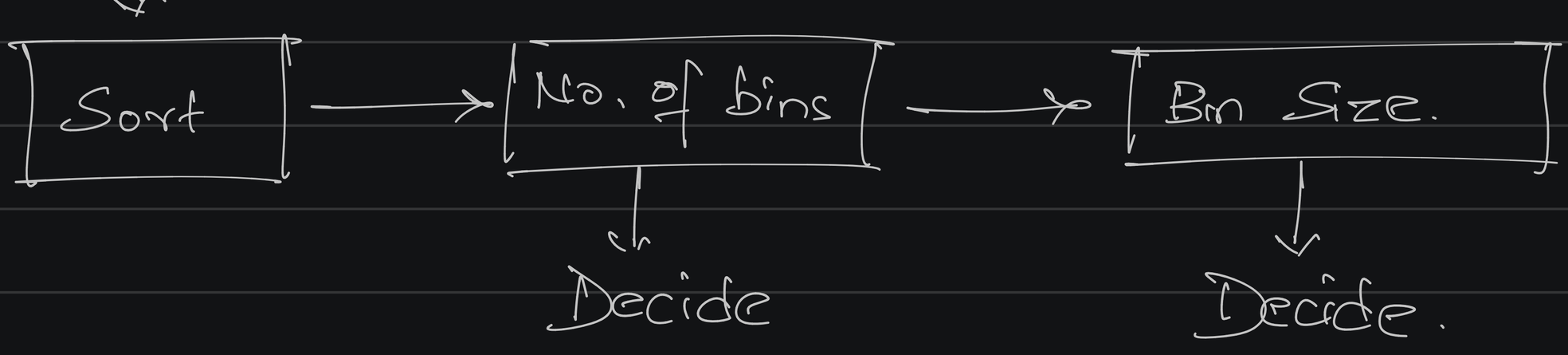
Day 2.

Agenda.

1. Histogram.
2. Measure of central tendency.
3. Measure of dispersion.
4. Percentiles & Quartiles.
5. 5 Number Summary (Box Plot).

1. Histogram. —

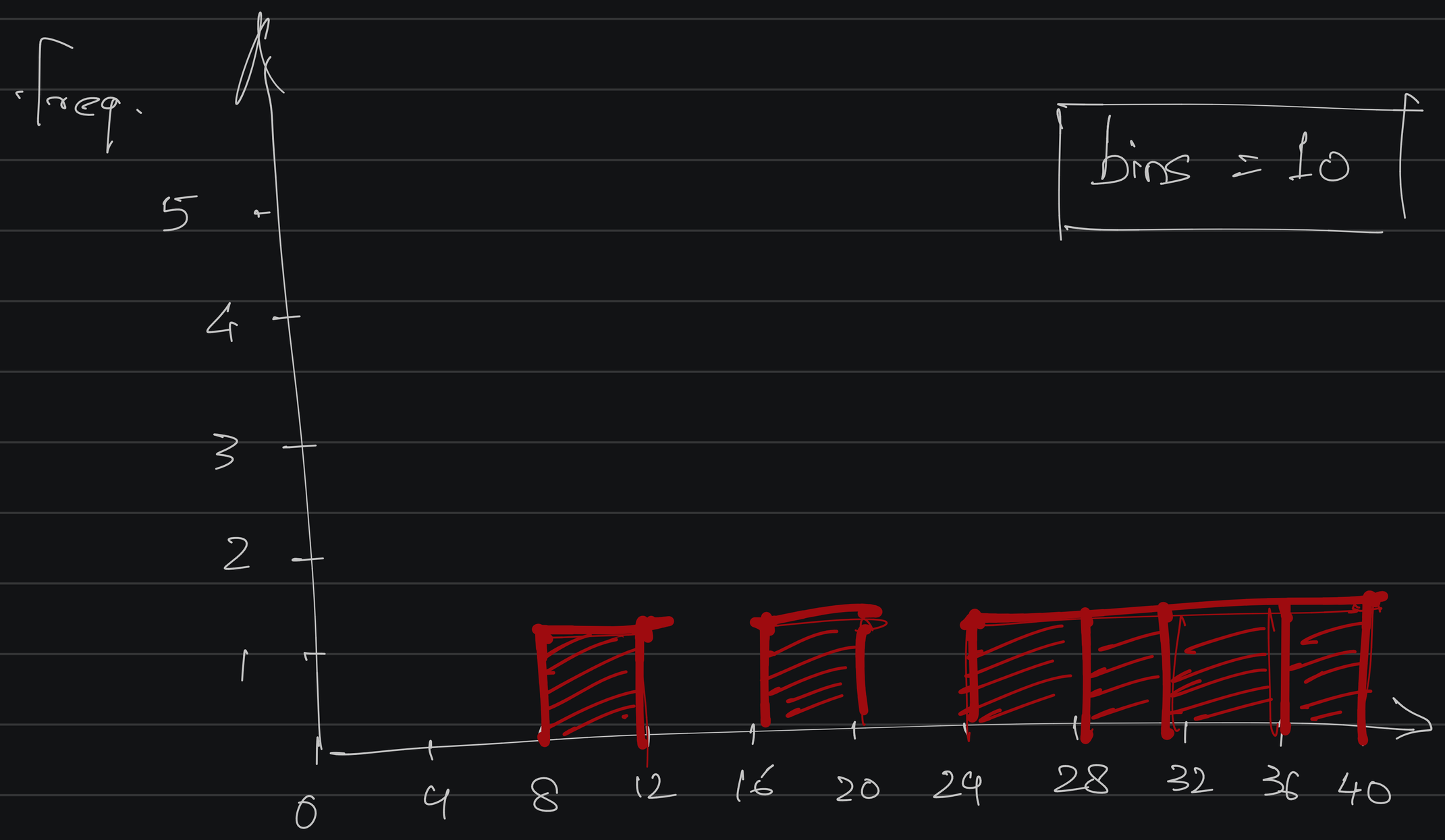
Ages. = { 10, 12, 14, 18, 24, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }



e.g. min = 10. bins = 10. binsize = $\frac{40}{10}$.

max = 40

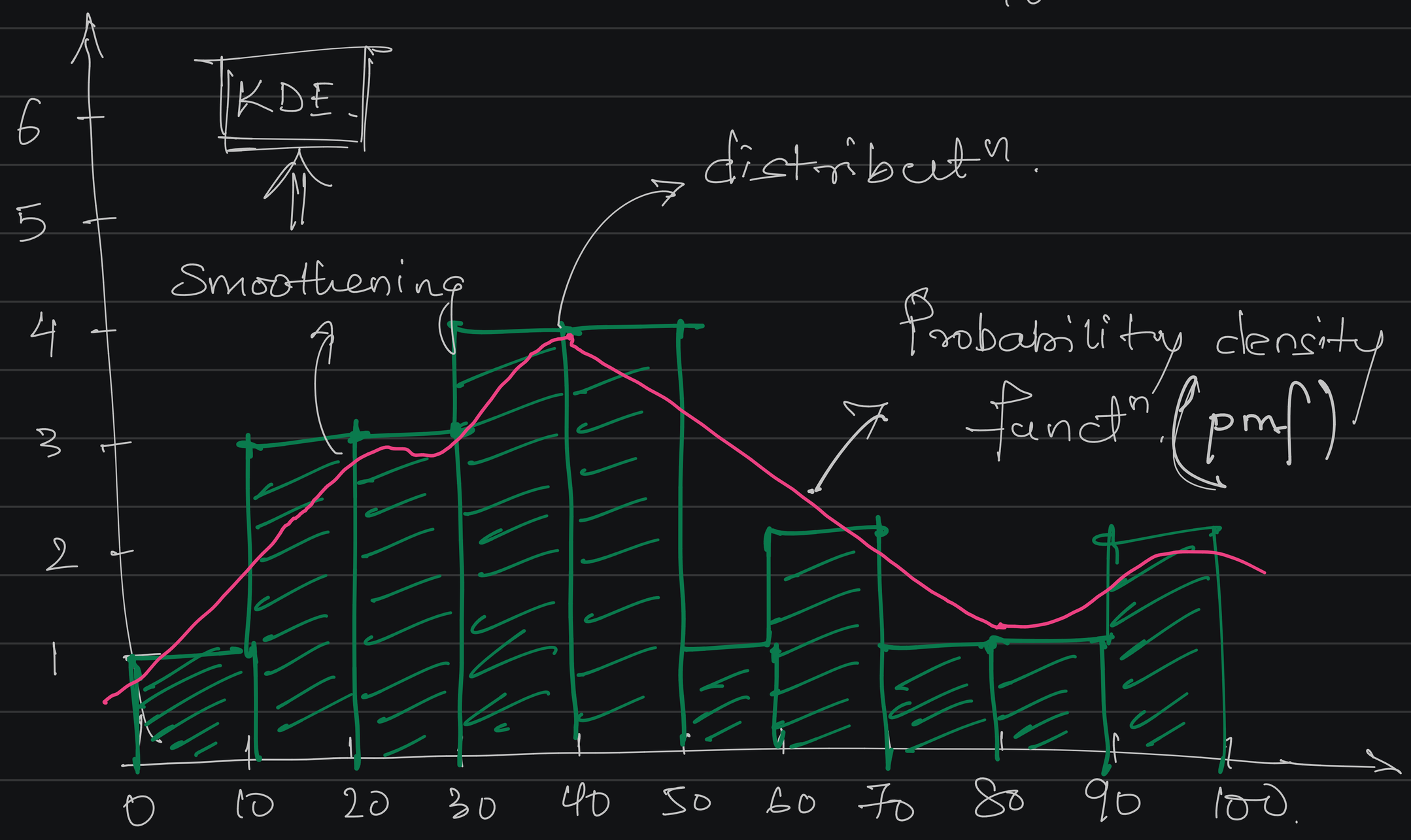
[10, 20, 25, 30, 35, 40]



bins = 10

For ages data, histogram :

Bins = 10 , Bin size = $\frac{100.}{10}$ = 10.



* Weigth. = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}.

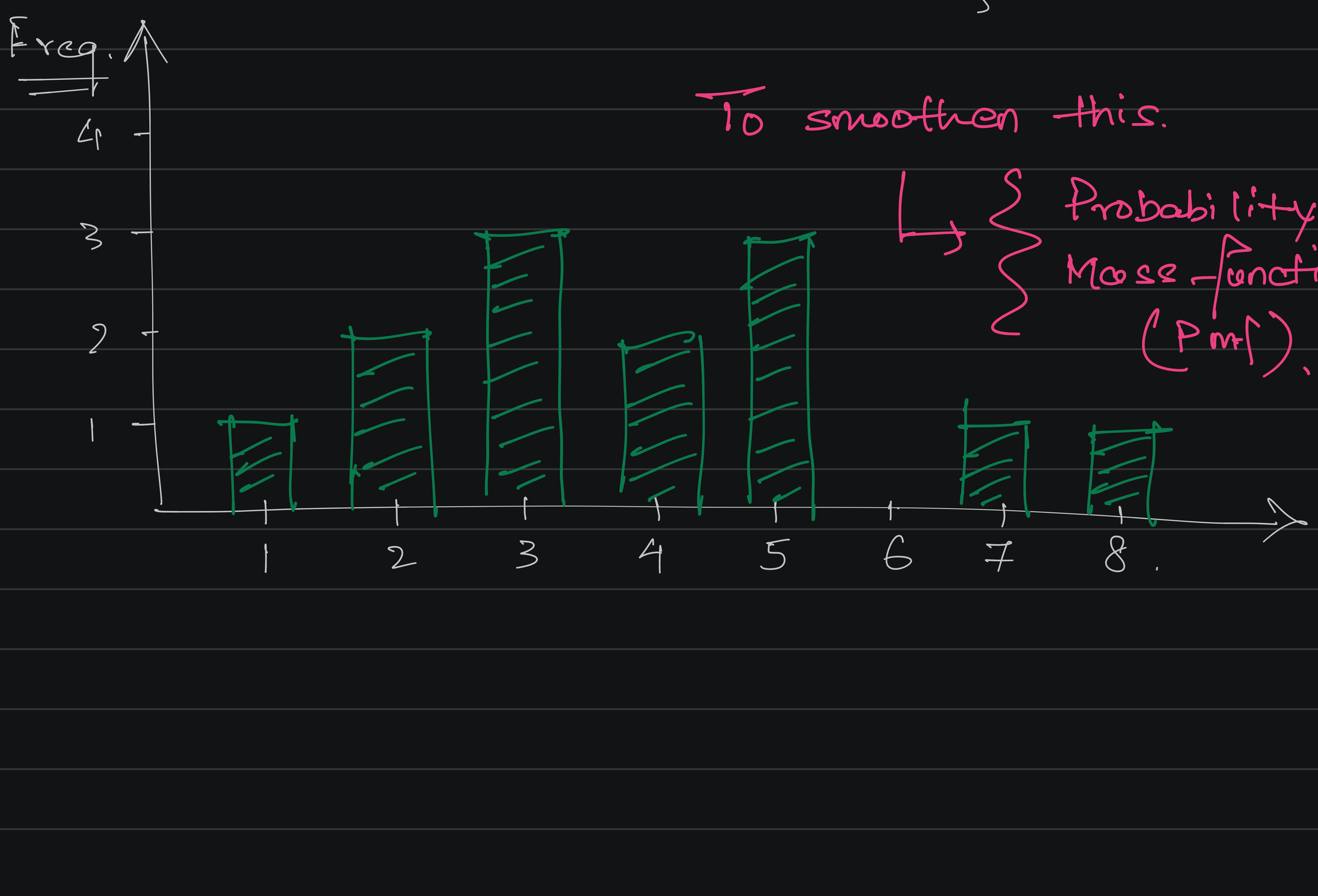
Bins = 10.

Bin Size = $\frac{95 - 30}{10} = \frac{65}{10} = 6.5$.

↳ Continuous variables.

* for discrete continuous variables.

No. of bank accounts. = {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}



To smoothen this.

↳ Probability Mass function (Pmf).

2. Measure of central tendency

i) Mean.
ii) Median.
iii) Mode

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position.

i) Mean $\Rightarrow X = \{1, 2, 3, 4, 5\}$

$$\text{Mean/Avg} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Population (N)

$$\text{Mean} \cdot \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample (n)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Population age $= \{24, 23, 2, 1, 28, 27\}$.

$N = 6$, $\mu = \frac{24+23+2+1,28,27}{6}$.

$\mu = 17.5$.

Sample = {24, 2, 1, 27}.

Sample mean $(\bar{x}) = \frac{24+2+1+27}{4}$.

$\bar{x} = 13.5$

Here, $N > n$.

For means, $\mu \geq \bar{x}$

or $\bar{x} \geq \mu$.

Practical Applicatn.

Age	Salary	Family Size.
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
NAN	NAN	NAN
—	—	—

MEAN → loss of info.

Replace NAN with mean

Median.

Age	Salary
24	45
28	50
29	NAN.
NAN	60
31	75.
36	80
NAN.	NAN.

Mean Age = 29.6.

Can replace NAN.

Mean salary = 62

Can replace. NAN.

Additional data = 80, 200.

outliers.

New mean. = 38.

New mean = 85.

Causing deviatn.

This is why Median is used.

Example.

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

$\bar{x} = 3.$

$\{1, 2, 3, 4, 5, 100\}$

$\bar{x} = 19.16.$

Outlier.

deviation due to outlier.

Steps to find out median

1. Sort the numbers

2. Find the central number.

(i) No. of elements are even:

We find avg. of central elements.

(ii) No. of elements are odd:

We find the central element.

Solved.

Central elements.

$\{1, 2, 3, 4, \boxed{5, 6}, 7, 8, 9, 10\}$ → Even.

no. of elements.

$\text{Median} = \frac{5+6}{2} = 5.5.$

Sorted.

{ 0, 1, 2, 3, 4, [5], 6, 7, 8, 100, 120 }

Odd no. of elements

Central element

Median.

Mean = 25.6.

(*) When there are outliers → Use median.

(!!!) Mode — Most frequent occurring element.

{ 1, 2, 2, [3, 3, 3], 4, 5 }

→ Mode = 3.

{ 1, [2, 2, 3, 3], 4, 5 }

→ Mode = 2, 3

Practical Importance

eg. Dataset :-

Types of flower.

Lily

Sunflower

Rose

Rose.

3. Measure of dispersion

i) Variance (σ^2)

ii) Standard deviation (σ)

i) Variance ⟶ Shows the spread of data.

Population variance (σ^2)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample variance (s^2)

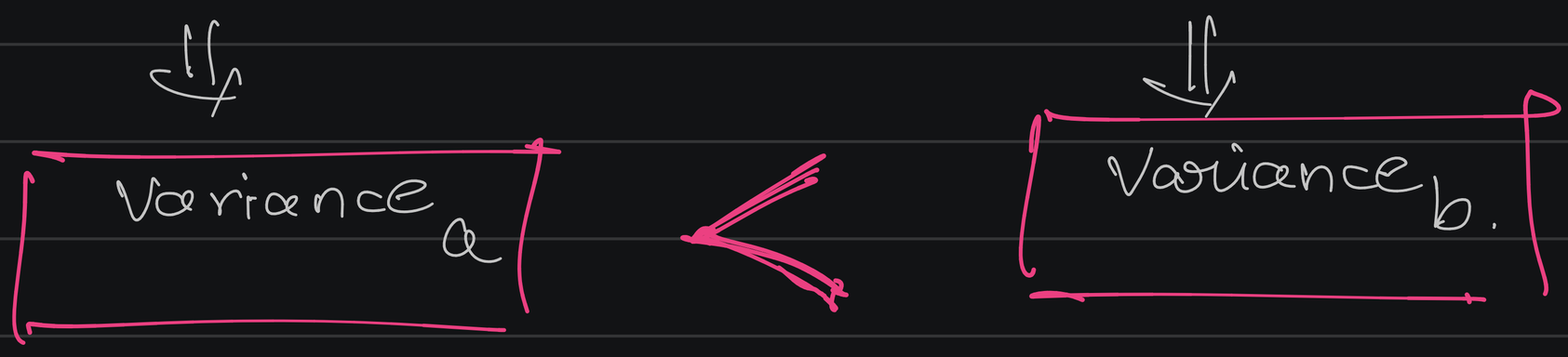
$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

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

$\mu = 3.$

$(x_i - \mu)$ = distance from the mean.

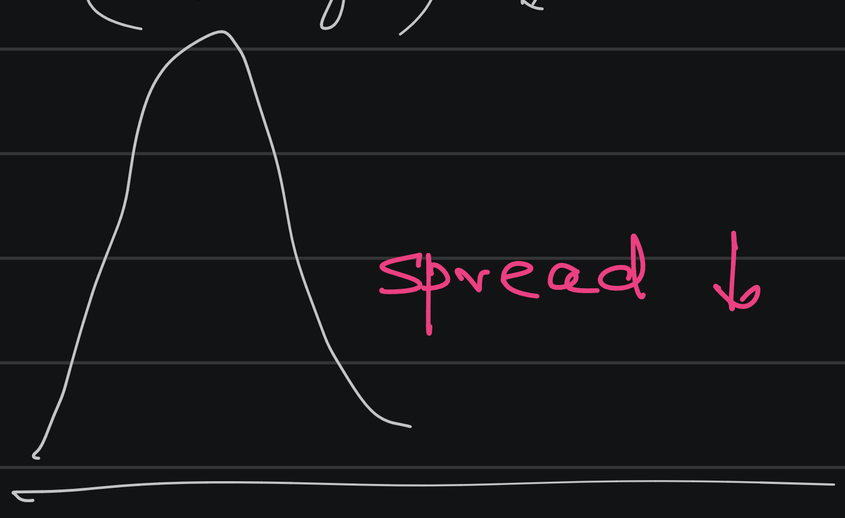
eg $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ $\{1, 2, 3, 4, 50, 60, 70, 100\}$

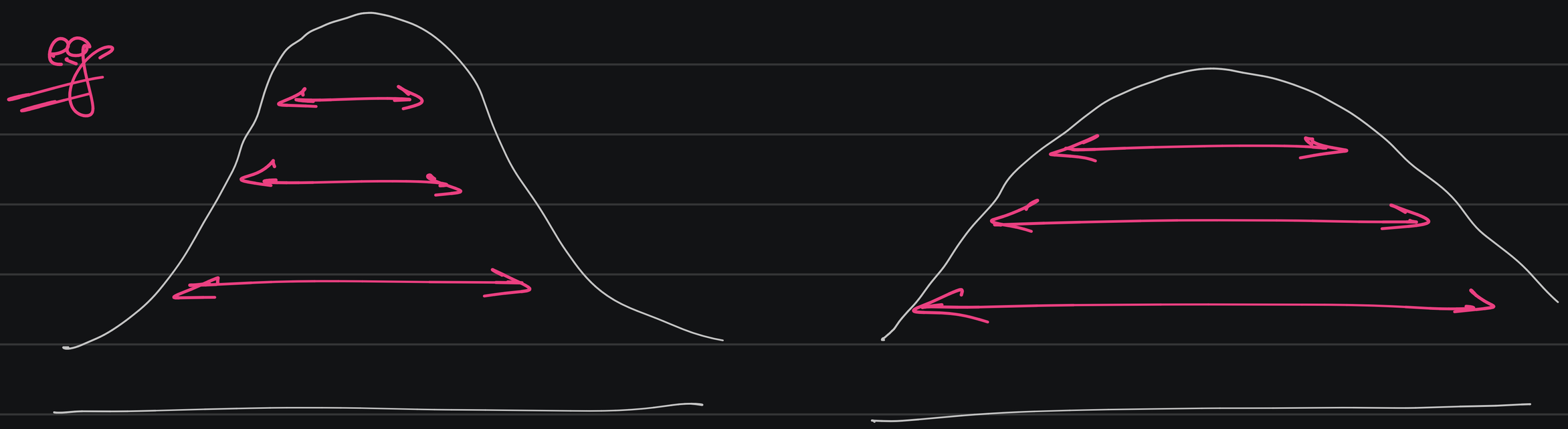


Mean ↓ Mean ↑

$\mu = 5.5.$ $\mu = 36.25$

$(x_i - \mu)^2$ ↓ $(x_i - \mu)^2$ ↑





spread (σ^2) < spread (σ^2)

Sample variance.

