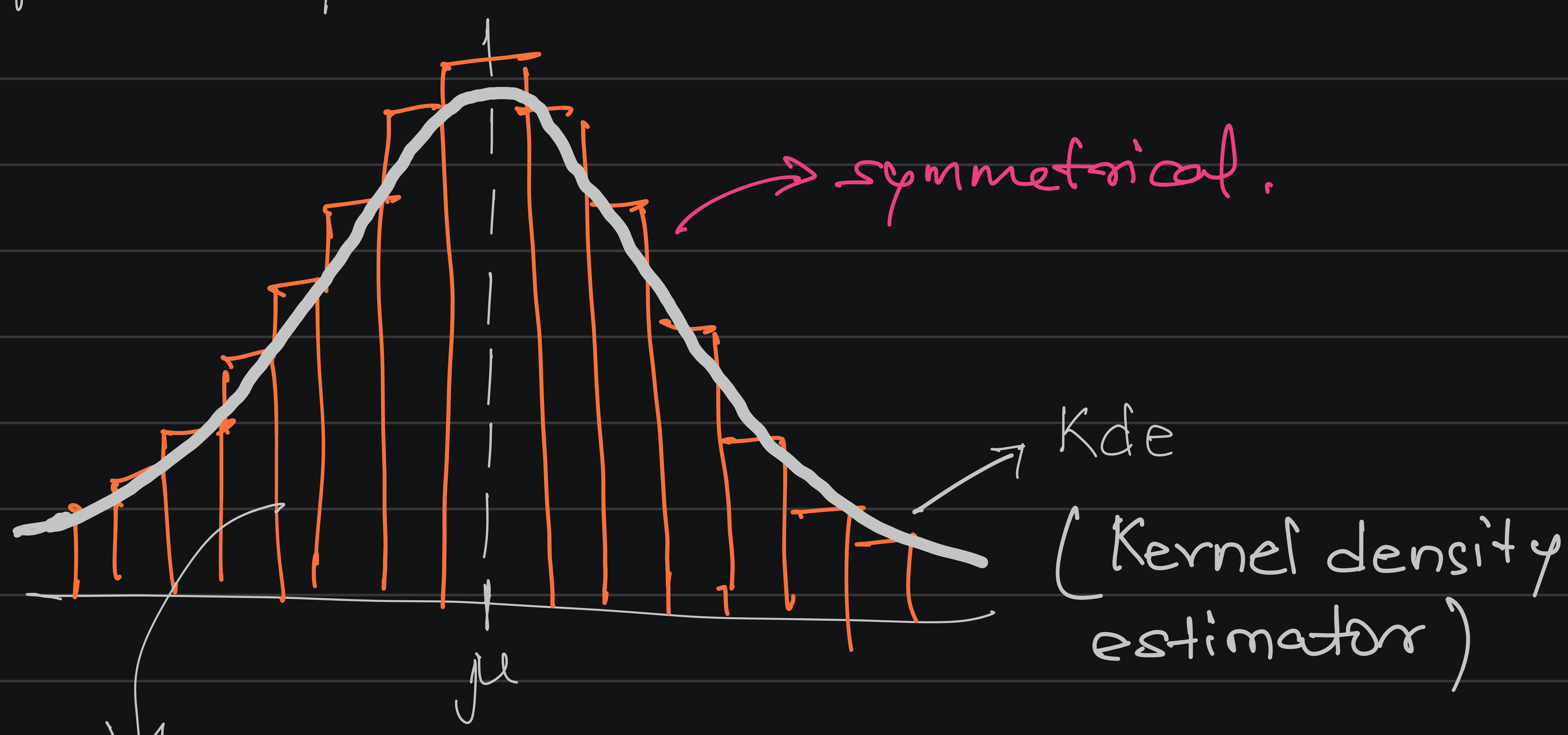
Day 3

Agenda.

1. Normal distribution.
2. Standard Normal distribution.
3. Z - Score
4. Standardization & Normalization

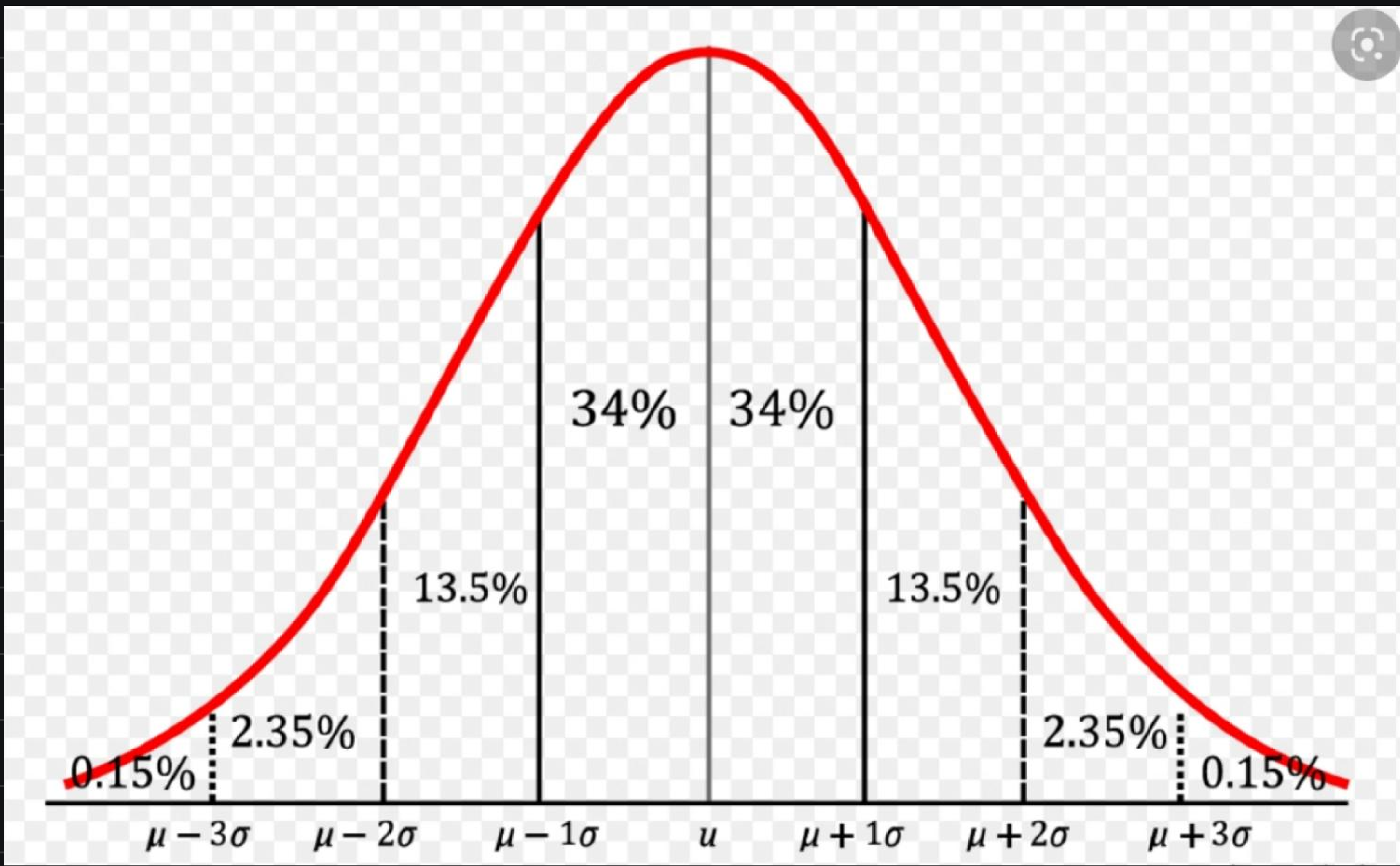
1. Gaussian / Normal distribution.