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

In real scenario,

$\sigma^2 > s^2 \longrightarrow$ from experiments.

to compensate for the difference,

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

'n' is adjusted to 'n - 1'.

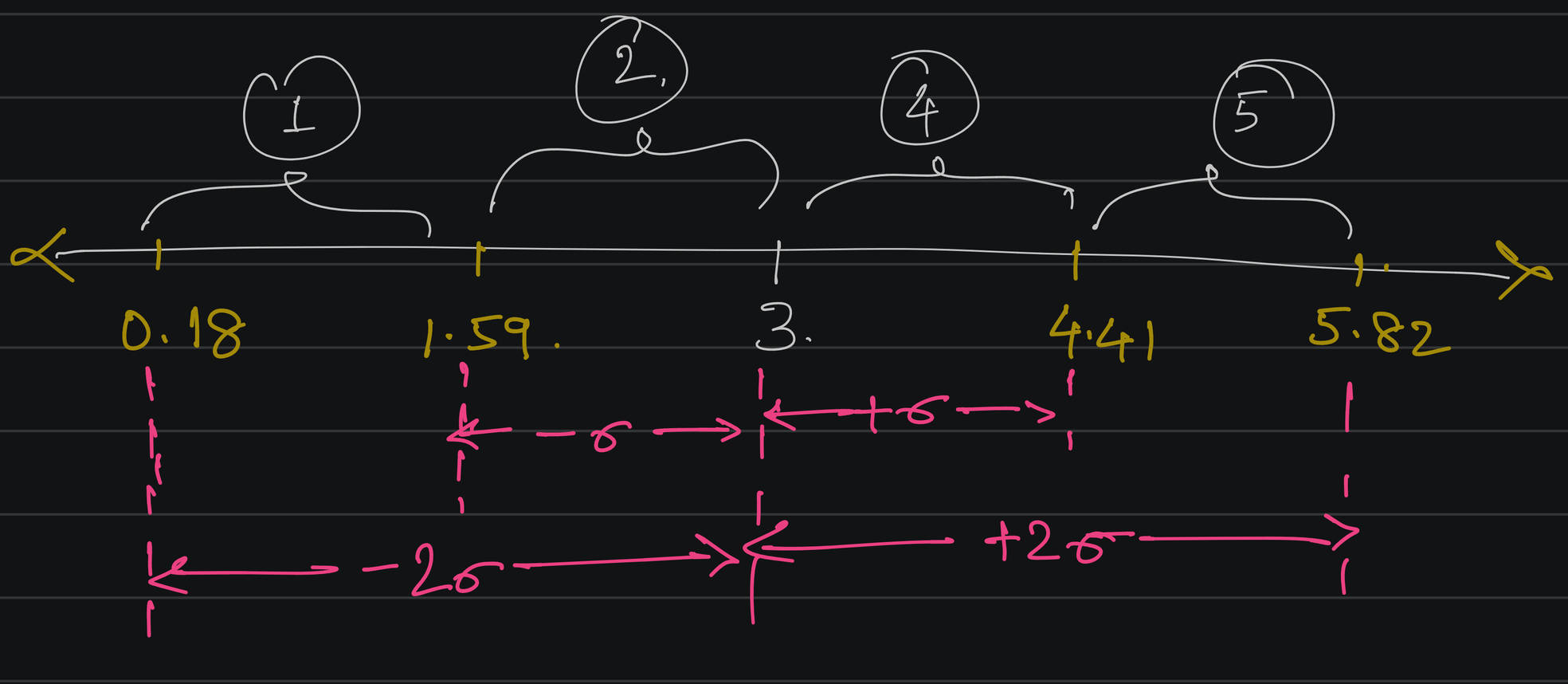
Standard deviation ($\sqrt{\sigma^2}$)

eg: $\{1, 2, 3, 4, 5\}$

$\sigma^2 = 2$

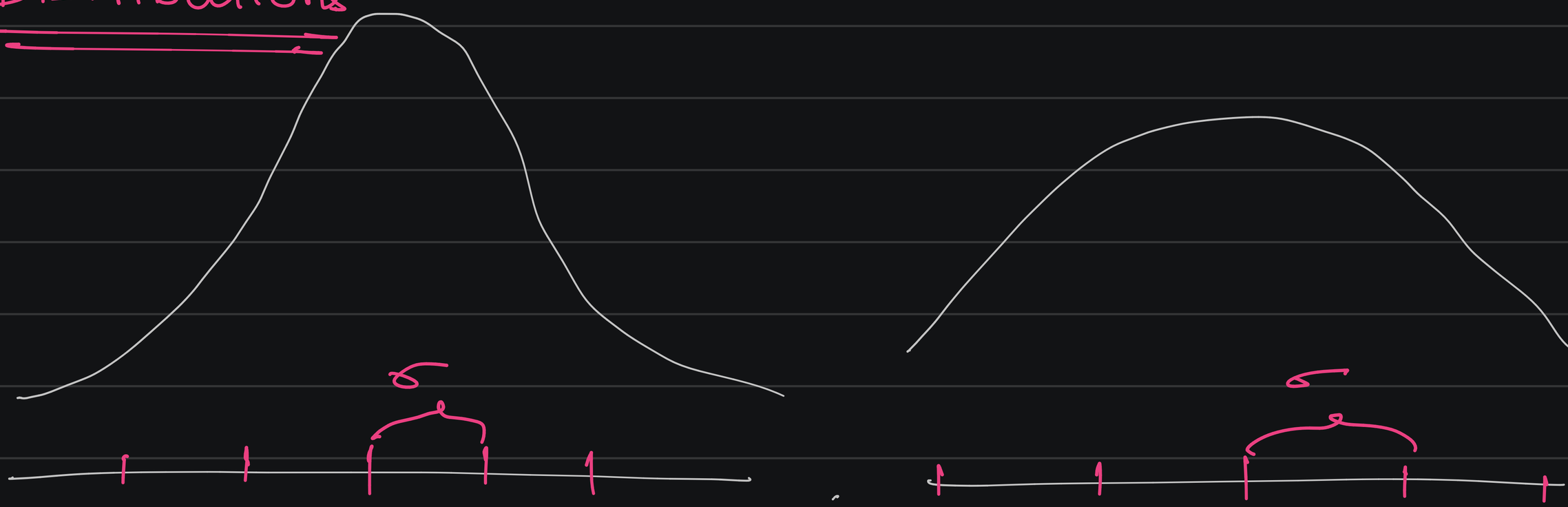
$\mu = 3.$

$\sigma = \sqrt{2} = 1.41.$



* Shows how many standard deviation away the element falls from the mean.

Distributions



Variance < Variance.

SD < SD.

4 Percentiles & Quartiles.

Percentage = {1, 2, 3, 4, 5, 6, 7, 8}.

Percentage (Even nos.) = $\frac{\text{No. of even nos.}}{\text{Total no. of nos.}}$

$= \frac{4}{8} = 0.5 = 50\%$

Percentile → Percentile is a value below which a certain percentage of observations lie.

eg 99 percentile = the person has got better marks than 99% of applicants.

5th index

eg. Dataset = { 2, 2, 3, 4, 5, 5, 5, 6, 7, 8,
8, 8, 8, 8, 9, 9, 10, 11, 11, 12 } Sorted.

Q What is the percentile rank of 10 ?

Percentile rank of x = $\frac{\text{No. of value below } x}{n}$ ylow

$= \frac{16}{20} =$ 80 percentile.

Percentile (8) $= \frac{9}{20} =$ 45 percentile

Percentile (6) $= \frac{7}{20} =$ 35 percentile.

(*) What is the value that exists at 25 percentile

$$\text{Value} = \frac{\text{Percentile}}{100} * n+1.$$

$$\text{Value} = \frac{25.}{100} * 21$$

$$= 5^{th} \text{ index}.$$

$$\text{Value.} = 5.$$

$$\text{Value} = \frac{95.}{100} * 21.$$

$$= 19.5 \longrightarrow 20^{th} \text{ element.}$$

$$\underline{\text{Value}} = 12.$$

5) 5 number Summary.

i) Minimum.

ii) first Quartile. (25 percentile) Q_1

iii) Median. (Q_2)

iv) Third Quartile (75 percentile) Q_3

v. Maximum.

Remove outliers. ←

↓

using Box Plot.

eg Dataset = 1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27

[Lower Fence ←——→ Higher Fence.]

Lower Fence = $Q_1 - 1.5(IQR)$

Higher Fence = $Q_3 + 1.5(IQR)$

$IQR = Q_3 - Q_1$

Inter-quartile range.

$$Q_1 = \frac{25}{100} * (n+1) = \frac{25}{100} \times 21$$

$= 5.25 \longrightarrow$ av. of 5^{th} & 6^{th} index.

$Q_1 = 3.$

$$Q_3 = \frac{75}{100} \times 21 = 15.75. \longrightarrow \text{av. of } 15^{th} \text{ \& } 16^{th} \text{ index de.}$$

$Q_3 = 7.5.$

lower Fence $= 3 - 1.5(4.5) = -3.65.$

Higher Fence $= 14.25.$

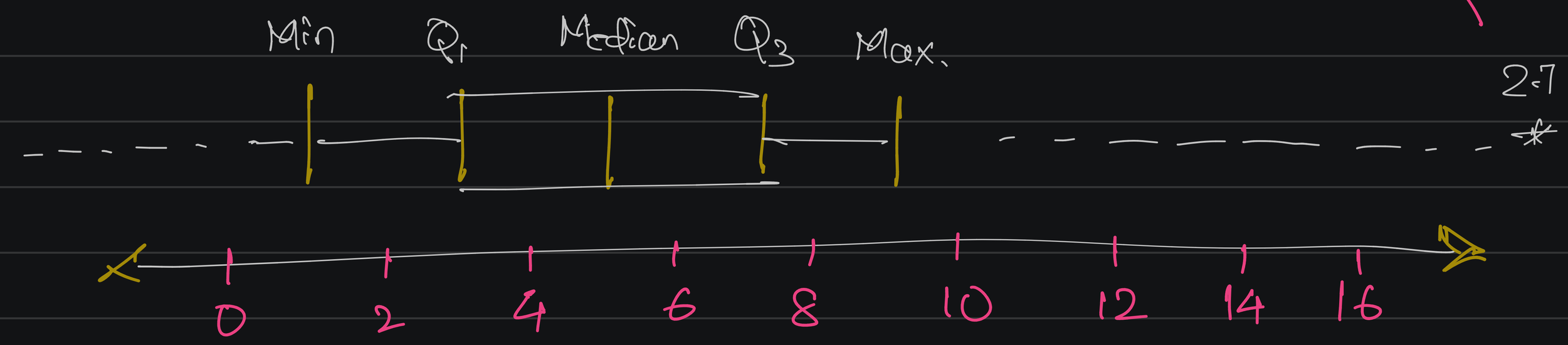
$-3.65 \longleftarrow$ values $\longrightarrow 14.25.$

So, 27 is the outlier.

5 number Summary :

1. Minimum = 1
2. Q_1 = 3.
3. Median = 5.
4. Maximum = 7.5.
5. Maximum = 9.

Outlier.



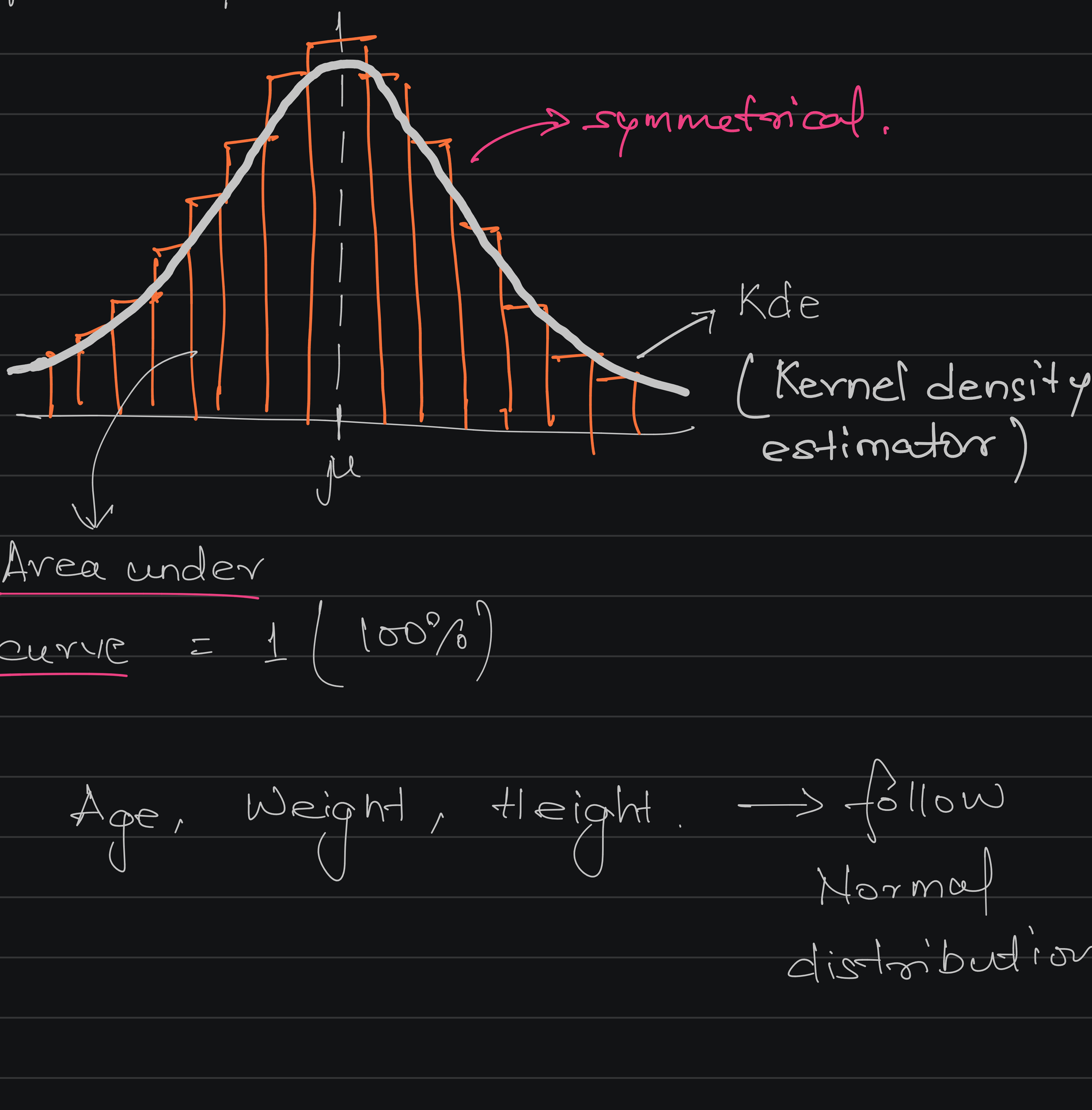
BOX PLOT

Day 3

Agenda.

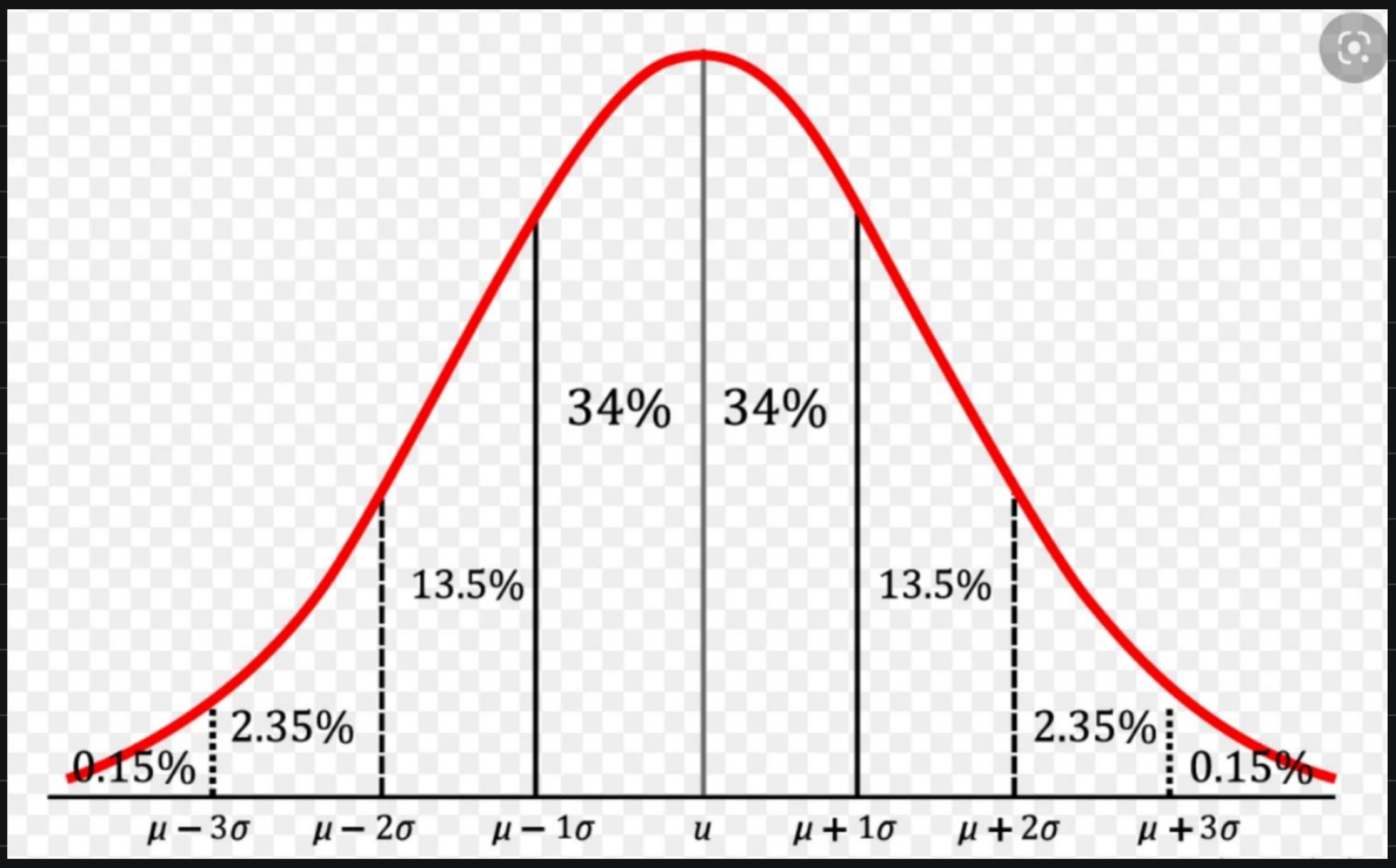
1. Normal distribution.
2. Standard Normal distribution.
3. Z - Score
4. Standardization & Normalization

1. Gaussian / Normal distribution.



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

(A) 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)

68 - 95 - 99.7% rule

→ Empirical Rule.

Q-Q plot ⟶ Can show if a distribution is Gaussian or not.

2. Standard Normal Distribution.

if $\underline{X}$ belongs to Gaussian Distribution (μ, σ)

⇓ can be converted into. using Z Score

$\underline{Y} \approx$ Standard Normal distribution

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

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

$\mu = 3$

$\sigma = 1.41$

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

$n = 1$, applied on every element

$\Rightarrow$ Standard Error.

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

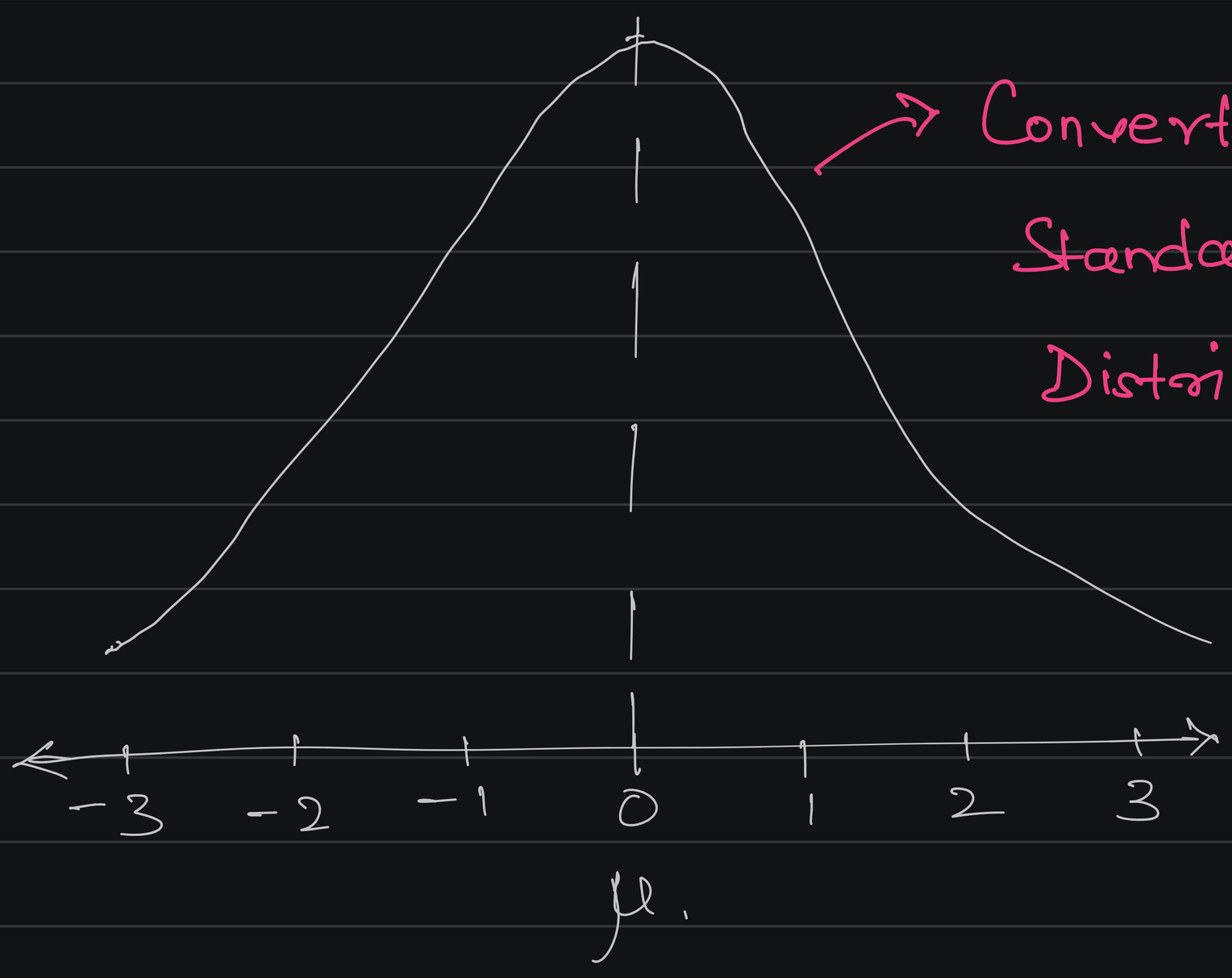
* $y = \{-1.414, -0.707, 0, 0.707, 1.414\}$

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

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

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

Distribution of y.