Area under curve = 1 (100%)

eg. Age, Weight, Height. → follow Normal distribution.

Empirical Rule of Normal distribution.



→ Assumptions

(i) 68% of data falls within 1 SD (Standard deviation)

(ii) 95% of data falls within 2 SD. (σ)

(iii) 99.7% of data falls within 3 SD (σ)

$\boxed{\text{68 - 95 - 99.7\% rule}}$

→ Empirical Rule.

(*)

$\boxed{\text{Q-Q plot}}$ → Can show if a distribution is Gaussian or not.

(2.) Standard Normal Distribution.

if x belongs to Gaussian Distribution(μ, σ)

⇓ can be converted into. using Z Score

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

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

$\mu = 3$

$\sigma = 1.41$

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

n = 1, applied on every element

⇒ Standard Error.

So, $\boxed{Z\ \text{Score} = \dfrac{x_i - \mu}{\sigma}}$

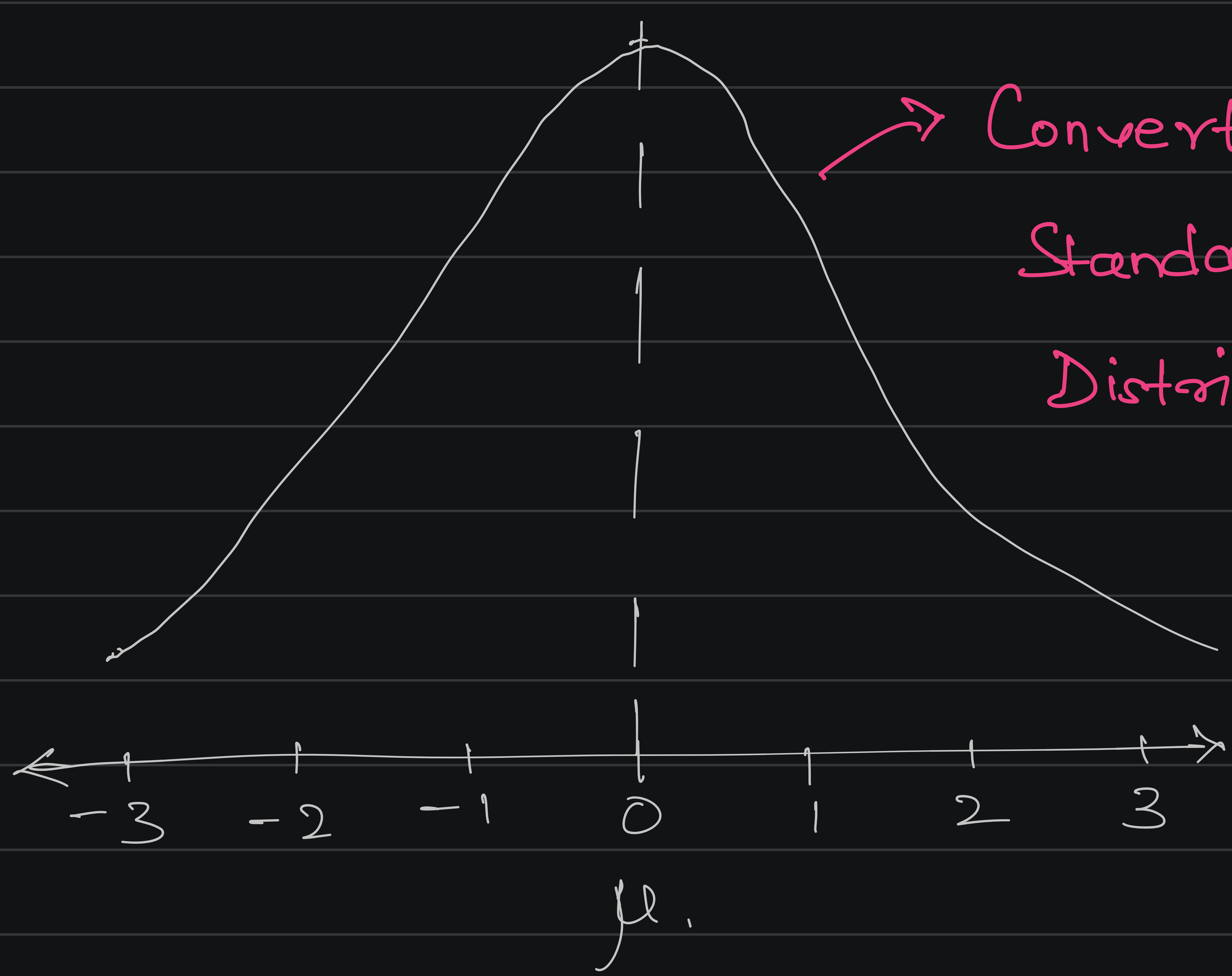
$y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

for 1st element Z-score $= \dfrac{1-3}{1.414} = -1.414$

for 2nd element Z-Score $= \dfrac{2-3}{1.414} = -0.707$

for 3rd element Z-score $= \dfrac{3-3}{1.414} = 0$

Distribution of y.