→ Converted into Standard Normal Distribution.

Why do we need to convert ?

3 features → Age (years), Weight (kg), Height (cm)

Age	Weight	Height
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
.		
28	83	180
29.	80.	175.

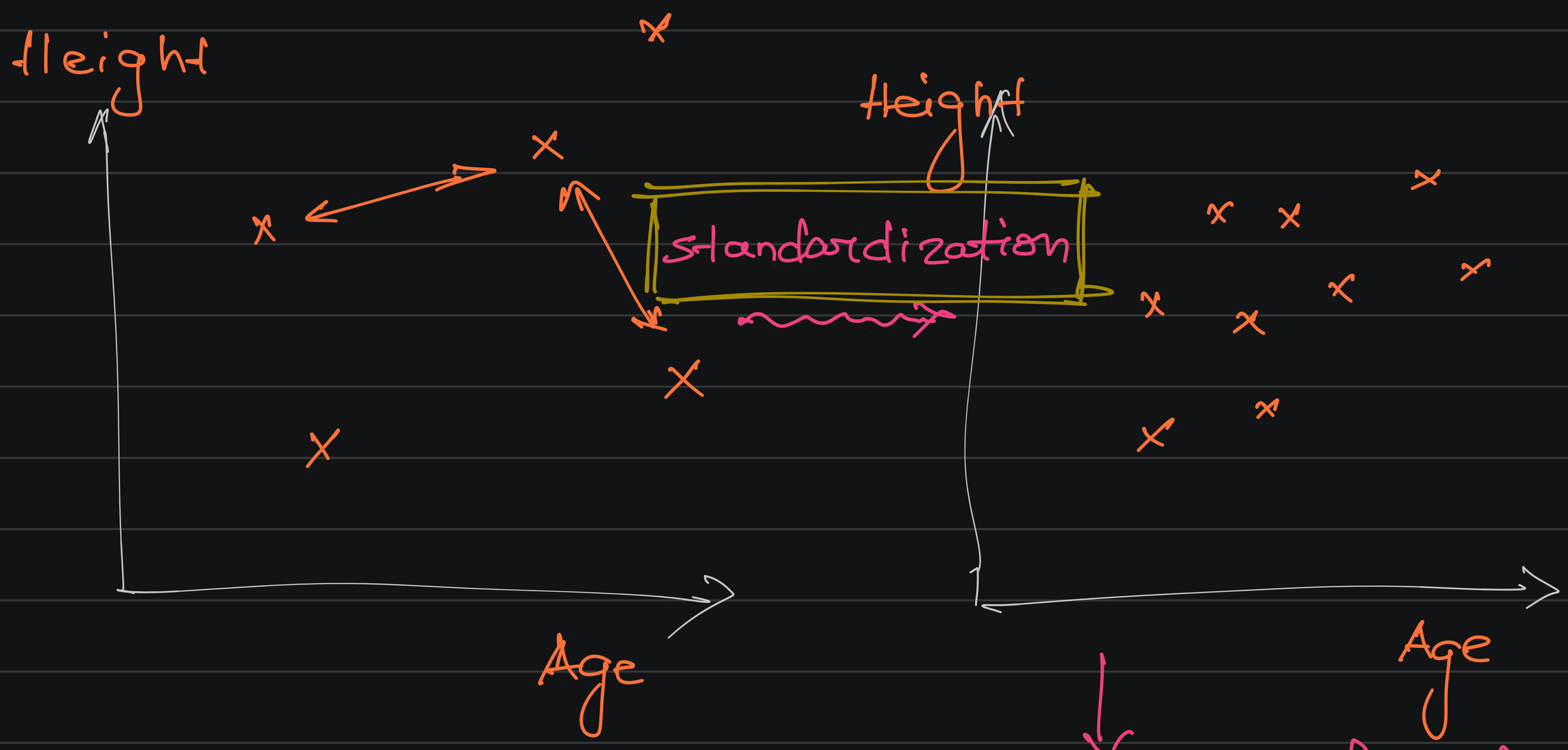
Units are different

↓

Value ranges differ a lot.

↓

Not favorable of ML models.



(*) 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 }

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

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 — transform values between 0 to 1.

Formula,

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

where,

$x_{max} = 1$

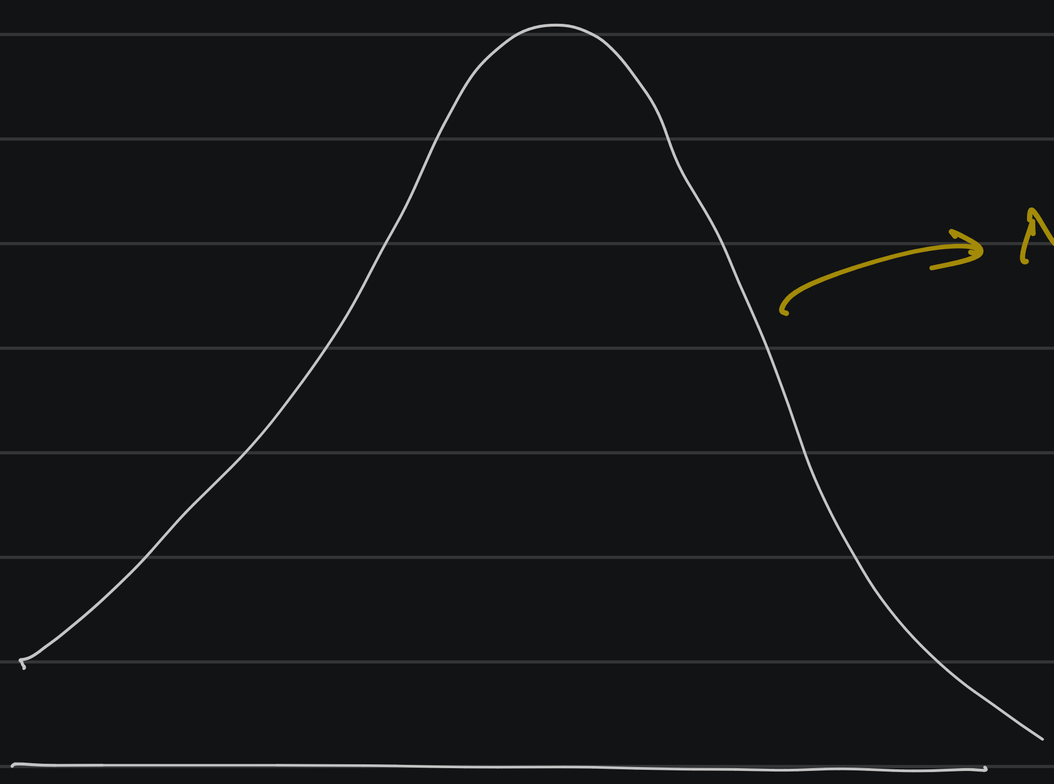
$x_{min} = 0$

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

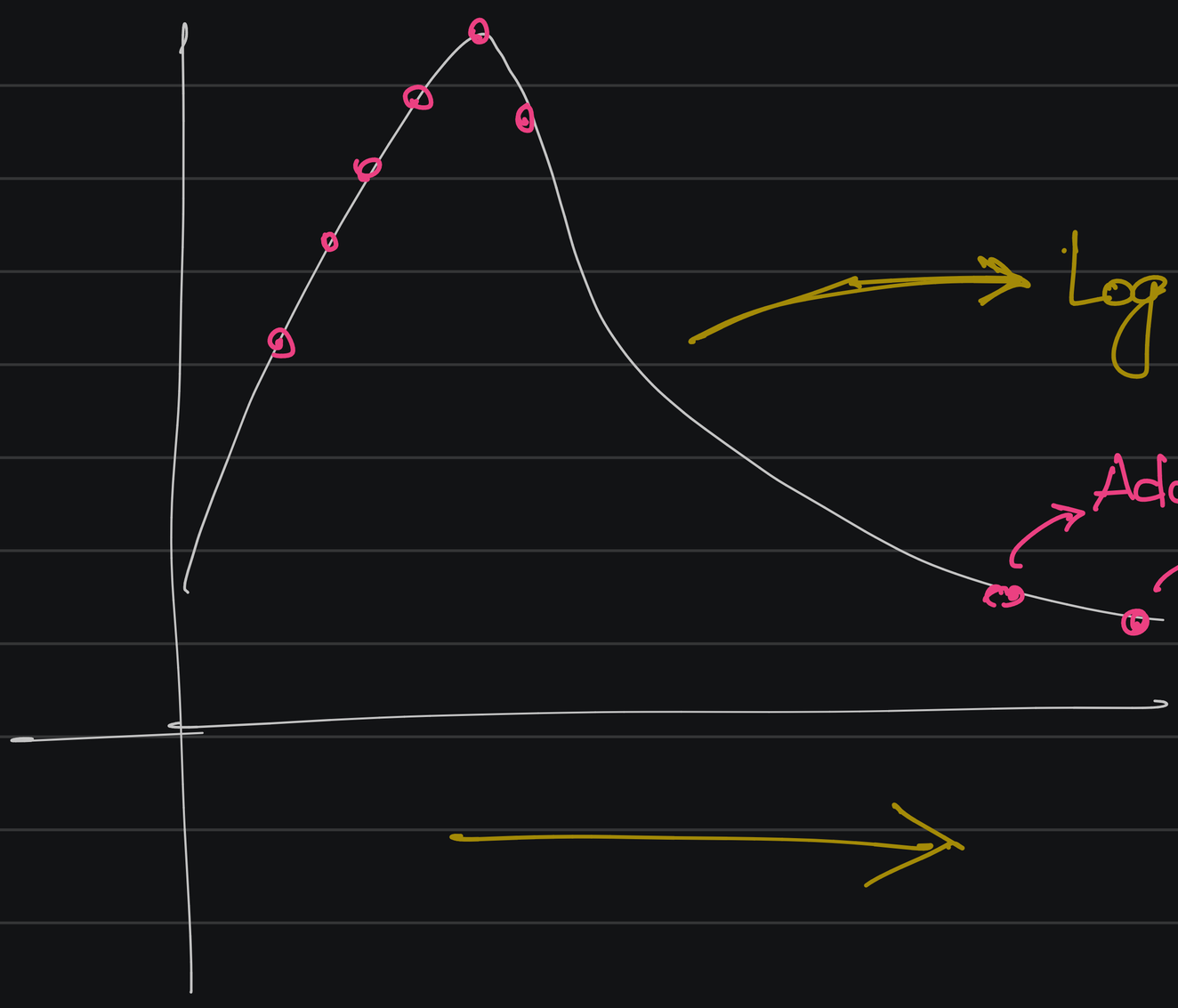
All values converted to 0 to 1.

Applied in Deep Learning

Log Normal Distribution.



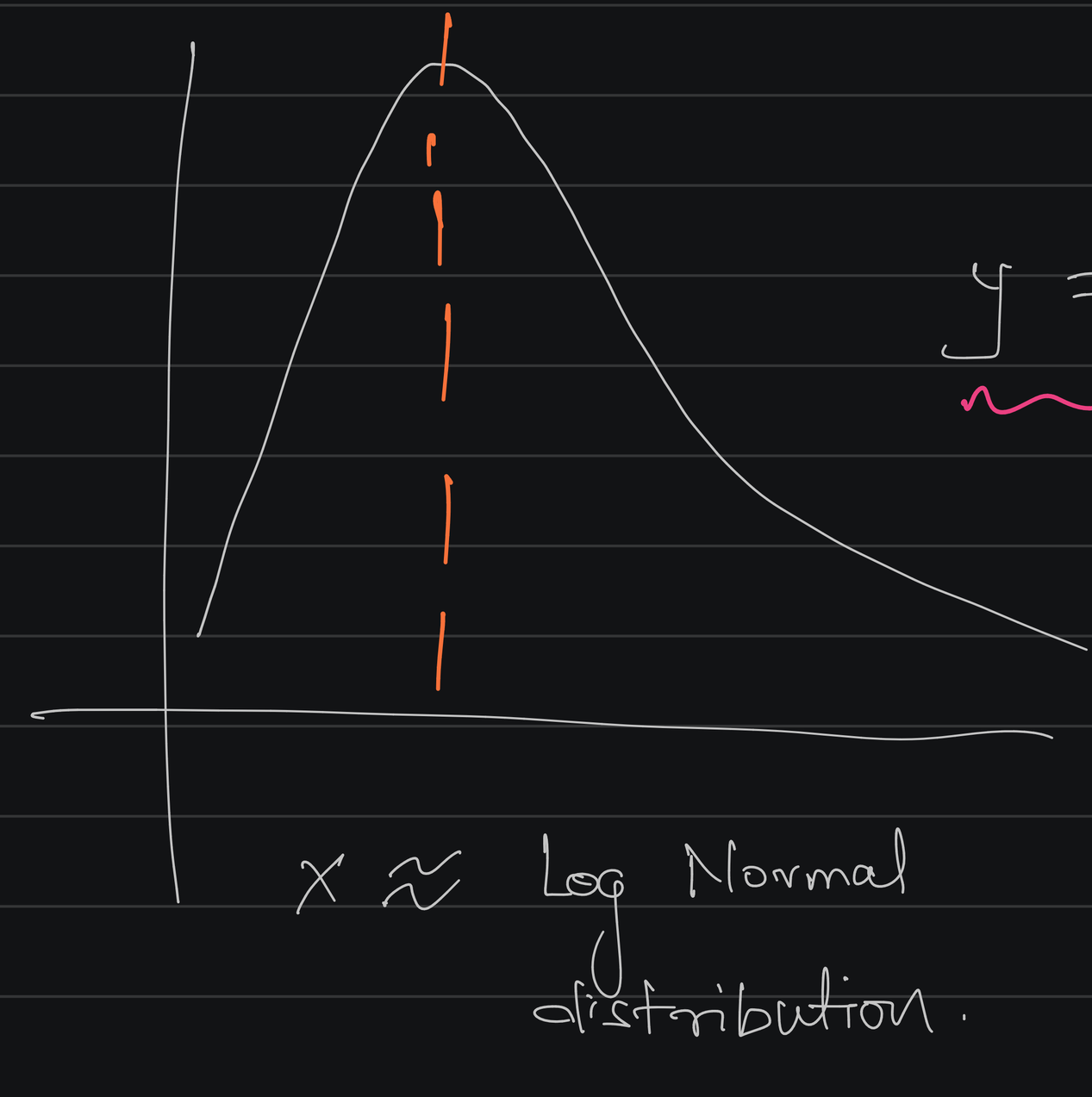
→ Normal distribution.

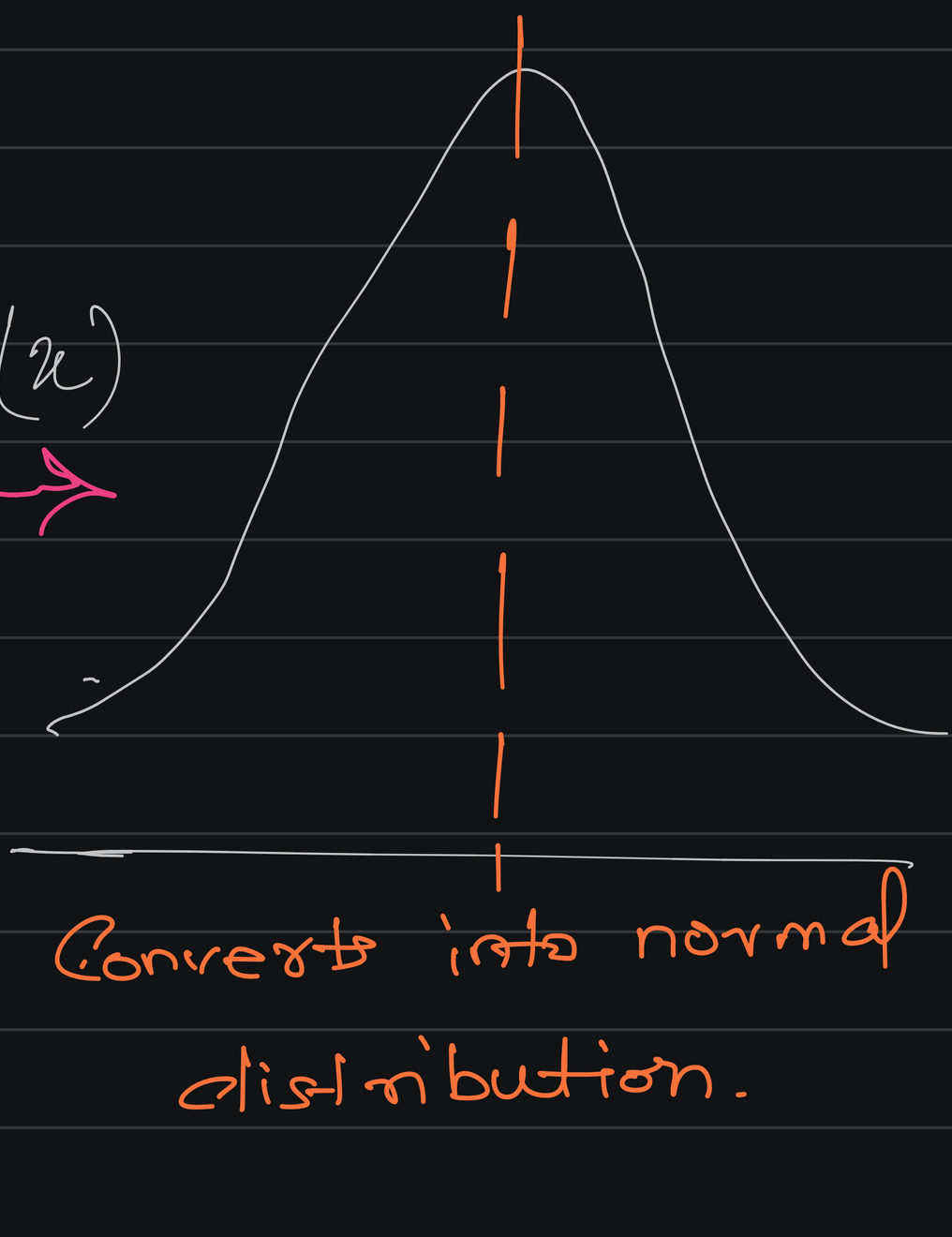


→ Log normal distribution.

eg - weath distribution.