Converted into Standard Normal Distribution.

Q Why do we need to convert ?

3 features →

Age (years)	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.

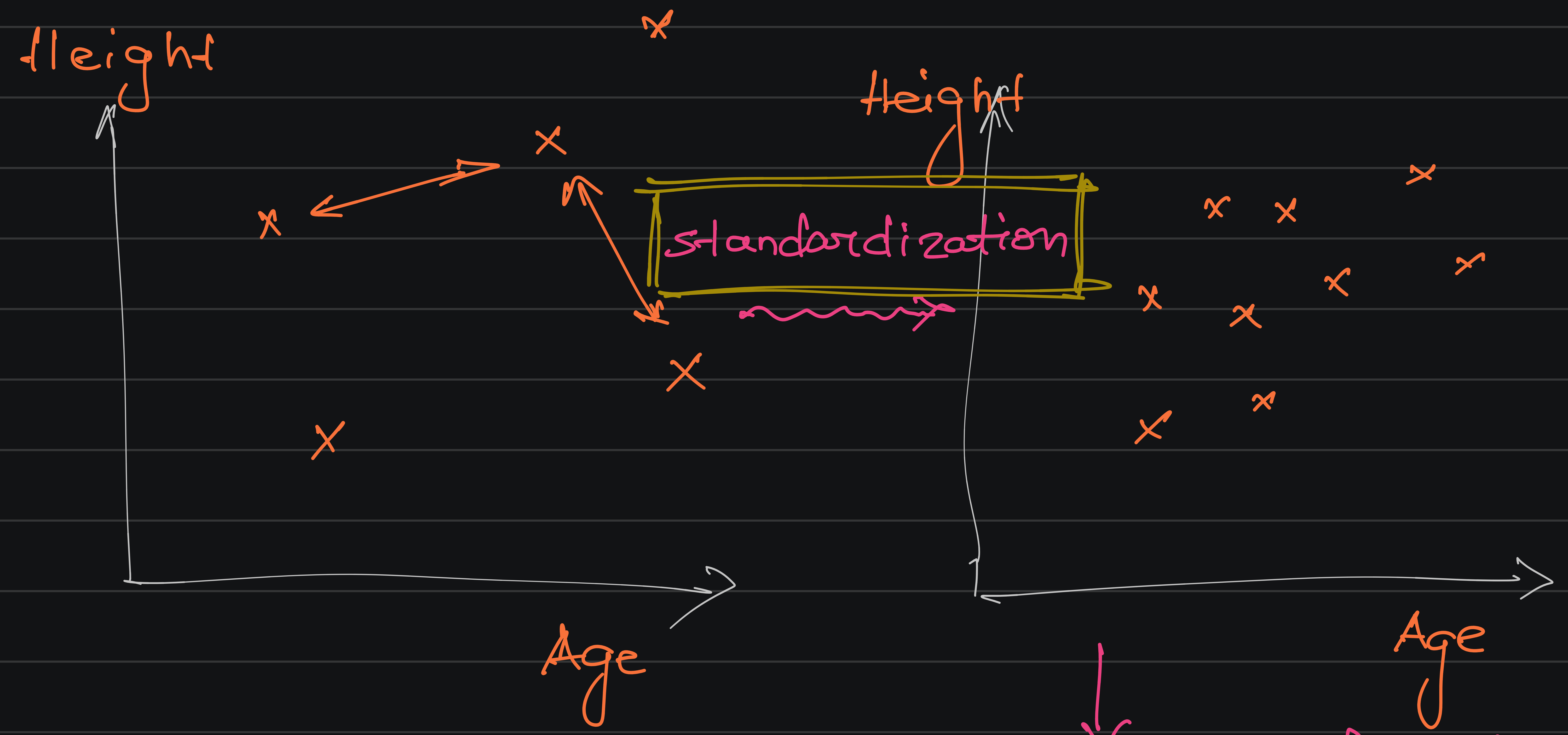
Units are different

⇓

Value ranges differ a lot.

↓

Not favorable of ML models.



More helpful for models.

★ Apply Z-score on all features to bring them all to the same scale (-3 to 3) & $\mu = 0$ & $\sigma = 1$

Feature Scaling

1) Standardization :- { z-score applied }

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

data = { -3 to +3 } 99.7% data

2 In Normalization → we give the range within which we need to convert the values.

eg. 1 Min-max scaler — tranform values between 0 to 1.