Length of comments

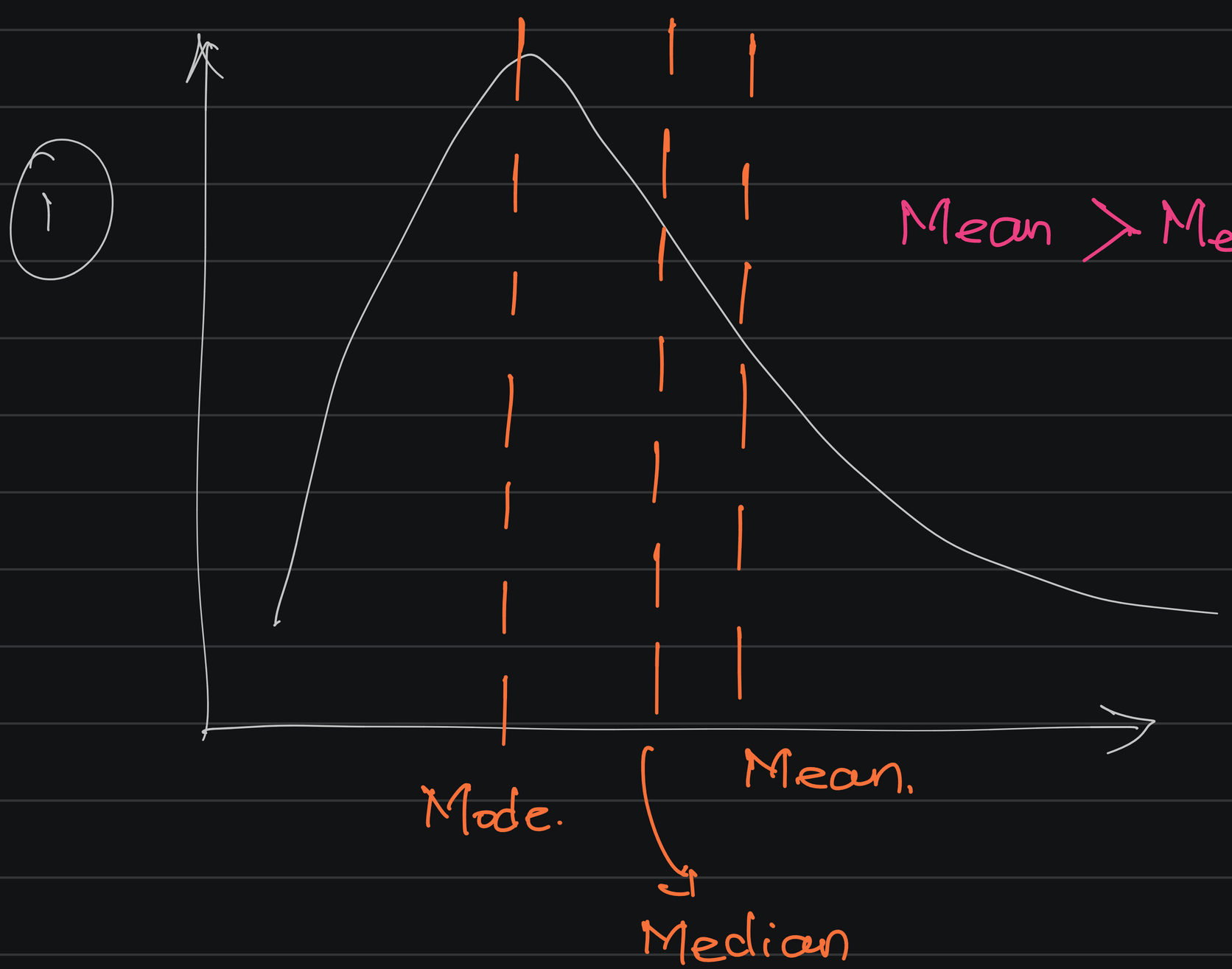




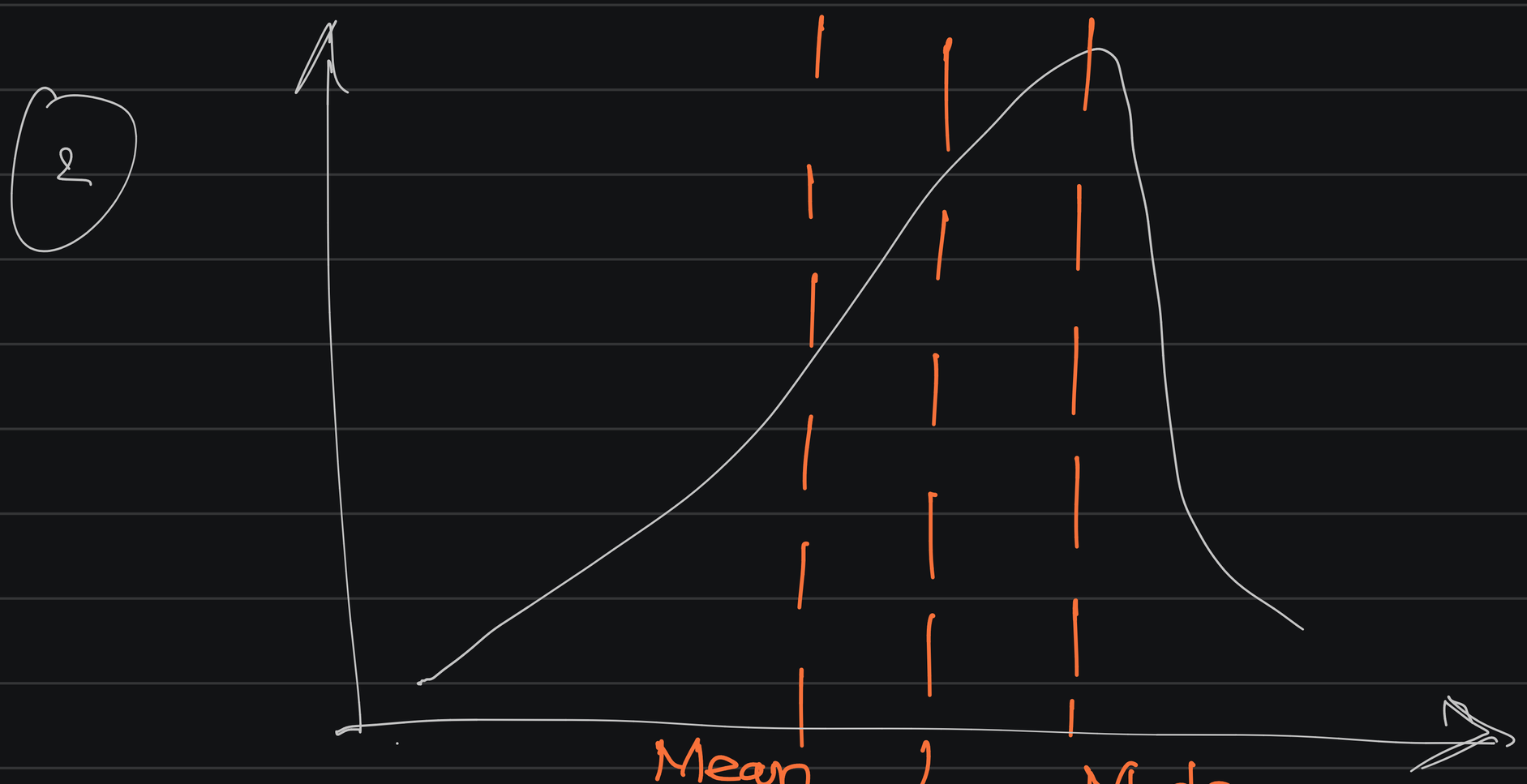
$X \approx$ Log Normal distribution.

Converts into normal distribution.

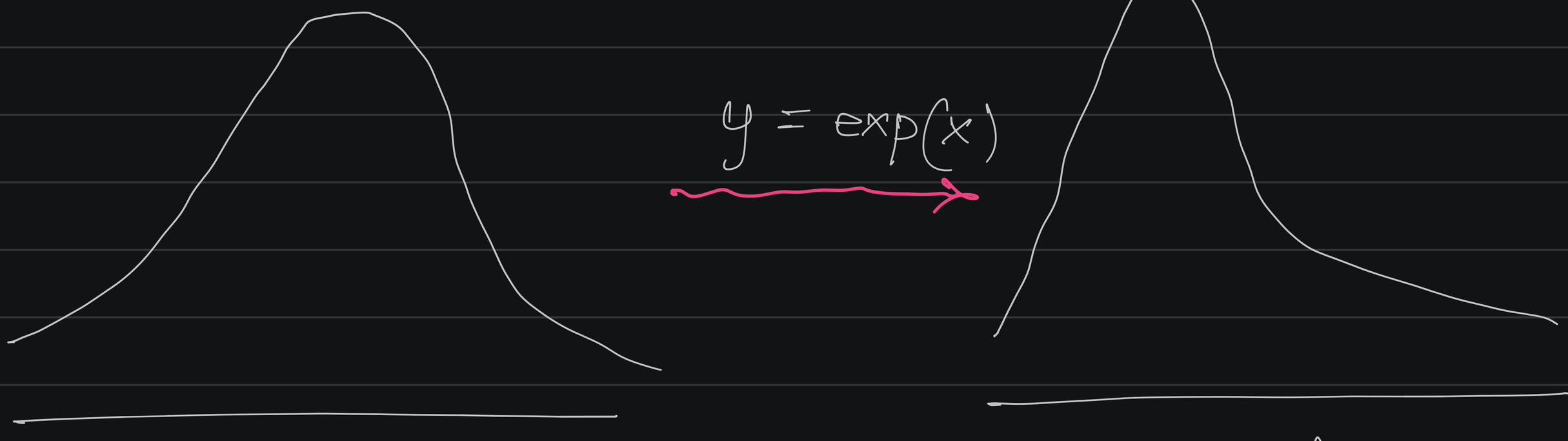
Q = What is the relationship between mean, median and mode for below distributions



Mean > Median > Mode.



Mean < Median < Mode.



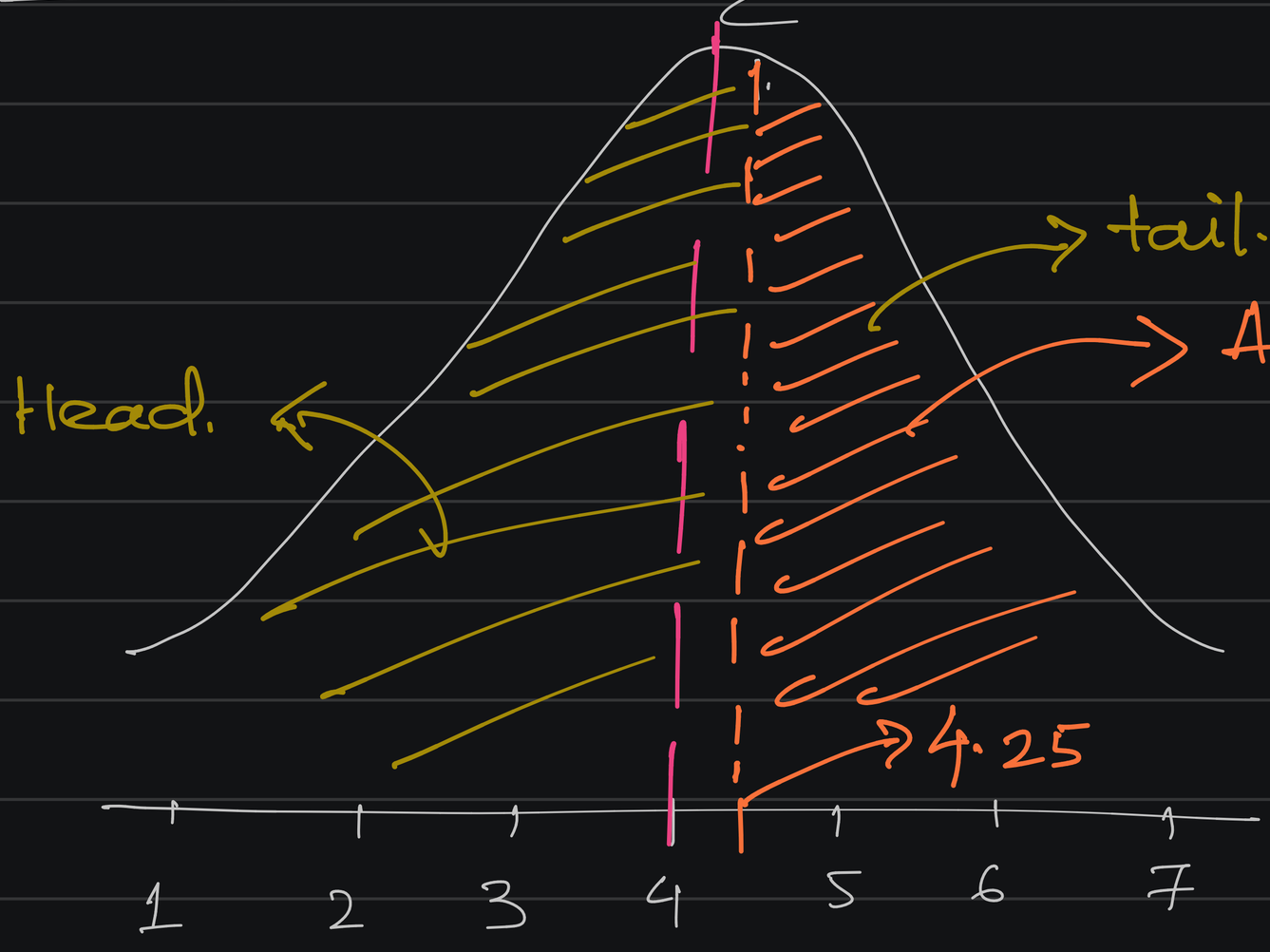
$X \approx$ Normal distribution.

(μ, σ)

Log normal distribution.

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

$\sigma = 1$



Question - What is the percentage of score that falls above 4.25 ?

Answer : (1.) Apply $Z\text{-score} = \frac{x_i - \mu}{\sigma}$

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

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

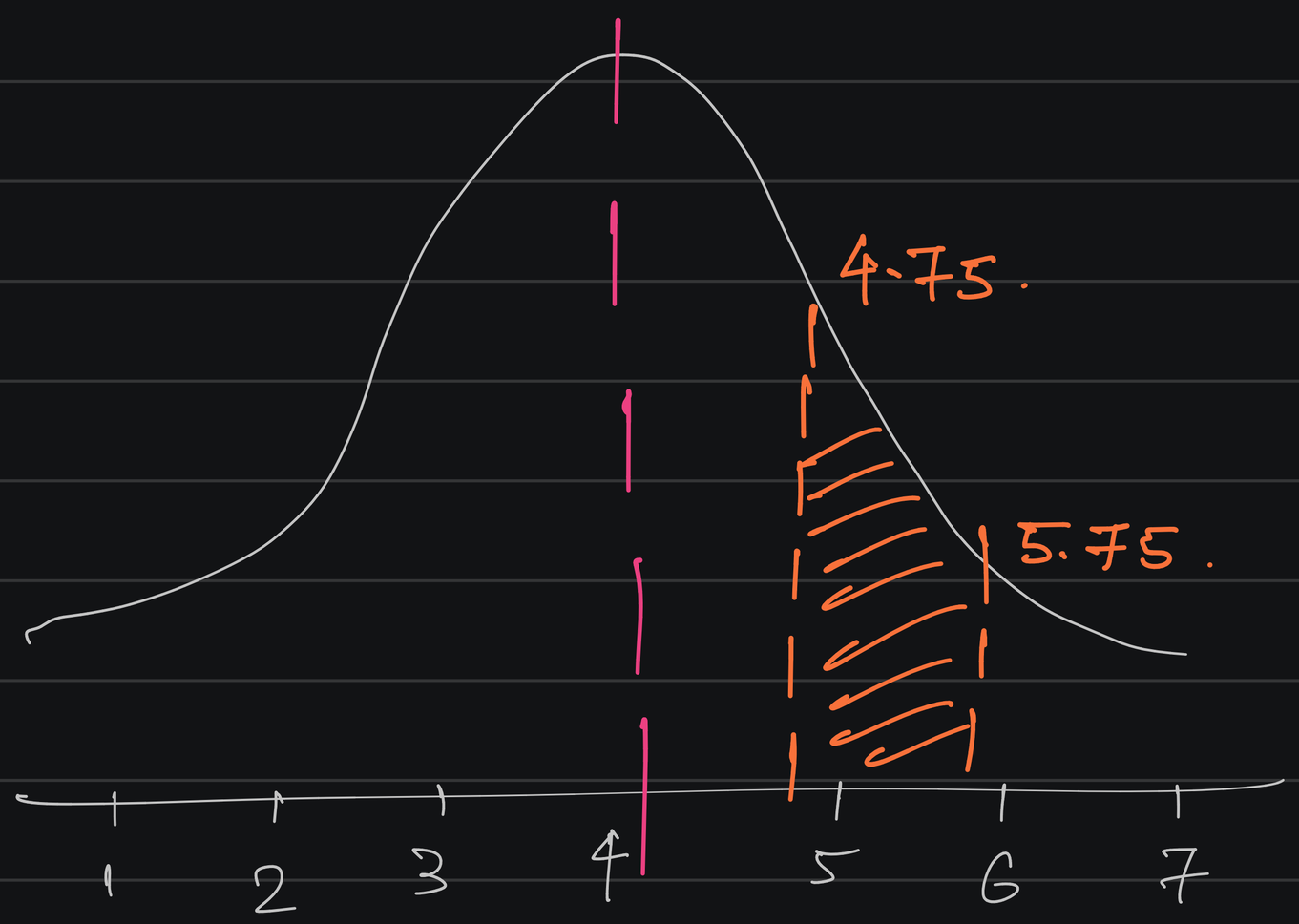
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.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% → 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).

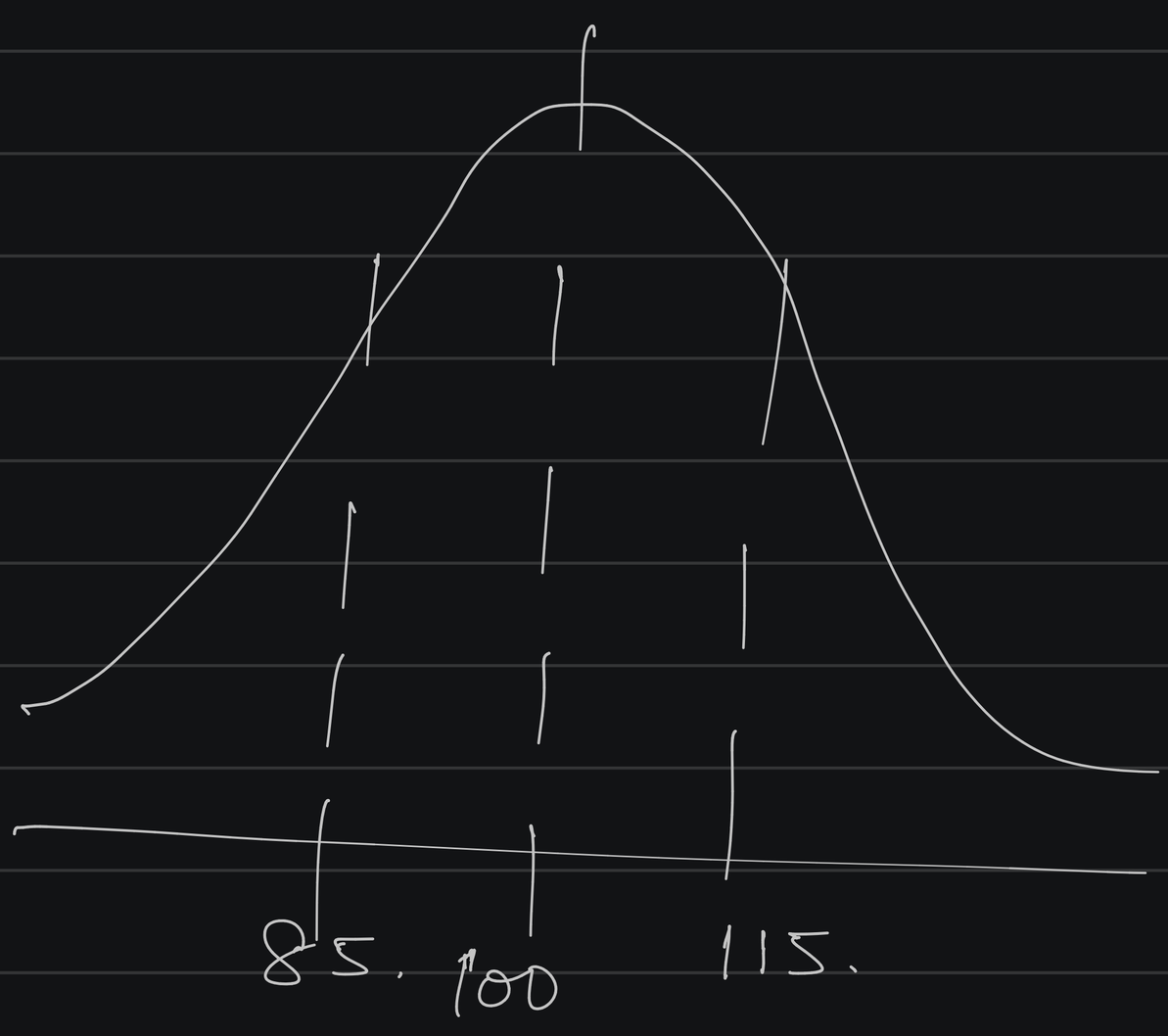
= 22.66 − 4.01 = 18.65%

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%



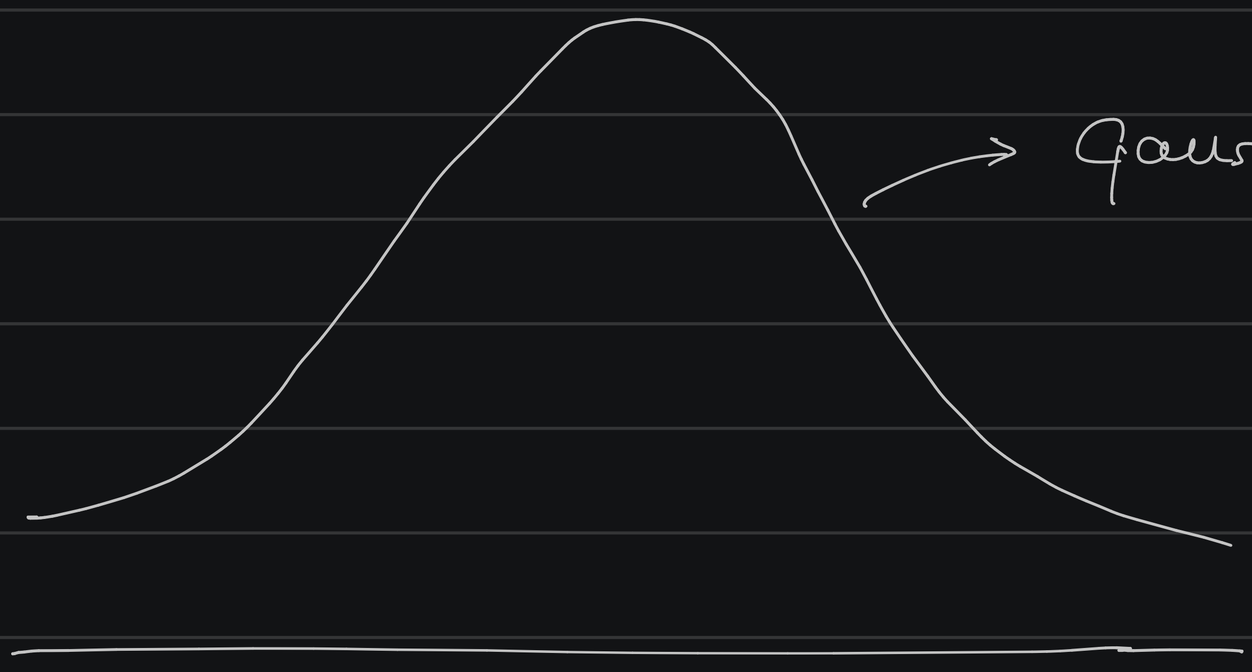
Day 4.

Agenda.

1. Central Limit theorem.
2. Probability.
3. Permutation & Combination.
4. Covariance, Pearson Correlation, Spearman Rank. Correlation.
5. Benouli's distribution.
6. Binomical distribution.
7. Power Law. (Pareto distribution)

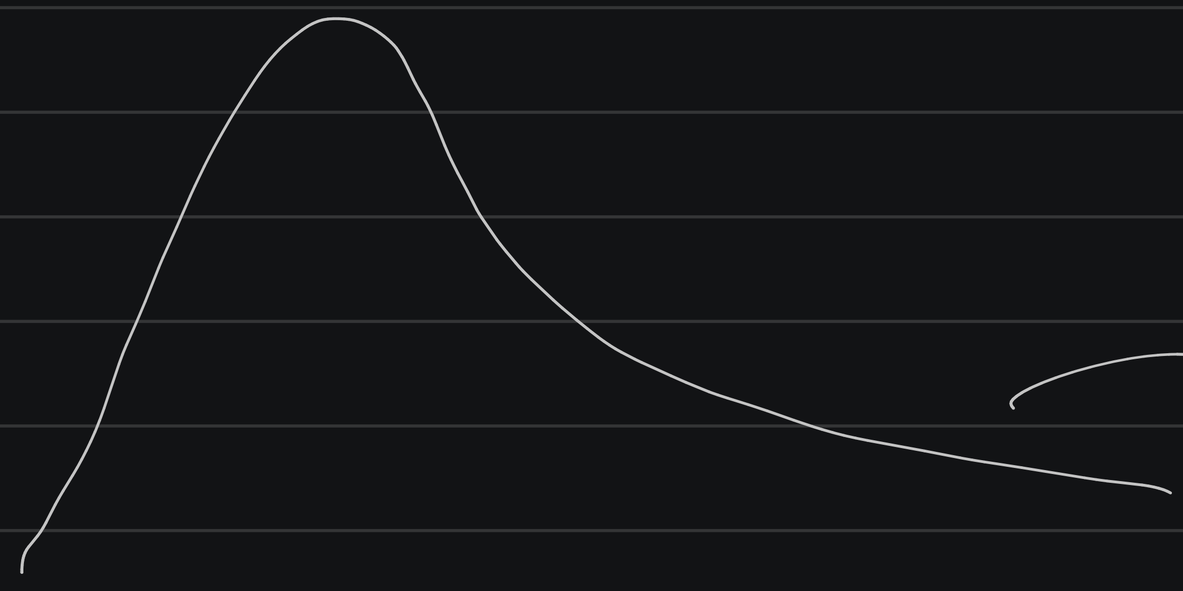
→ Not covered on Day 4.

1. Central Limit theorem.



→ Gaussian distribution.

Or.



→ log normal distribution

Requirement

→ Value ∝ accuracy

→ $n \geq 30$

(*) if we consider 'm' samples of 'n' elements from any distribution :-

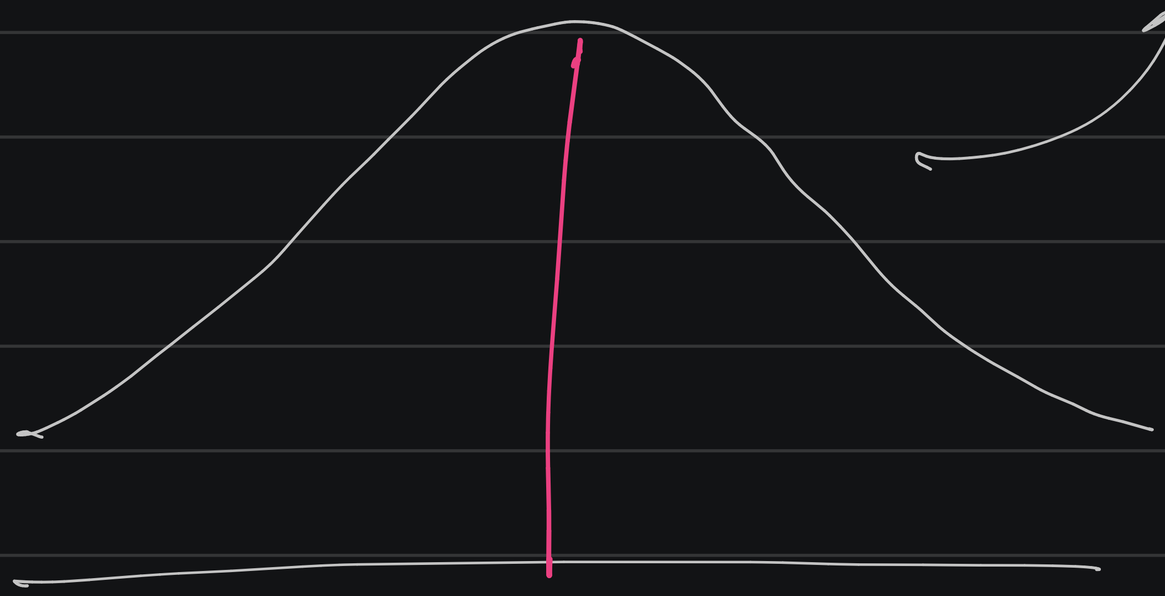
Sample$_1$ ⇒ { $x_1, x_2, x_3 \ldots\ldots x_n$ } → $\bar{x}_1$

Sample$_2$ ⇒ { $x_1, x_2, x_3 \ldots\ldots x_n$ } → $\bar{x}_2$

⋮

Sample$_m$ ⇒ { $x_1, x_2, x_3 \ldots\ldots x_n$ } → $\bar{x}_m$

The means will follow a gaussian distribution :-



As the distribution is now a gaussian distribution All the assumptions about gaussian distribution is valid now.

(2.) Probability — Probability is a measure of the likelihood of an event

eg: Tossing of coin (fair)

P(H) = 0.5 → Head.

P(T) = 0.5 → Tail.

SHOLAY →

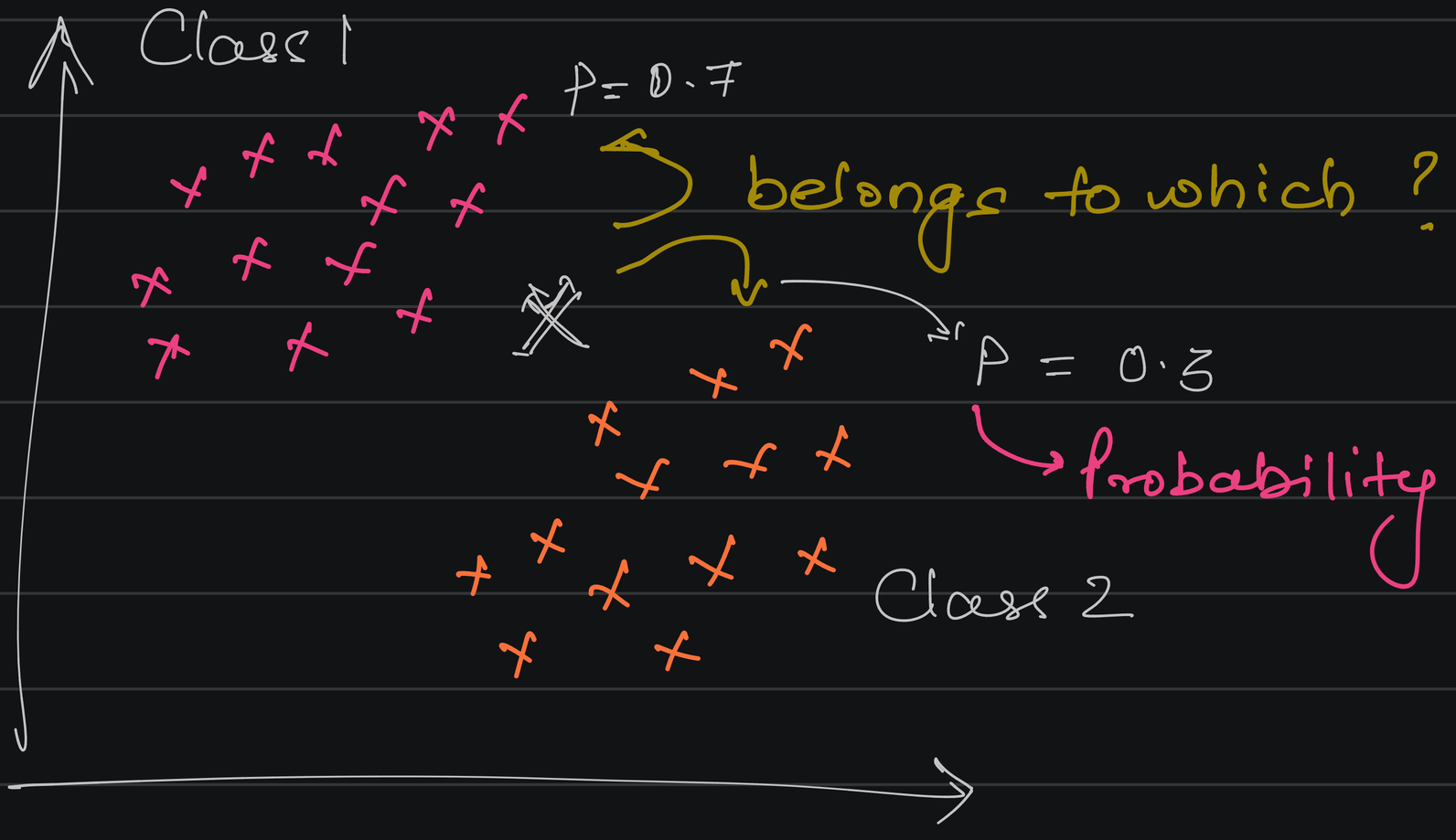
COIN. → P(H) = 1.

⇓

Unfair coin.

Rolling a dice — $P(1-6) = \frac{1}{6}$

Use in ML : Classification Problem.



(i) Mutual Exclusive event.

Two events are mutually exclusive if they cannot occur at the same time.

e.g. Rolling of dice, Tossing a coin.

(ii) Non-mutually Exclusive event.

Two events can occur at the same time.

e.g. Picking randomly a card from a deck of cards, two events "heart" and "king" can be selected.

Mutually exclusive

Q: What is the probability of coin landing on heads or tails ?

⇓

Addition rule for mutual exclusive events.

$$P(H) = 0.5.$$

$$P(T) = 0.5.$$

$$P(H \text{ or } T) = P(H) + P(T)$$

$$= 0.5 + 0.5 = 1.$$

Q: What is the probability of getting 1 or 3 or 6 in a dice roll ?

$$P(1-6) = 1/6$$

$$P(1 \text{ or } 3 \text{ or } 6) = P(1) + P(3) + P(6)$$

$$= 1/6 + 1/6 + 1/6$$

$$= \frac{3}{6} = 1/2$$

Non-mutual Exclusive event

Problem 1. = Bag of marbles : 10 Red, 6 Green

3 (R & G)

When picking randomly from a bag of marbles what is the probability of choosing a marble that is red or green?

Addition rule for non-mutual exclusive

Ans = Even 1.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(R \text{ or } G) = \frac{13}{19} + \frac{9}{19} - \frac{3}{19}$$

$$= \frac{19}{19} = 1.$$

Deck of cards. ⟶ Probability of choosing 'Heart' or 'Queen'.

$$P(\text{Heart or Queen}) = P(H) + P(Q) - P(H \cap Q)$$

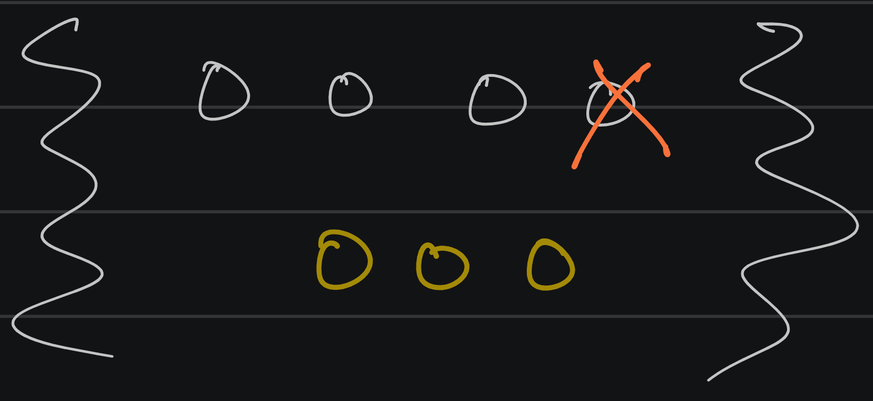
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52}$$

$$= \frac{16}{52} = \frac{4}{13}$$

Multiplication Rule.

(i) **Dependent events.** Two events are dependent if they affect one another.

Bag of marbles



$$P(\text{White}) = \frac{4}{7}$$

1st draw.

⟶ $P(\text{Yellow}) = \frac{3}{6}$

2nd draw.

Independent Event:-

e.g. Q What is the probability of rolling a 5 & then a 3 with a normal six sided dice ?

Ans. = $P(1 \text{ to } 6) = \frac{1}{6}$

Multiplication rule for independent events.

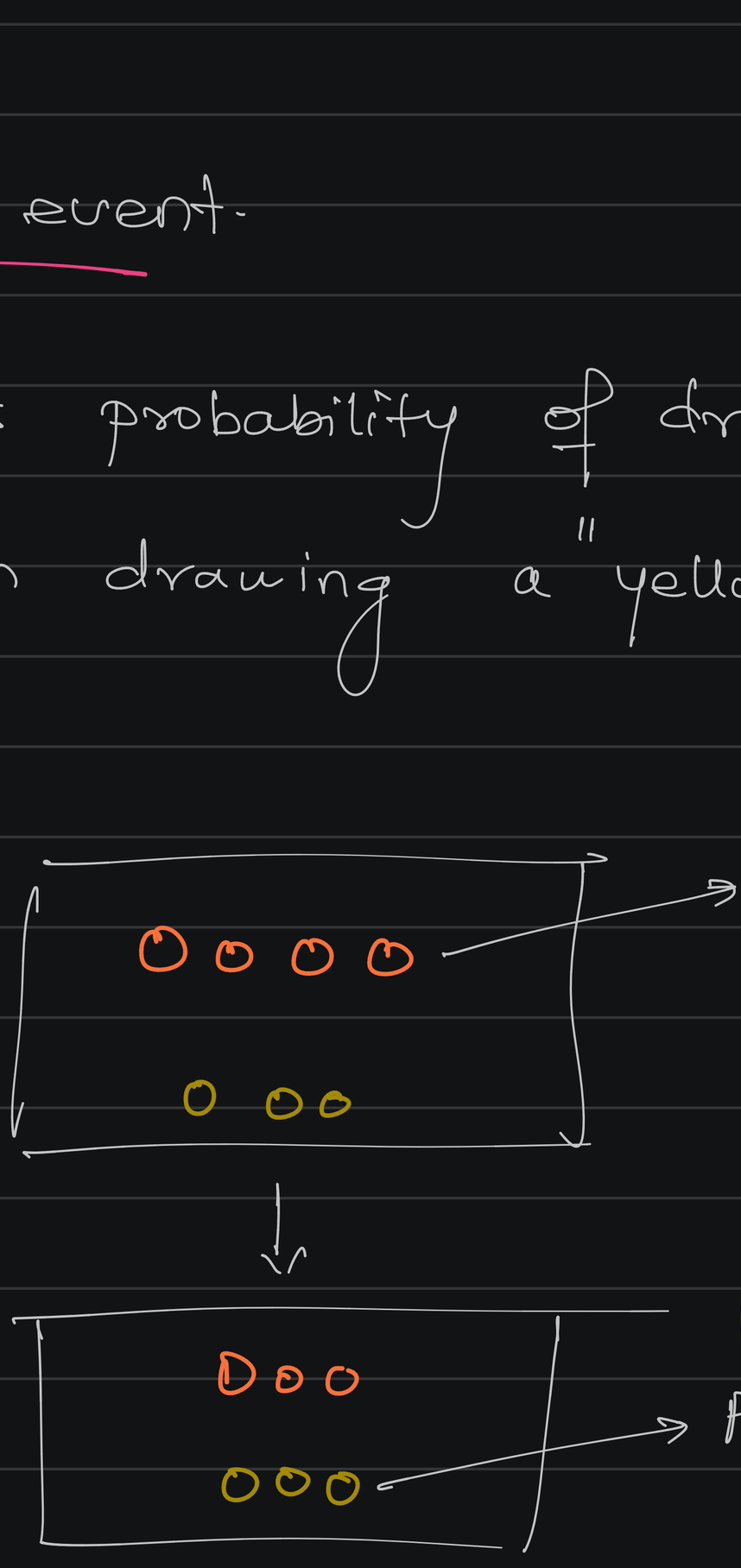
$$P(A \text{ and } B) = P(A) * P(B)$$

$$= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Dependent event.

e.g Q What is probability of drawing a "Orange" and then drawing a "yellow" marble from the bag?

Bag:



$$P(O \text{ and } Y) = P(O) * P(Y/O)$$

$$= \frac{4}{7} \times \frac{3}{6} = \frac{2}{7}.$$

3.

(i) Permutation.

eg.

Chocolates.

{ dairy milk, kit kat, milky bar, Sneakers, 5-star. }

Arrange these in the available spacel.

5 4 3

Possibilities.

Total no. of ways = $5 * 4 * 3$.

= 60 ways.

* With permutation, order matters.

Formula.:

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

n = total no. of objects.

r = # of selection

(from previous example) $P = \frac{5!}{(5-3)!} = \frac{5 \times 4 \times 3 \times 2!}{2!}$

$= 60.$

(ii) Combination. Repeatation will not occur. (Only unique combinations allowed)

Formula.:

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

(from previous example) $C = \frac{5!}{3! \times 2!} = \frac{5 \times 4^{2} \times 3!}{3! \times 2 \times 1}$

$= 10.$

Used in Dream 11. a lot.

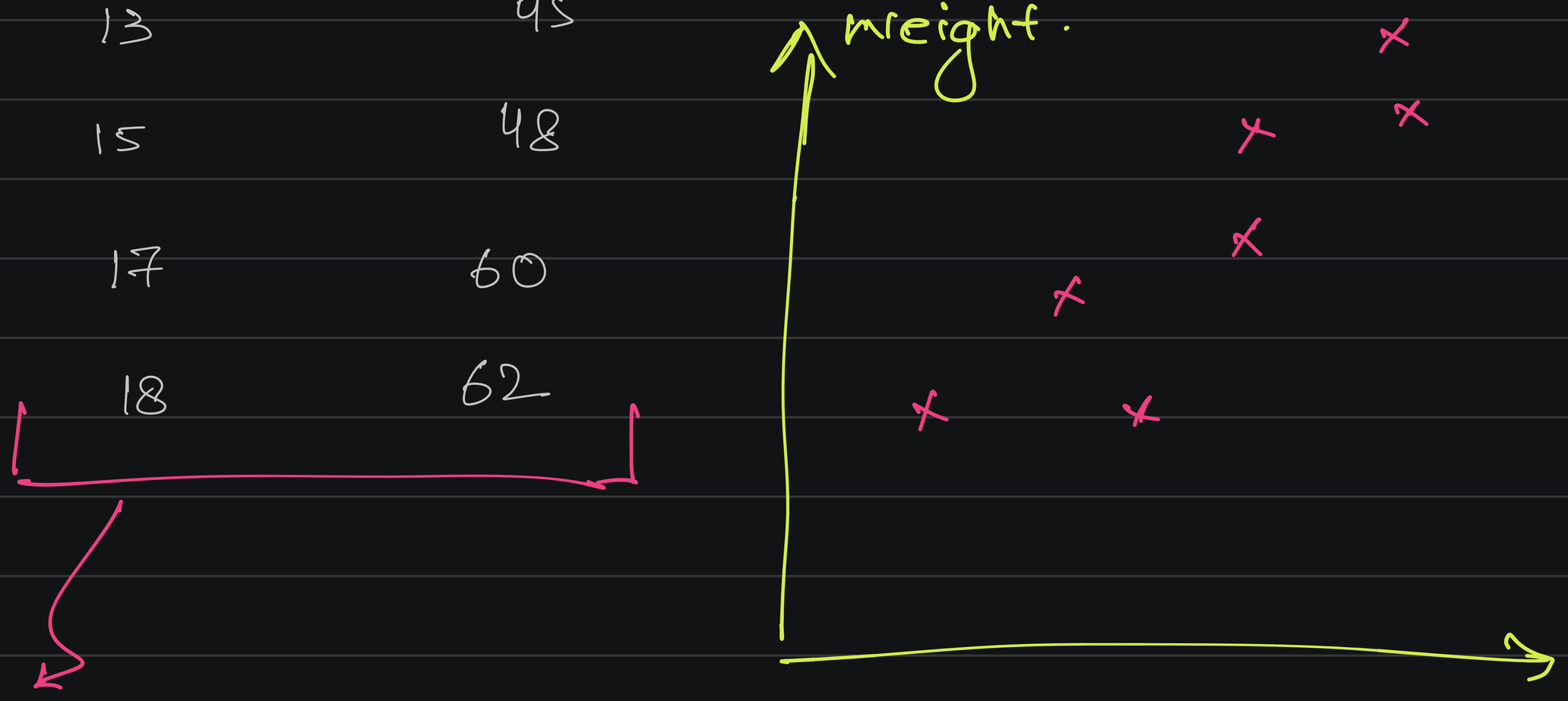
4

(i) Covariance → important for feature selection.

Age	Weight
12	40
13	45
15	48
17	60
18	62

Age ↑ Weight ↑

Age ↓ Weight ↓



Let's quantify the relationship x & y using mathematical operations.

Formula:

$$Cov(x,y) = \frac{\sum(x_i - \bar{x}) * \sum(y_i - \bar{y})}{n-1}$$

Note. :

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \frac{\sum (x_i - \bar{x}) * \sum (x_i - \bar{x})}{n-1}$$

So, $\boxed{Cov(x,x) = \sigma^2(x)}$ → Interview question.

- for previous example,

$\bar{x} = 15$ $\quad$ $\bar{y} = 51$

$Cov(x,y) = 24.$ → Positive value.