Formula,

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

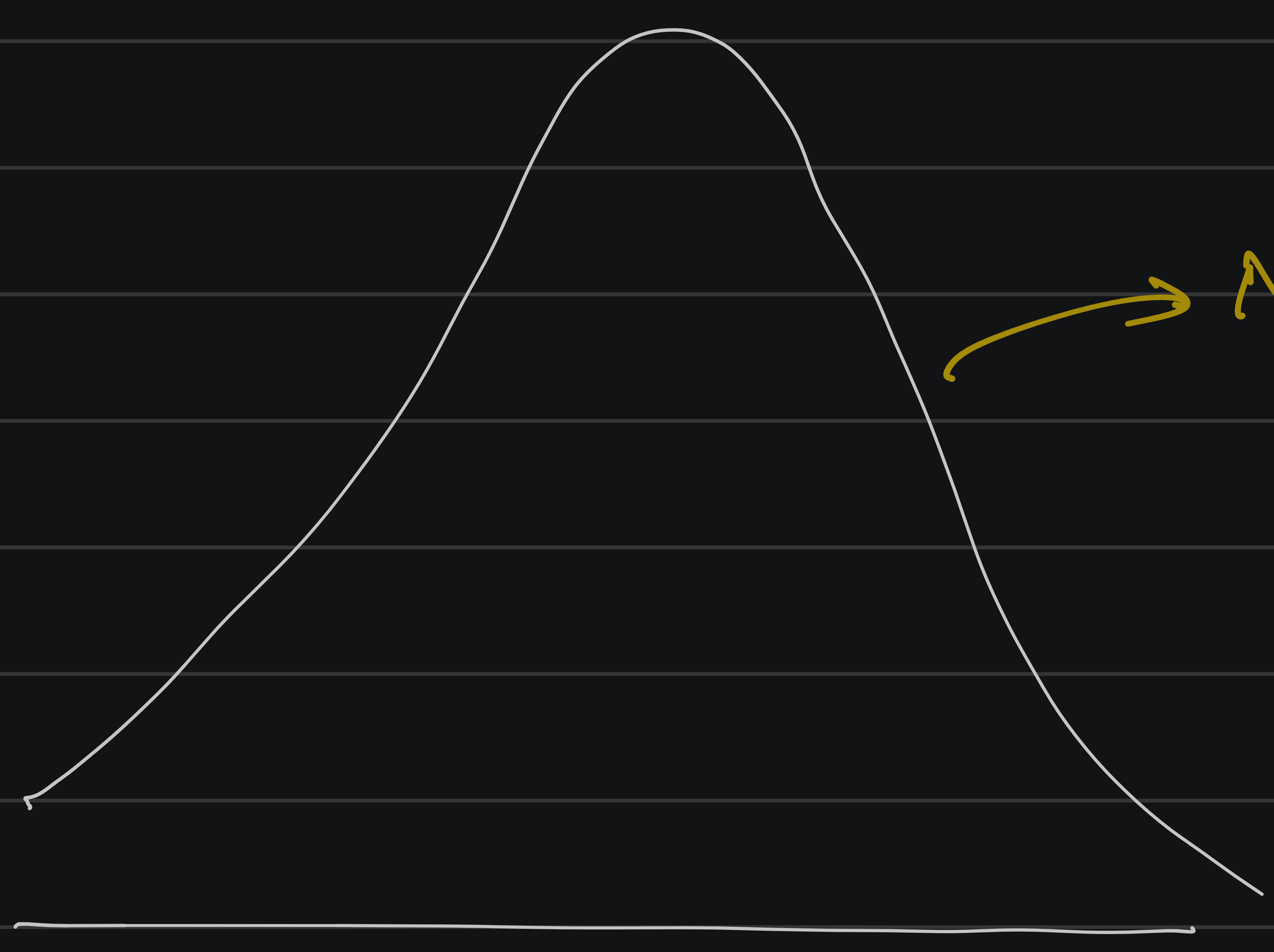
where,
$x_{max} = 1$
$x_{min} = 0$

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

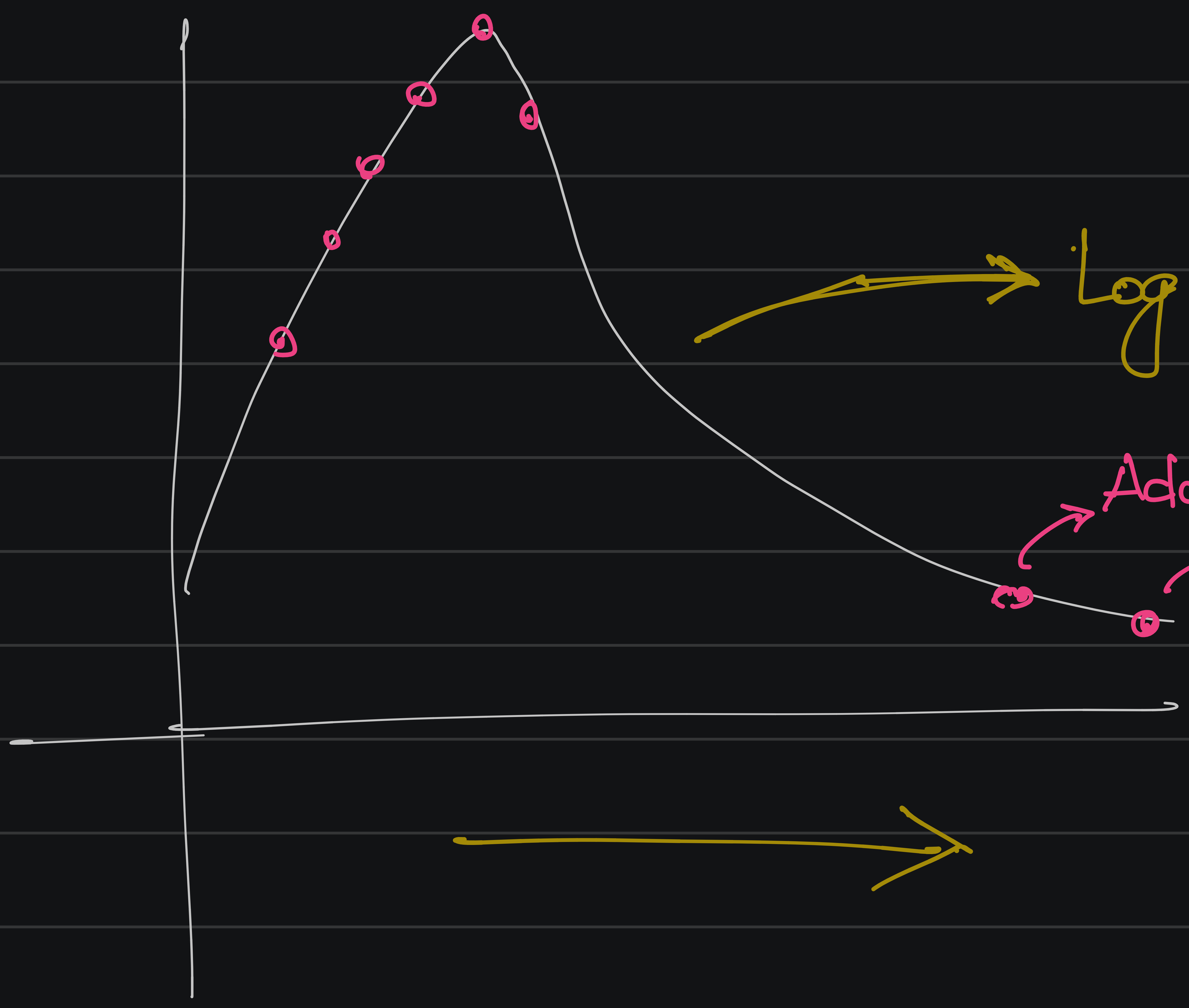
All values converted to 0 to 1.