↓

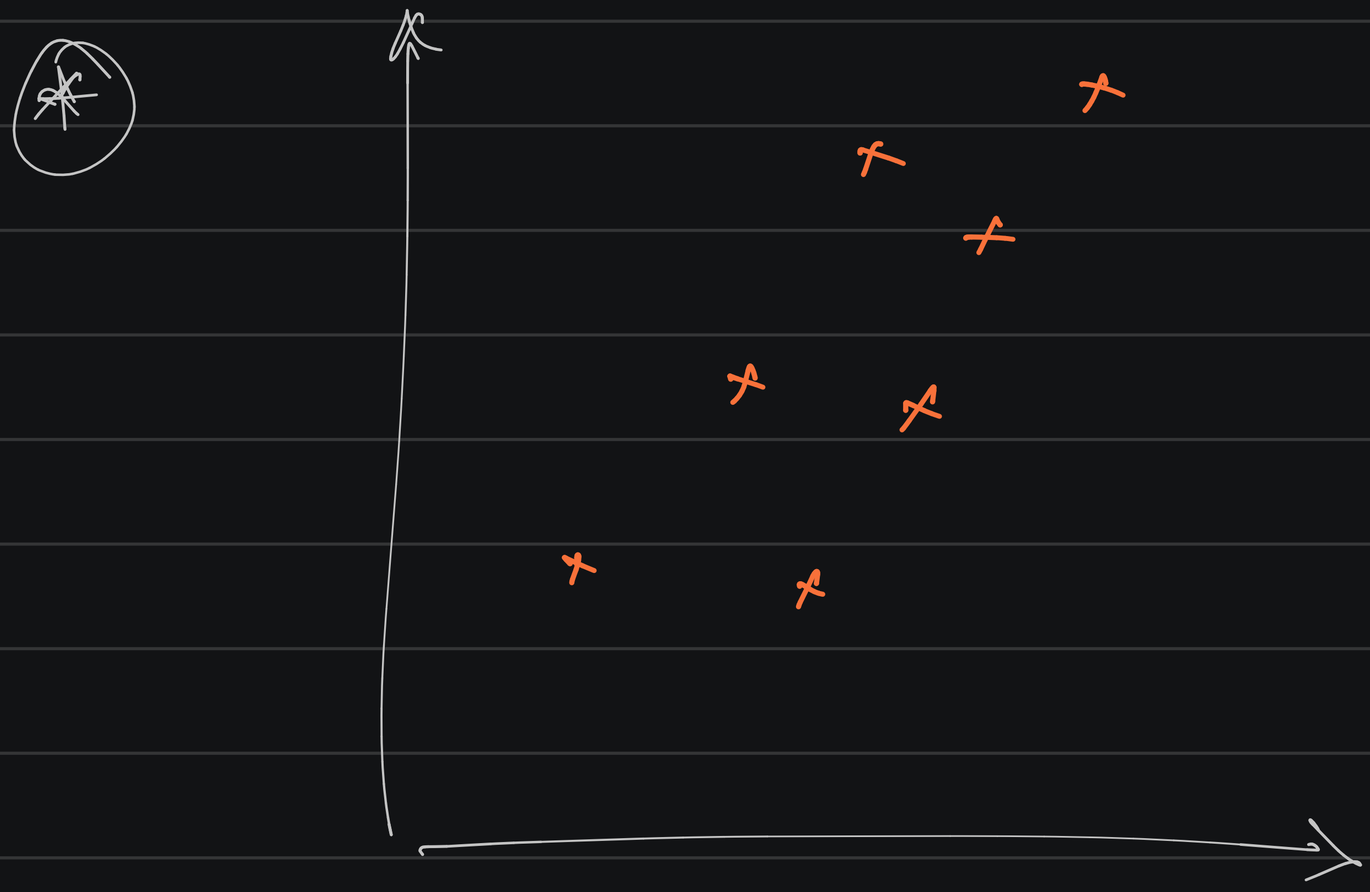
Age ∝ Weight.

if Negative value,

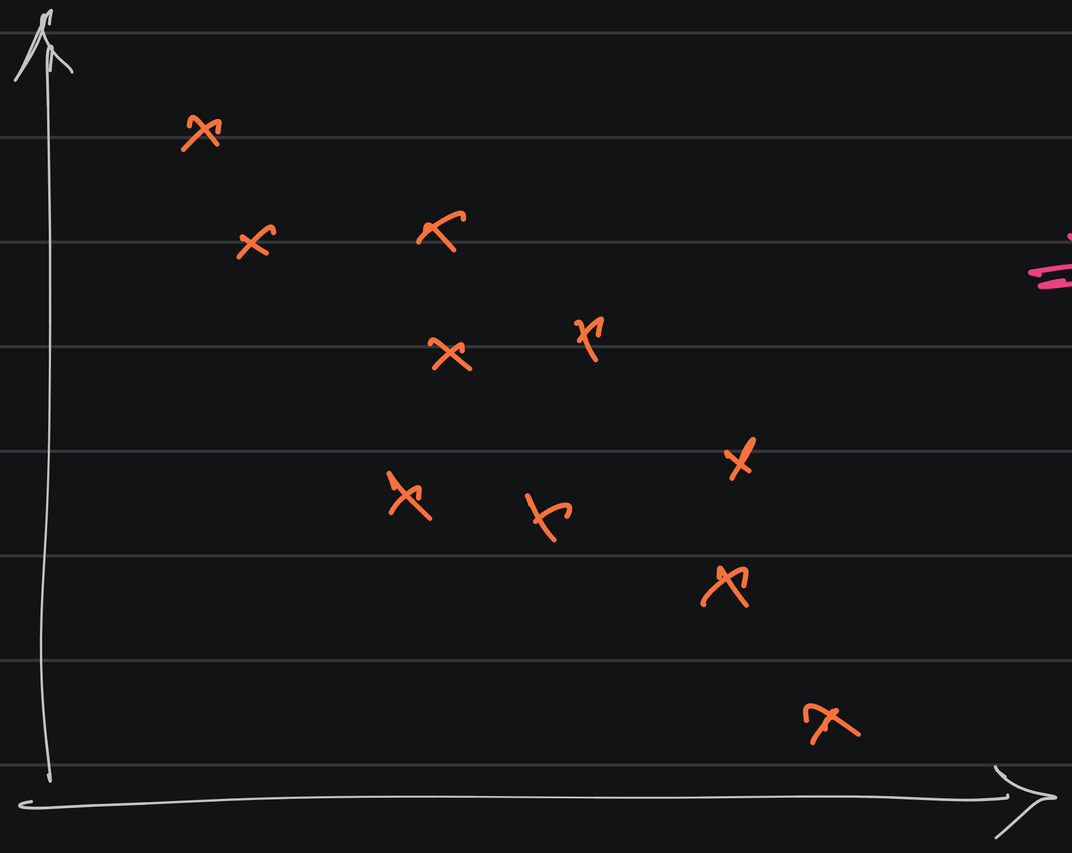
$x \propto \frac{1}{y}$.

↳ proportionality sign.

if Covariance = 0, no relation between x and y.



⇒ +ve Covaiance.



⇒ -ve covariance

⇒ No relation.

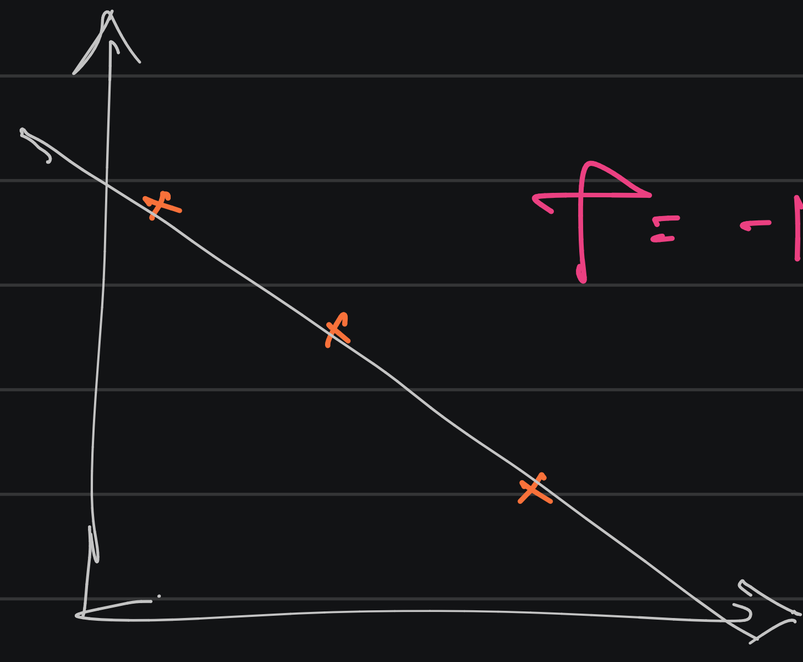
(ii) Pearson correlation coefficient. → tries to restrict the correlation between −1 and 1.

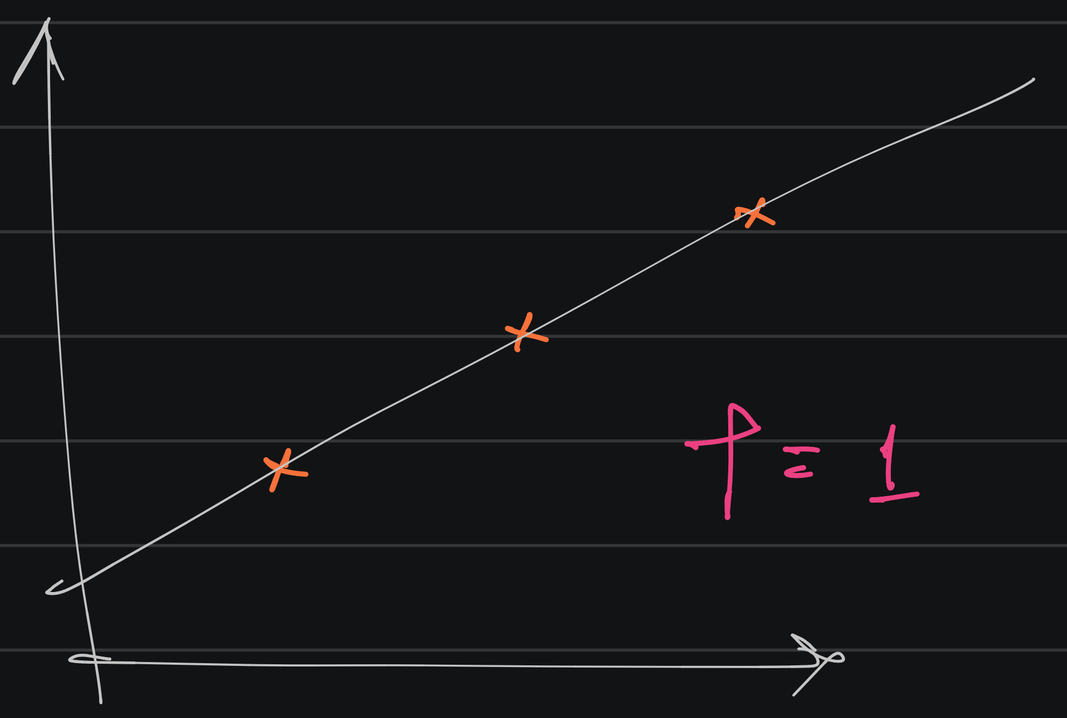
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

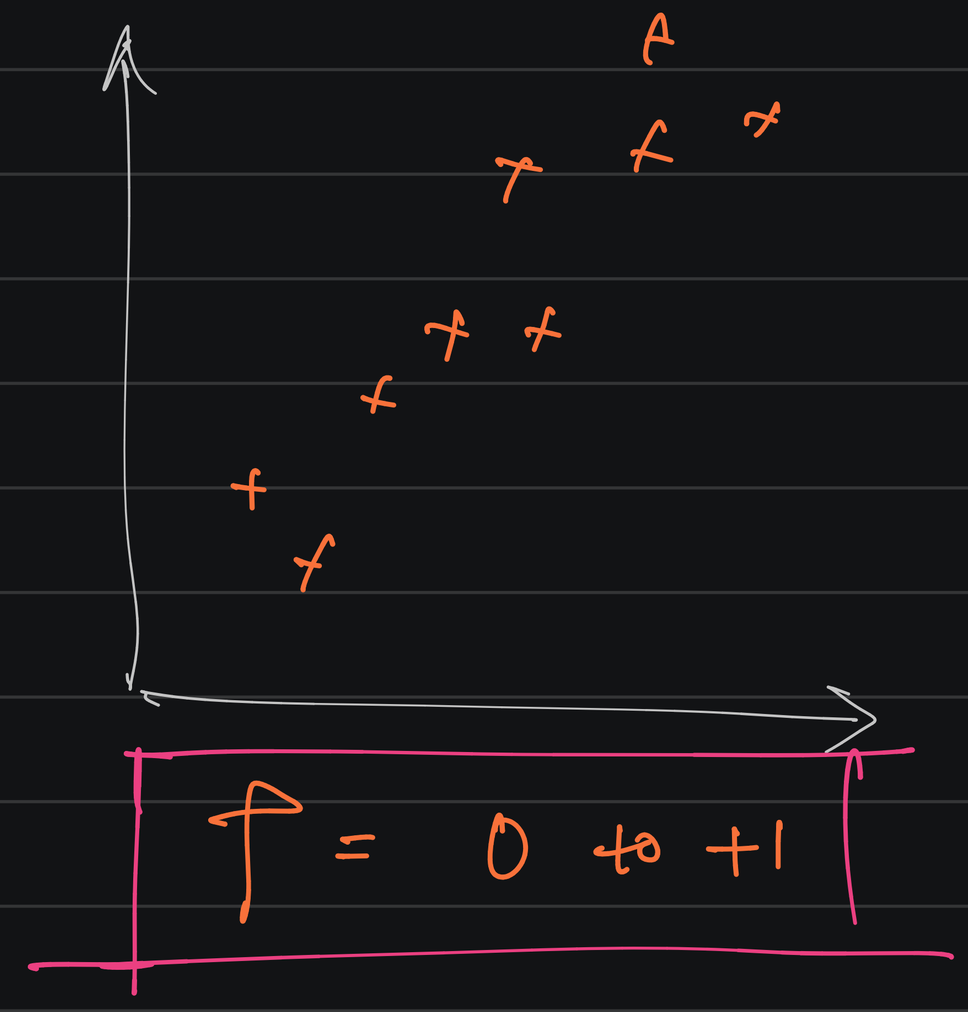
rho

Move value towards +1 → ↑ positive correlation.

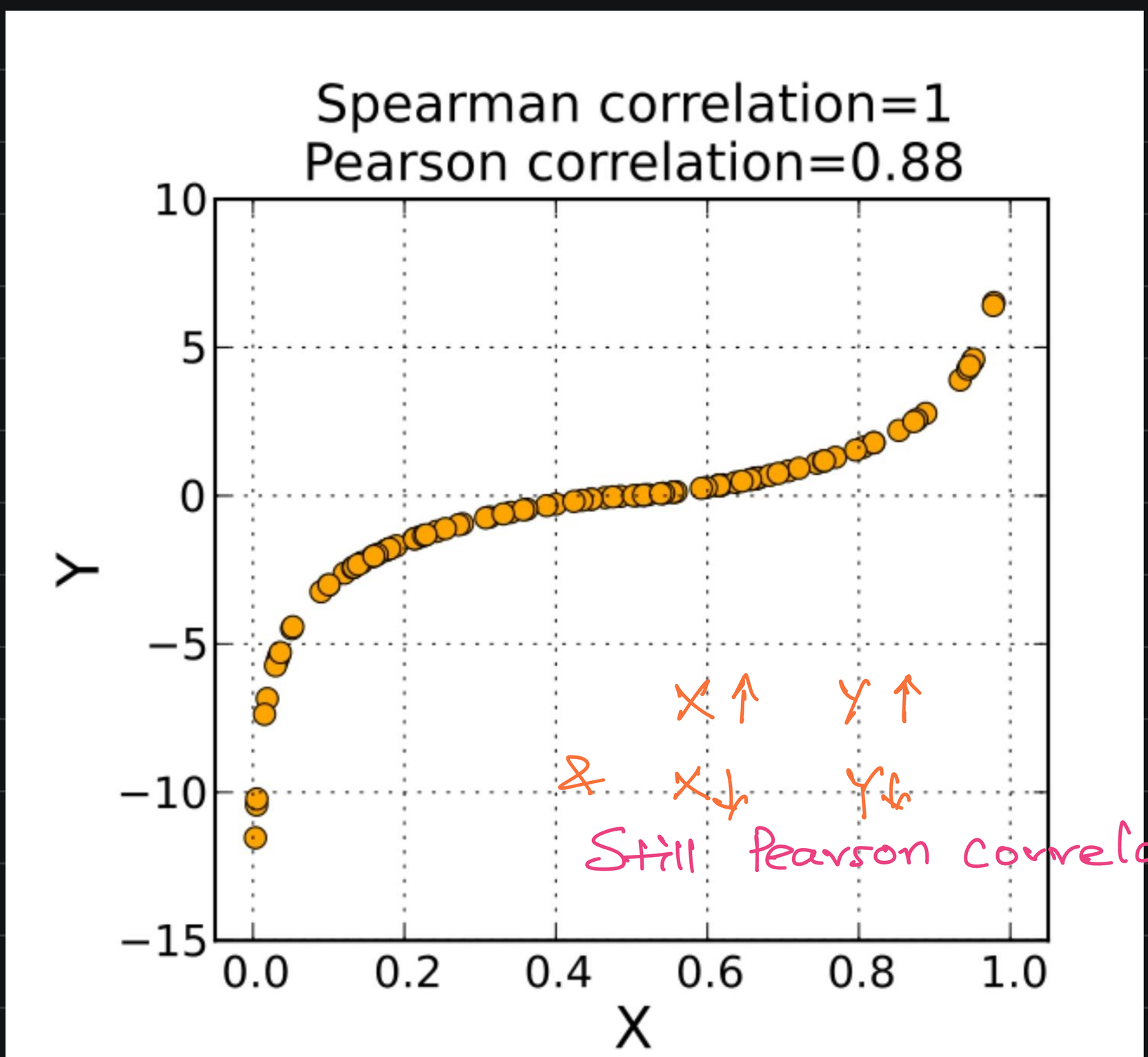
Move value towards −1 → ↑ negative Correlation.







(bc 111) Spearmann's rank correlation.



* Pearson correlation ⟶ does not work.

on linear data.

⇓

So, spearman correlation

is used.

Assigns ranks.

Formula.,

Rank of x/y.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

e.g.

x	y.
10	4
8	6
7	8
6	10.

R(x).	R(y)
4	1
3	2
2	3
1	4.

Use ranks to calculate spearman Correlation

Why correlation is used ?

Important -feature if highly corelated

Keep.

features. ⇒ X Y Z Output

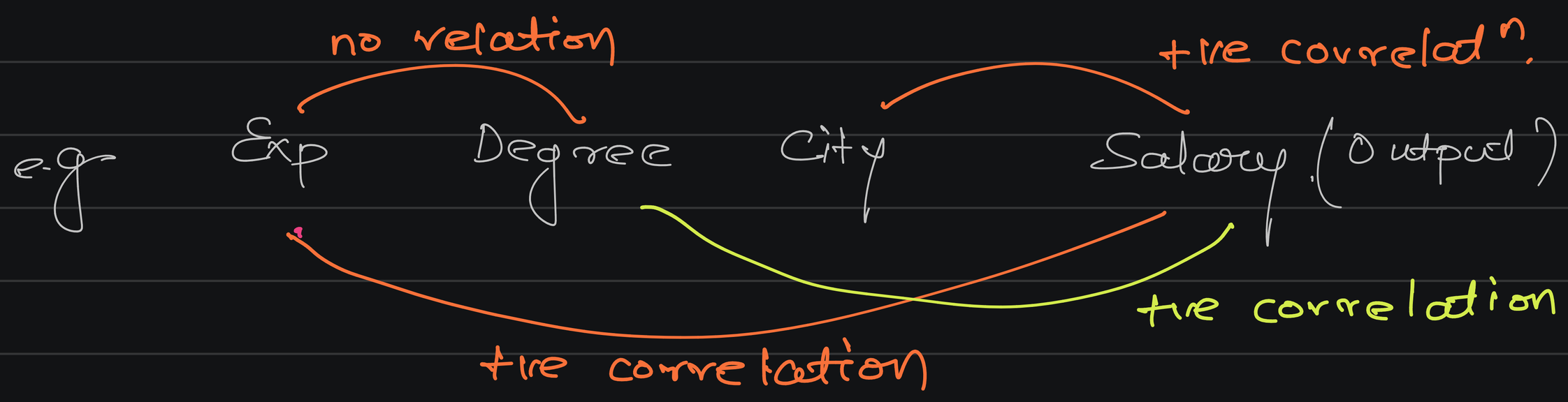
Highly correlated (95%)

low correlation

Not important -features

Can be dropped.

One of the features can be dropped as both have similar impact.


no relation
+ve correlation
e.g.
Exp
Degree
City
Salary (Output)
+ve correlation
+ve correlation

Day 5.

Inferential Statistics

1. Hypothesis testing.
2. p-value.
3. Confidence interval
4. Significance value.

- Z-test
- t-test
- Chi square test
- Annova test (F-test)

3 distributions

i. Bernoulli's
ii. Binomial.
iii. Power Law.

→ Transformation.

⊛ Inferential Statistics.

⇓

Hypothesis Testing is used.

Steps of Hypothesis testing

(1.) Define Null Hypothesis — eg The person is not a criminal.

by default → Null hypothesis → TRUE

(2.) Define Alternate hypothesis :— Opposite to null hypothesis

eg The person is a criminal.