Applied in Deep Learning

Log Normal Distribution.



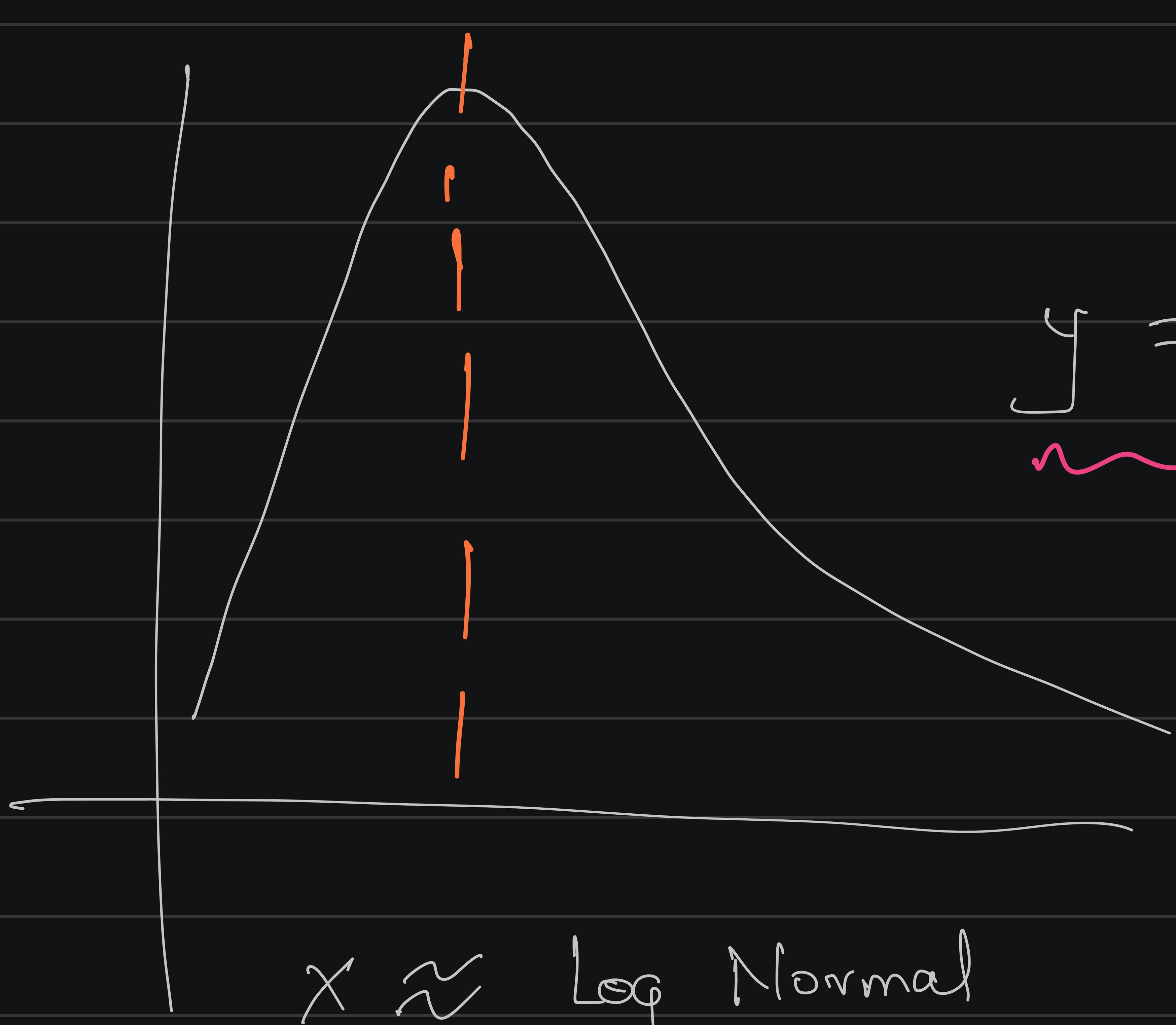
Normal distribution.



Log normal distribution.

eg - weath distribution.