(3) Perform experiments

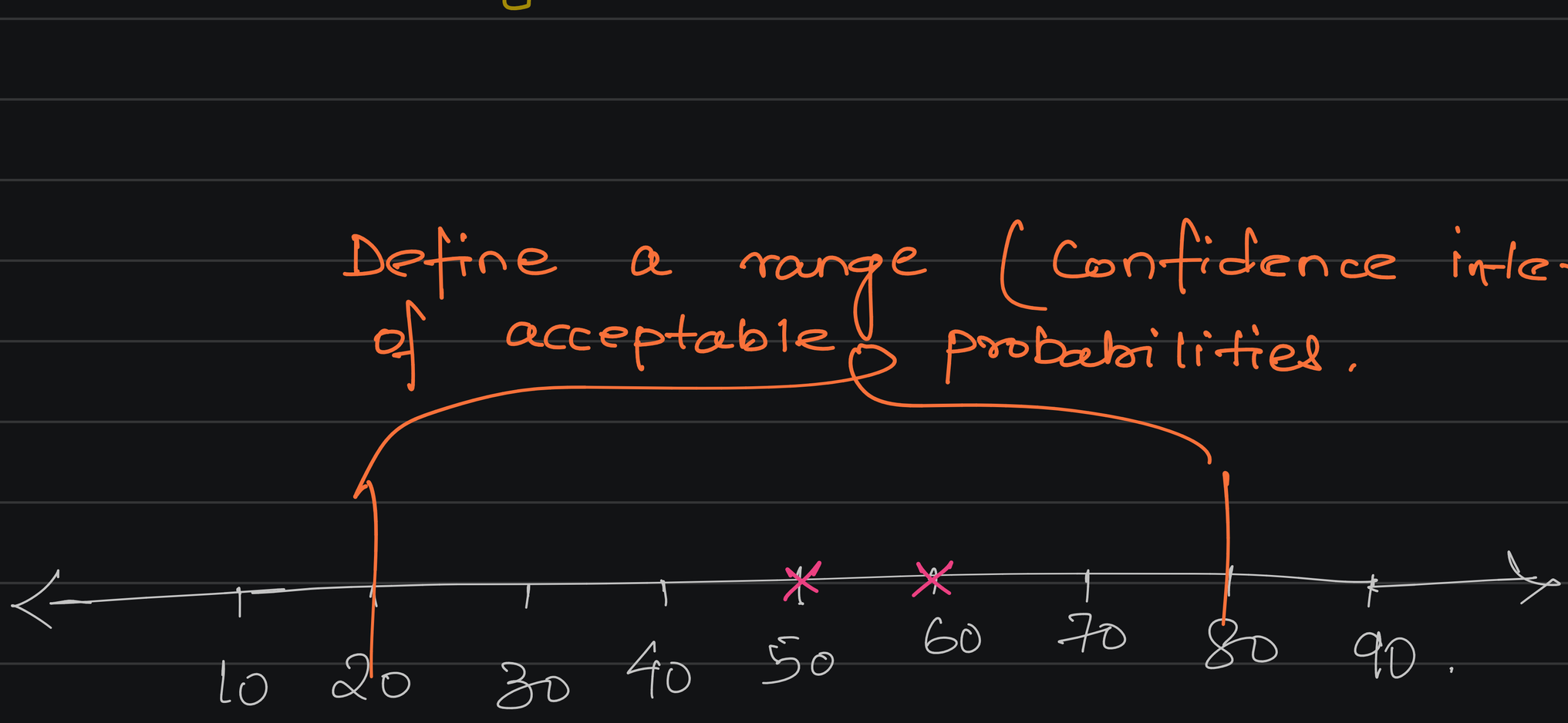
Exp. = "Coin is fair or not"

Null hypothesis ⟶ Coin is fair.

Alternate hypothesis ⟶ Coin is not-fair.

Experiment: Coin tossed 100 times.

$\mu = 50$, $SD(\sigma) = 10$.



→ rounds

Result: (1) 50 times head. — Fair coin

(2) 60 times head. — Fair coin.

* if probability goes beyond Confidence interval ⟶ Unfair Coin

(3.) 70 times head. ⟶ Ask domain expert.

↓

Define Confidance Interval.

$CI = [20 - 80]$

↳ Coin is fair.

So, null hypothesis can be accepted.

(4.) 10 times head. ⟶ Un-fair coin.

(Outside C.I.)

So, null hypothesis will be rejected.

& alternate hypothesis can be accepted.

(4.) Conclusion.

↳ We fail to reject the null hypothesis.
(within C.I.)

↳ We reject the null hypothesis.
(outside C.I.)

e.g 2 Person is criminal or not (murder).

1. Null hypothesis — Person is not criminal.

2. Alternate hypothesis — Person is criminal.

3. Experiment / Proofs — DNA, finger print, weapons, footage.

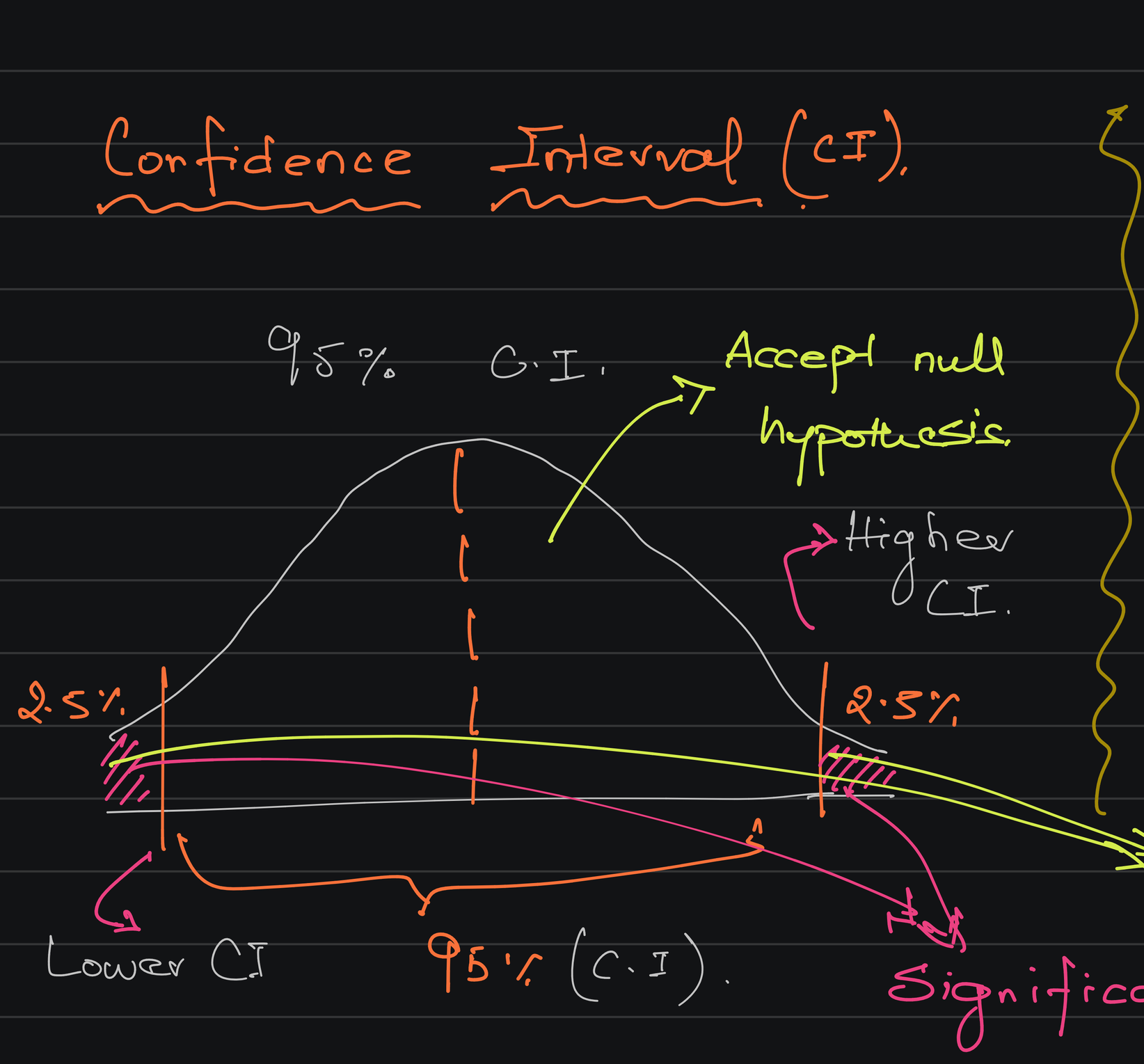
↓

Conducted by judge.

4. Conclusions — Accept / Reject null hypothesis

Confidence Interval (CI).

Significance value



if C.I. = 95%

S.V = 1 − 0.95.

= 0.05.

Reject null hypothesis

Point Estimate — the value of any statistic that estimates the value of a parameter is called point estimate.

Statistic. ← $\bar{x}$ → estimate. → parameter.

eg. $\bar{x}$ (sample mean) → μ (population mean)

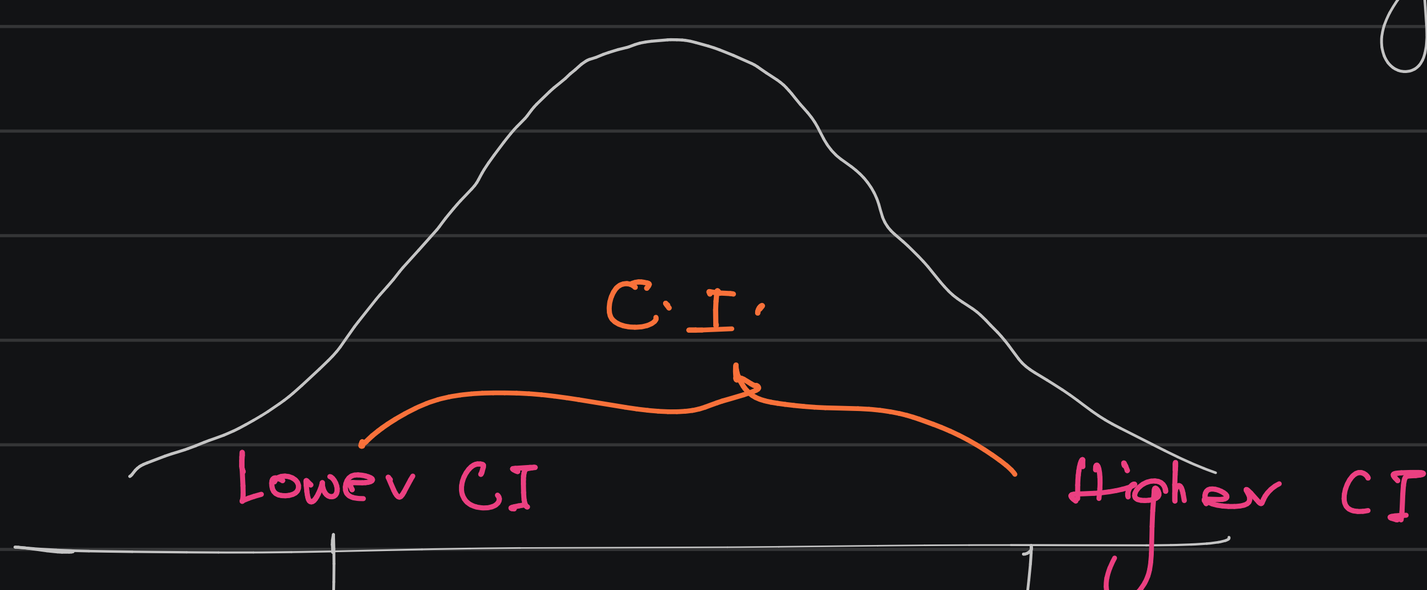
Point Estimate ± Margin of error. = Parameter. (population mean)

As, $\bar{x} \leq \mu$.

or $\bar{x} \geq \mu$.

Lower CI = Point Estimate − Margin of error

Higher CI. = Point estimate + Margin of error



(+) Margin of error = $Z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$

σ → population SD

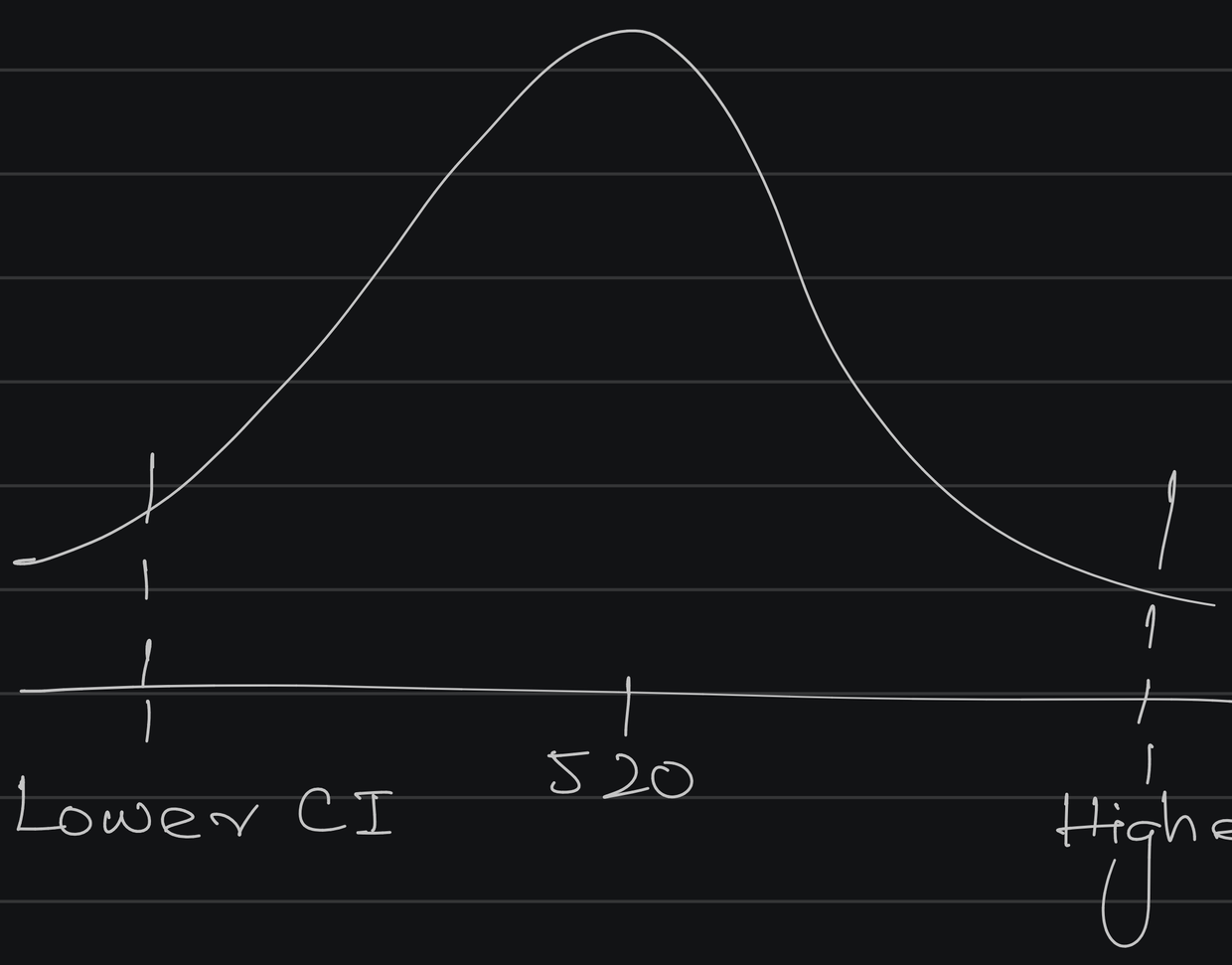
$\frac{\sigma}{\sqrt{n}}$ ⇒ Standard error.

$Z_{\alpha/2}$ → significance value.

Q On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a population SD of 100. Construct a 95% CI about the mean?

Ans $n = 25$, $\bar{x} = 520$, $\sigma = 100$

CI = 95%

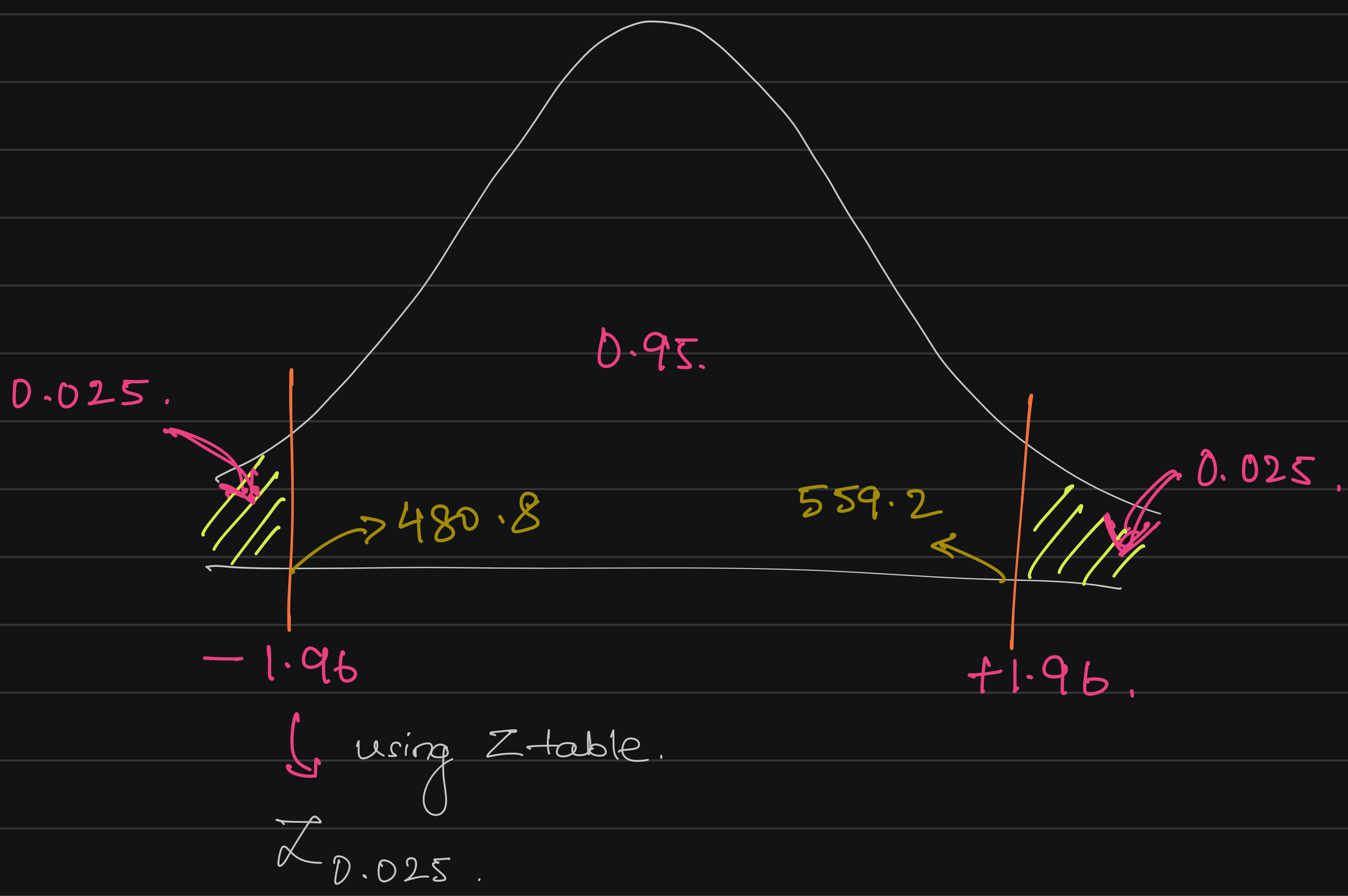


$$\text{Lower CI} = \text{Point estimate} - \text{Margin of error.}$$

$$= 520 - Z_{0.05/2} \frac{\sigma}{\sqrt{n}}.$$

$$= 520 - Z_{0.025} \frac{100}{\sqrt{25}}.$$

$$= 520 - (1.96 \times 20) = 480.8$$



↳ using Z-table.

$Z_{0.025}$.

$$\text{Higher CI} = 520 + 1.96 * 20 = 559.2$$

$$\therefore \text{C.I.} = \{ 480.8 \text{ to } 559.2 \}$$

Q. $\bar{x} = 480$, $\sigma = 85$, $n = 25$.

$CI = 0.90. (90\%)$

Find the C.I range.

Lower C.I = Point estimate − Margin of error.