Length of comments



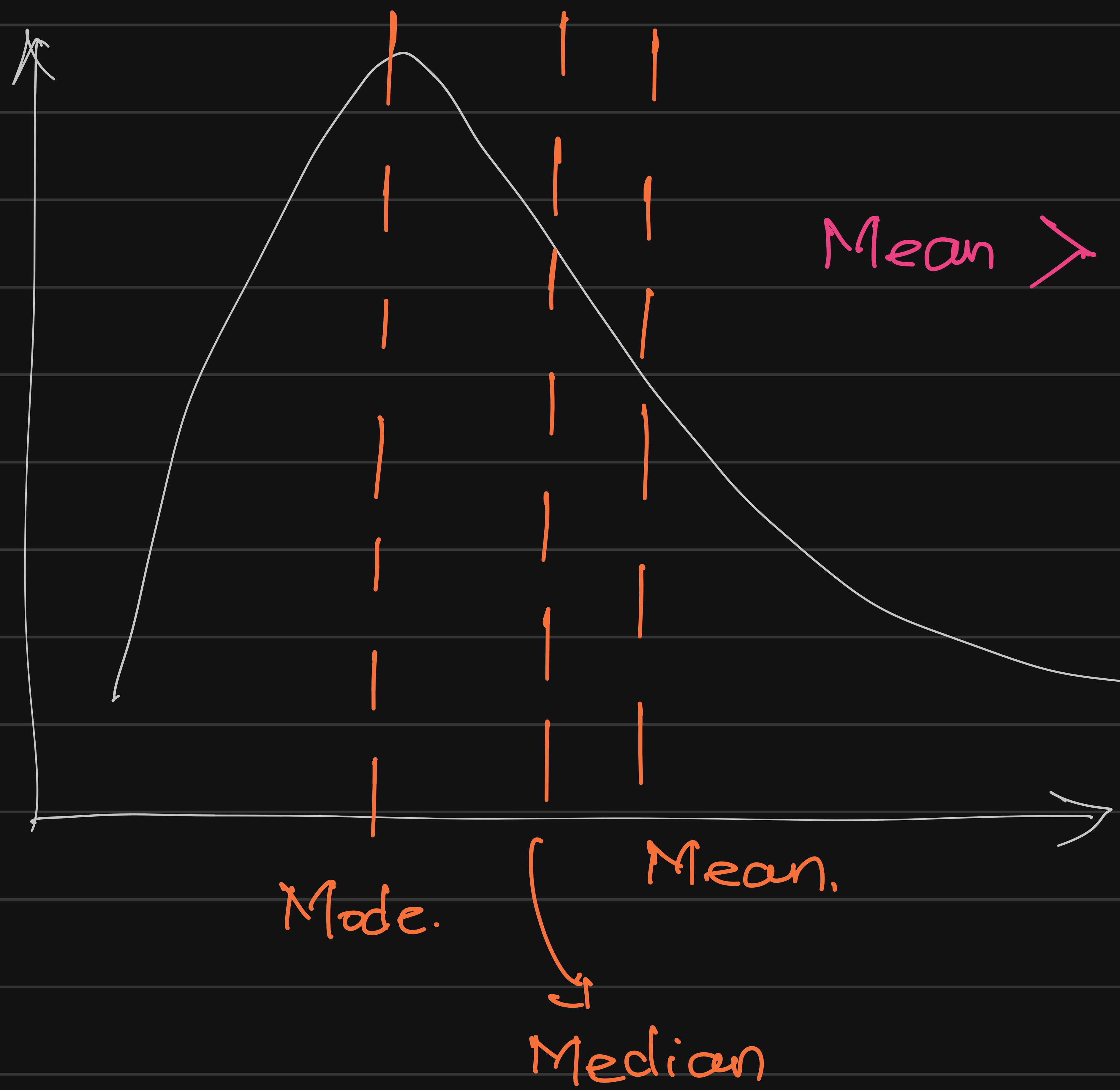
$X \approx$ Log Normal distribution.

$$y = \ln(x)$$

Converts into normal distribution.

Q = What is the relationship between mean, median and mode for below distributions

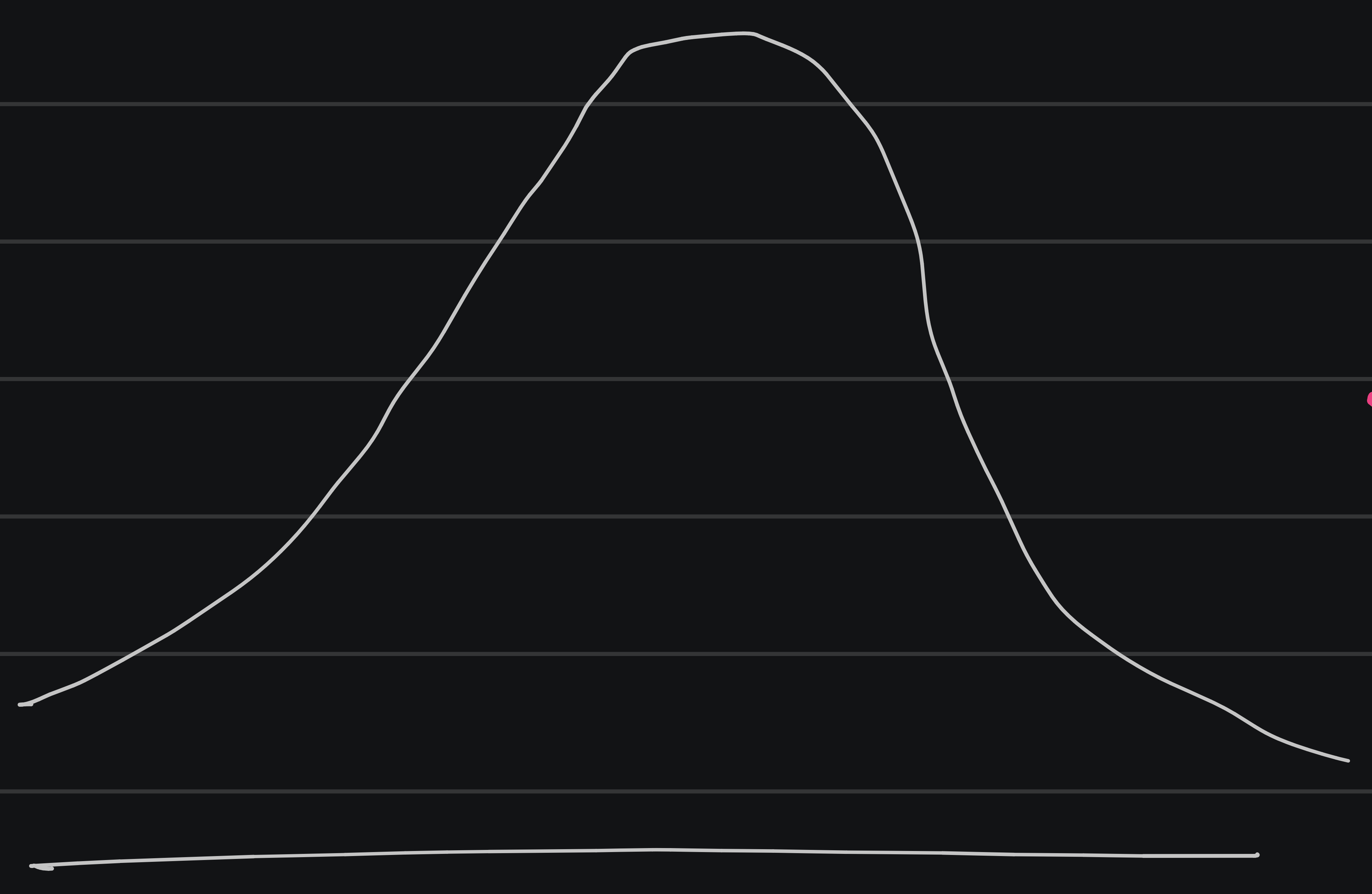
1



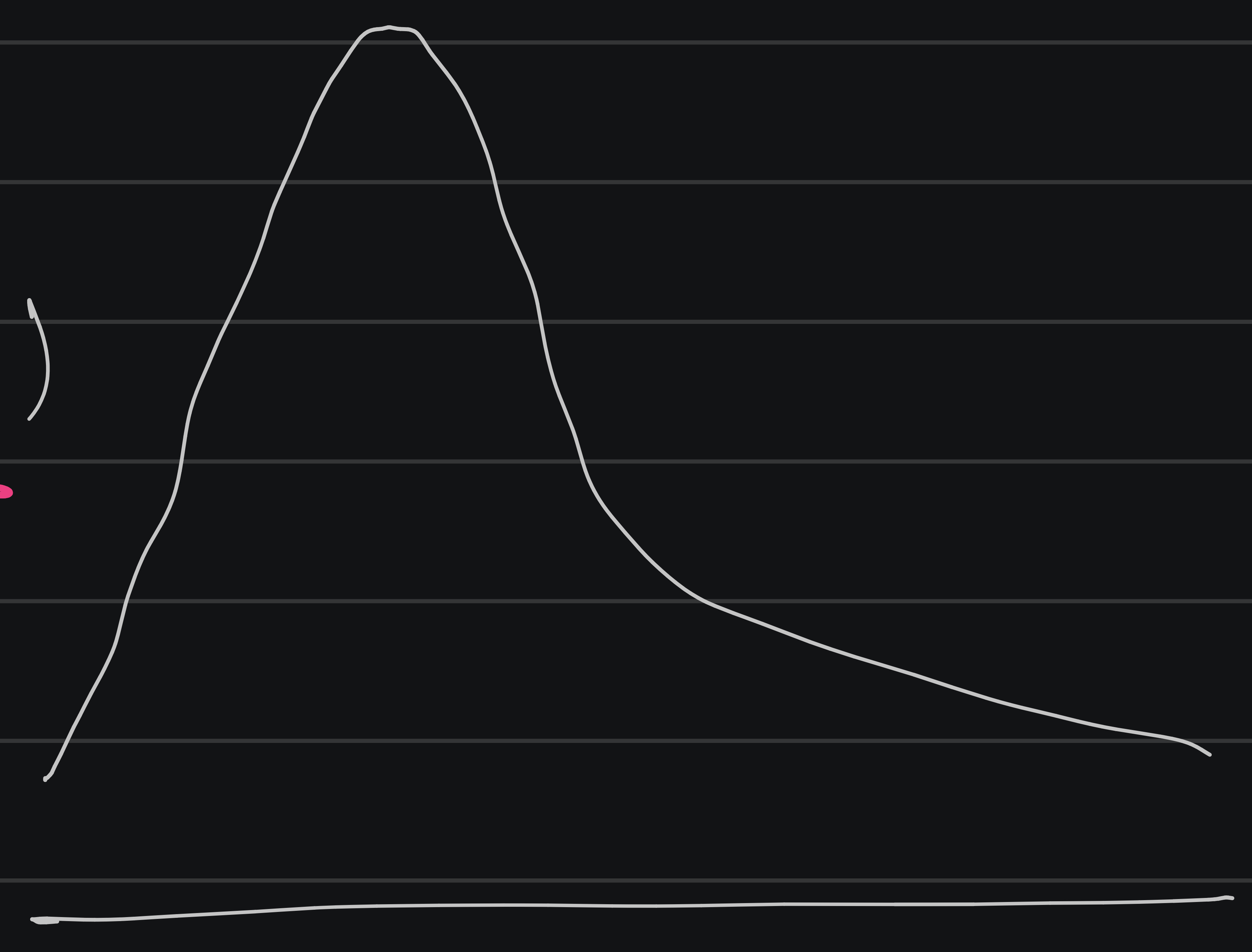
Mean > Median > Mode.

2

Mean < Median < Mode.



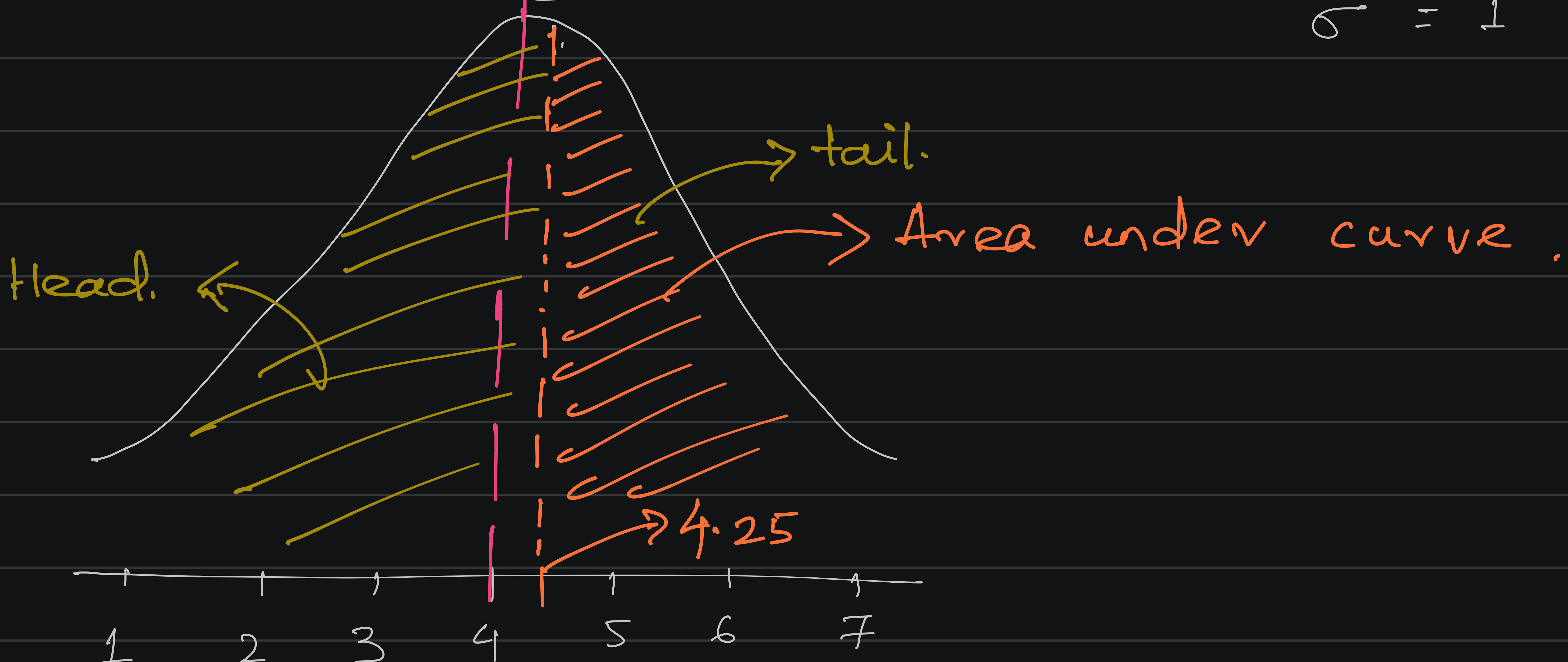
$y = \exp(x)$



$X \approx$ Normal distribution (μ, σ)

Log normal distribution.

<u>Problem</u>: $x = \{1, 2, 3, 4, 5, 6, 7\}$, $\mu = 4$, $\sigma = 1$



<u>Question</u> — What is the percentage of score that falls above 4.25 ?

<u>Answer</u>: 1. Apply Z-score $= \dfrac{x_i - \mu}{\sigma}$

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

(2.) Use Z-table to find area under curve.

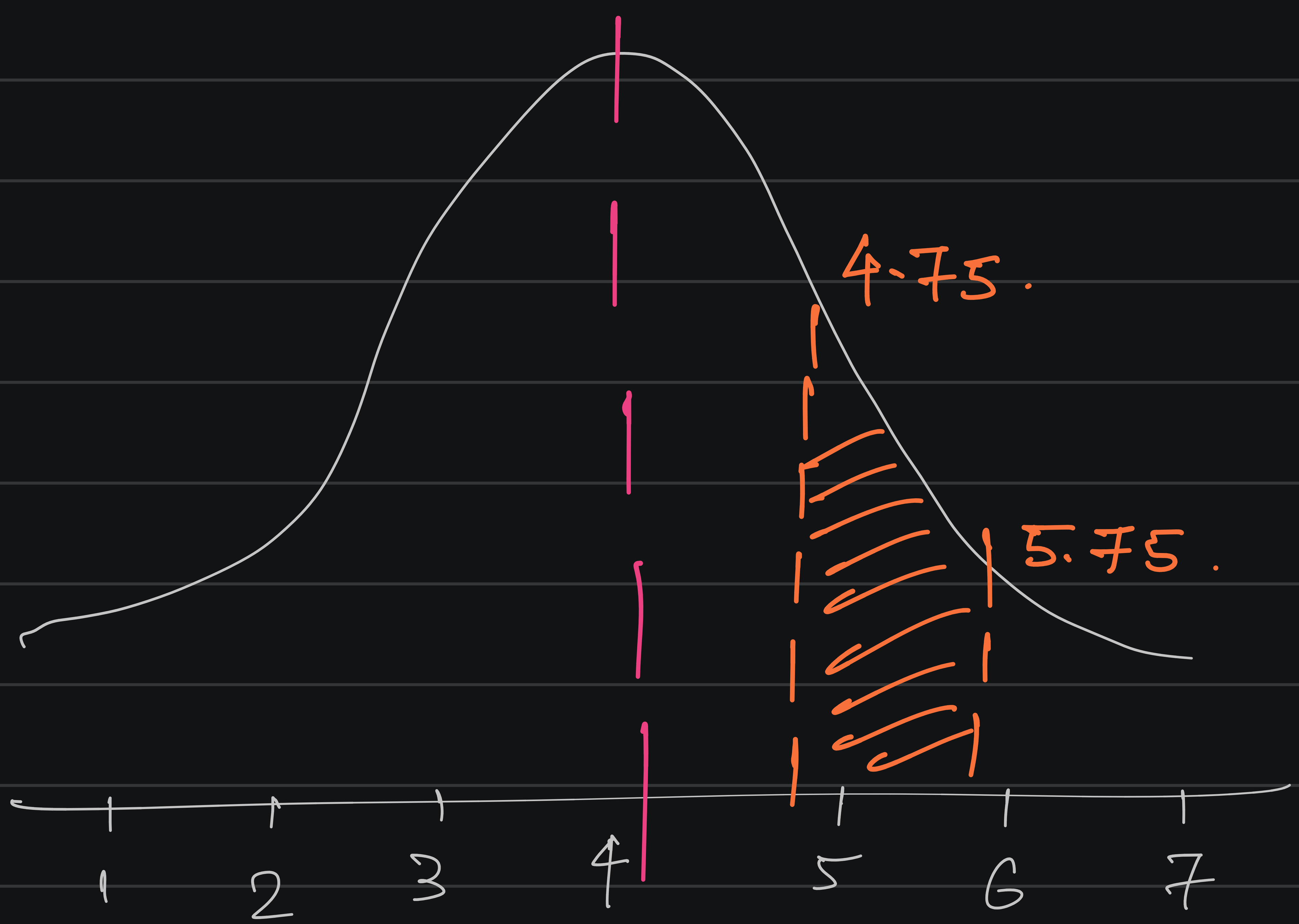
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189

0.59 $\Rightarrow$ 59% $\rightarrow$ Area under curve for head.

Area under curve (tail) = (1 − 0.59)

= 41%

Q Area under curve between 4.75 & 5.75.



Z Score (5.75) = 1.75. = 77.34% (Head)

Z Score (4.75) = 0.75. = 95.99% (Head)

Area under curve (5.75 tail) = 22.66%

Area under curve (4.75 tail) = 4.01%

Area under curve (5.75 − 4.75).

= 22.66 − 4.01 = **18.65%**

Q In India the average IQ is 100 with standard deviation of 15. What is the percentage of population would you expect to have an IQ.

1. Lower than 85. — 16%
2. Higher than 85. — 84%
3. Between 85 & 100. — 34%