$= 480 - Z_{0.1/2}\left(\frac{85.}{\sqrt{25}}\right)$

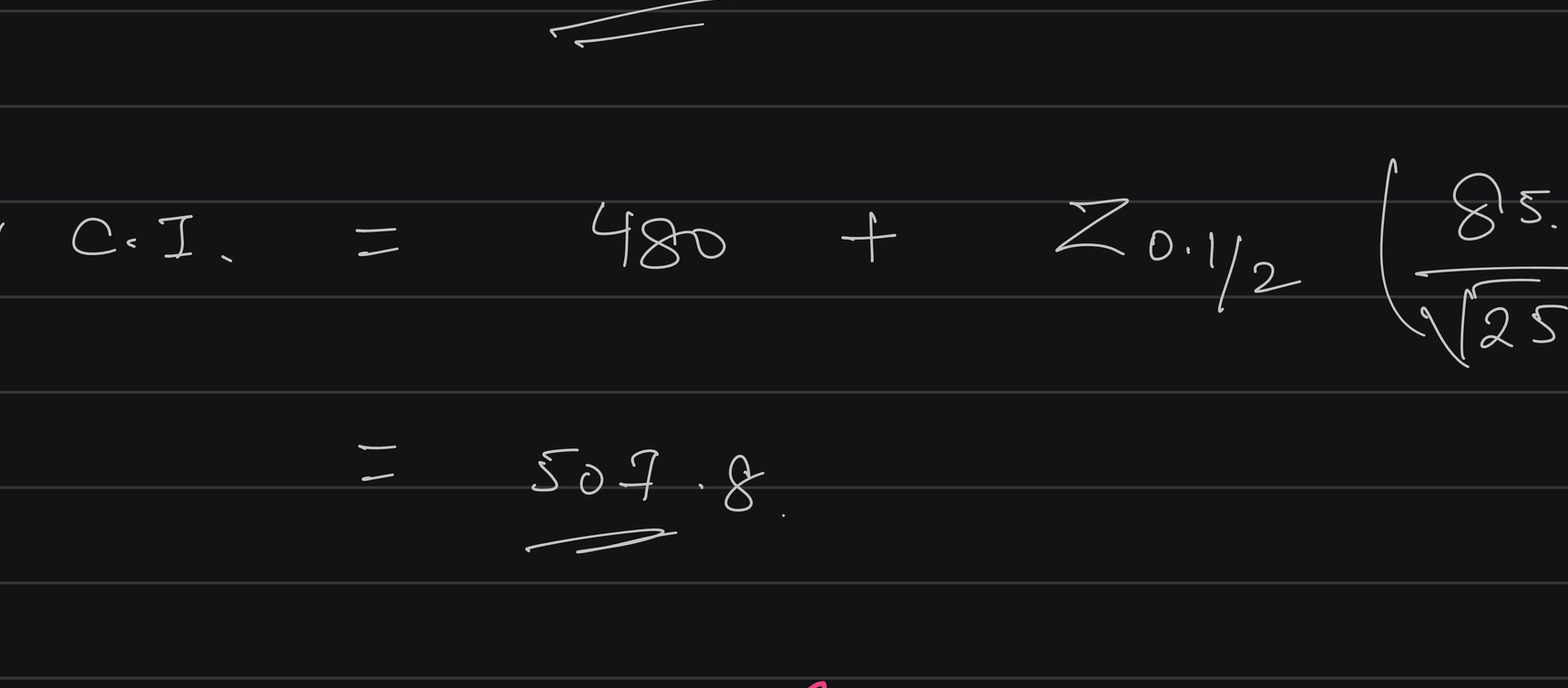
$= 480 - Z_{0.05} * \frac{85.17}{8.}$

$= 452.12$

Higher C.I. $= 480 + Z_{0.1/2}\left(\frac{85.}{\sqrt{25}}\right)$.

$= 507.8$.

C.I range = 452.12 to 507.8.



Q On the quants test of CAT exam, a sample of 25 test takers has a mean of 520, with a sample standard deviation of 80. Construct 95% C.I. about the mean?

Ans. = $\bar{x} = 520$, $s = 80$, C.I. = 95%

S.v. = $1 - 0.95 = 0.05$.

$$\bar{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

t - test

Degree of freedom = $n - 1 = 25 - 1$

$= 24$.

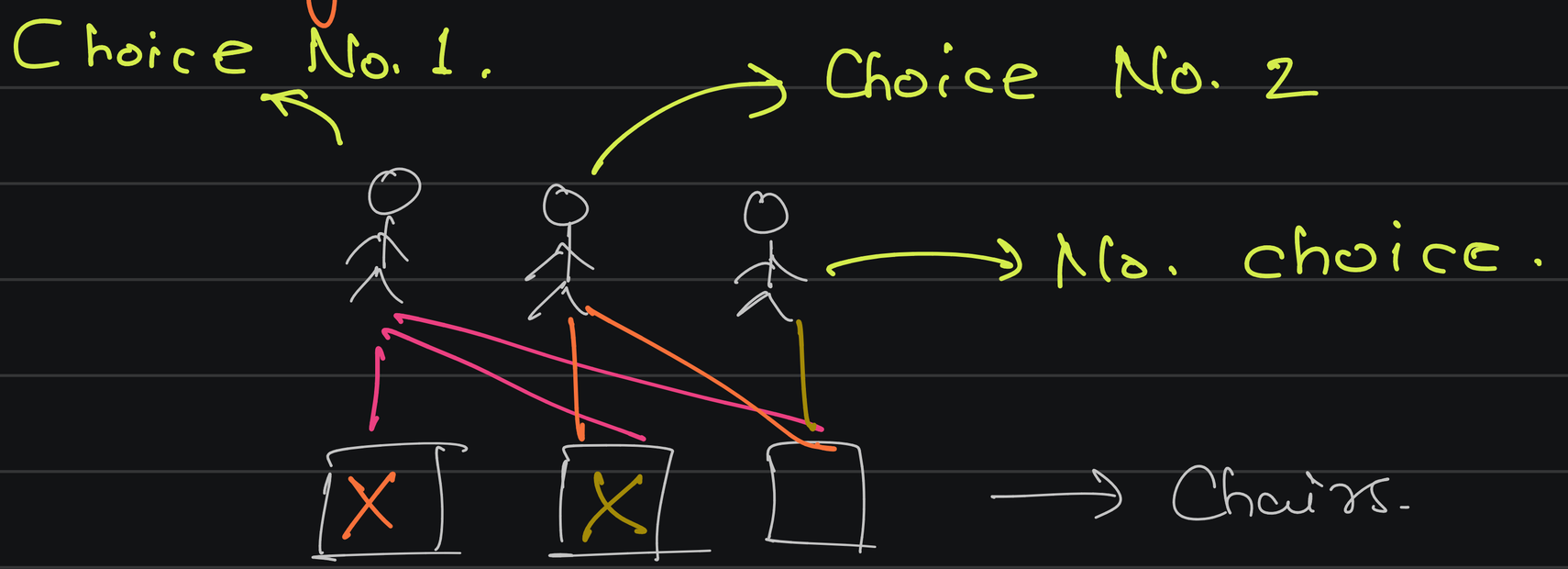
Lower C.I = $520 - \frac{t_{0.05}}{2}\left(\frac{80}{\sqrt{25}}\right)$

$= 520 - t_{0.025}\left(\frac{80}{5}\right)^{16}$.

$= 520 - (2.064 * 16)$

Lower CI = 486.976. - from t - table

✳ Degree of Freedom.



∴ In a sample of 'n' elements,

Degree of Freedom = n - 1.

Higher CI = 520 + 2.064 * 16.

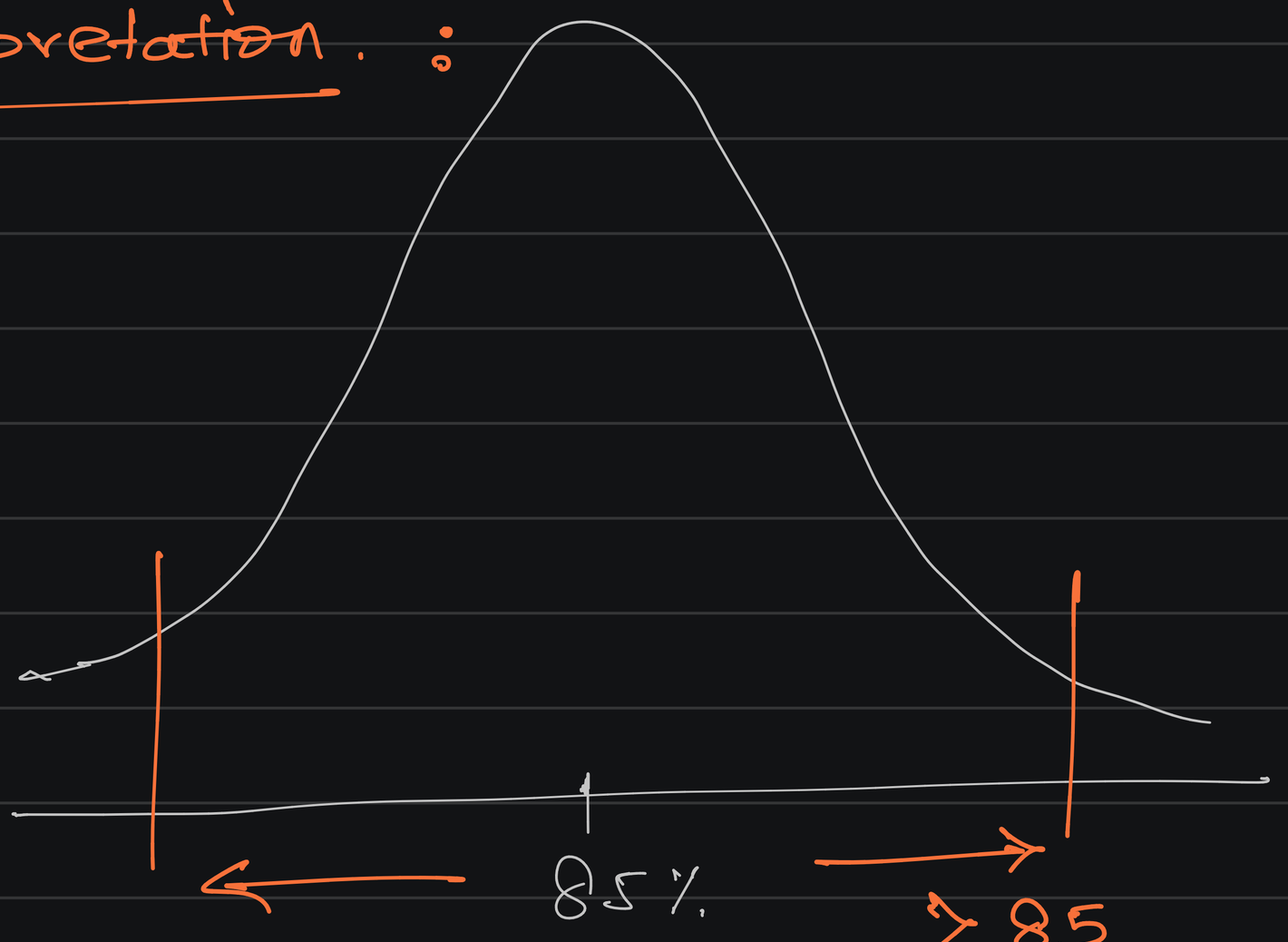
= 553.024

(1) 1-tail and 2-tail test.

Q = Colleges in Town A has 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88% with a standard.

deviation of 4%. Does this college has a different placement rate with CI = 95%?

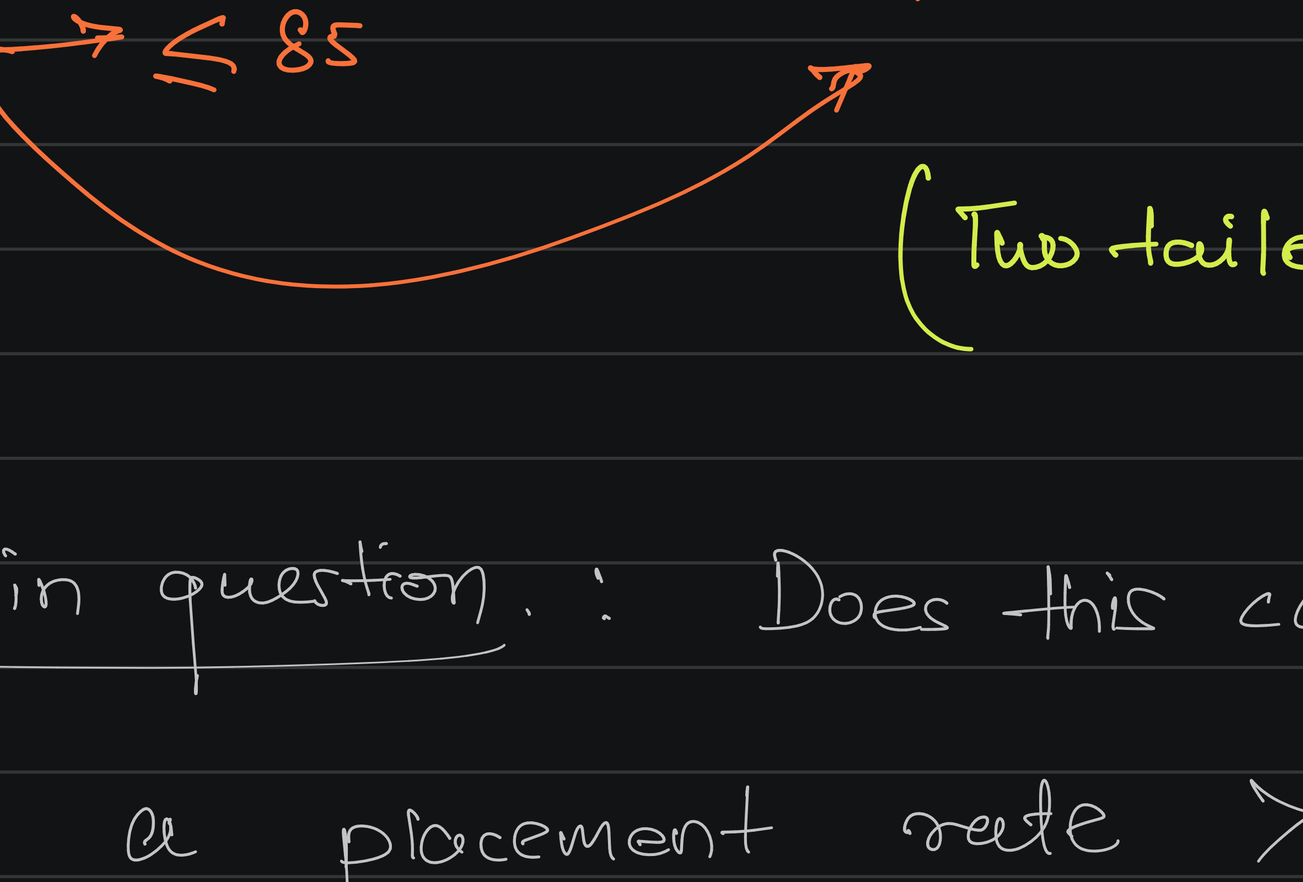
Interpretation. :



can be : → ≤ 85

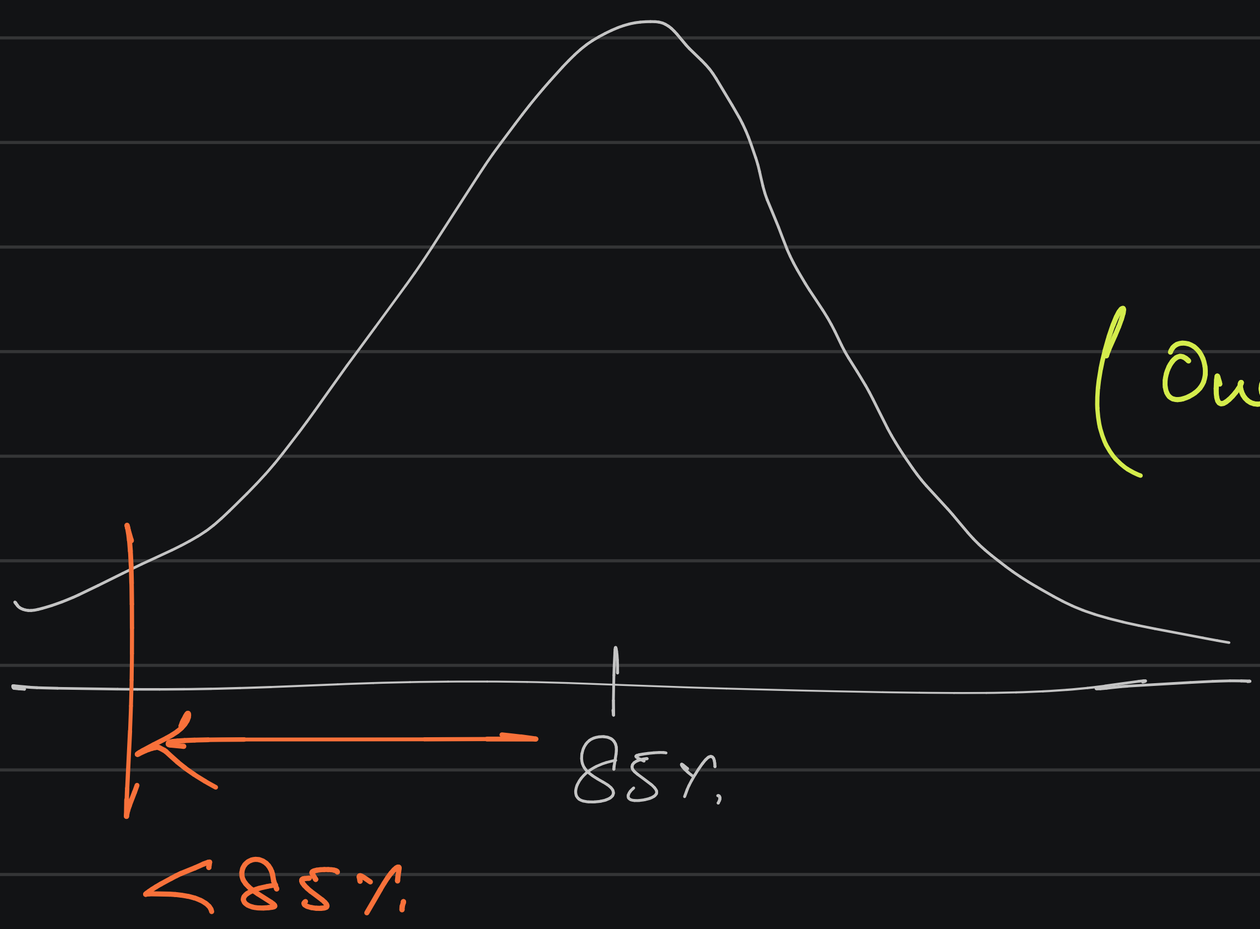
(Two tailed test.)

Change in question. : Does this college has a placement rate > 85%.



(One-tailed test)

if. Can placement rate be < 85%.



(One tailed test)

Hypothesis testing.

1. Z-test.

2. t-test.

Q = A factory has a machine that fills 80ml of baby medicine in a bottle. An employee believes, the av. amount of baby medicine is not 80ml. Using 40 samples he measures the average amount dispersed by the machine to be 78ml with a SD = 2.5.

(a.) State null and alternate hypothesis.

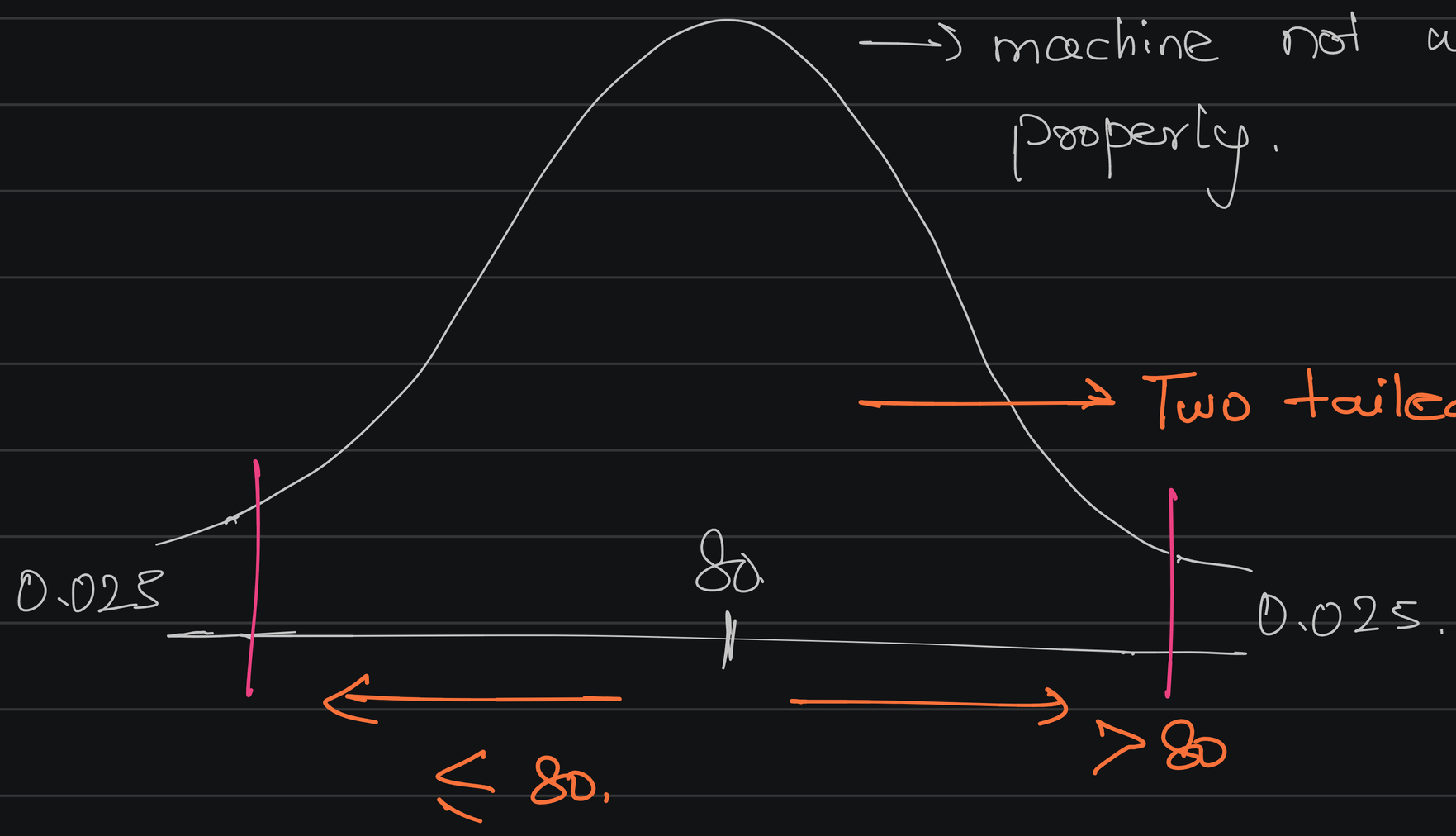
(b.) At CI = 95%, is there enough evidence to support Machine is working properly or not.

Ans. Null Hypothesis : $\mu = 80$ ml.

→ machine working properly.

Alternate hypothesis : $\mu \neq 80$ ml.

→ machine not working properly.



Step-2 : CI = 0.95, $\alpha = 0.05$

Step 3. : $n = 40$.

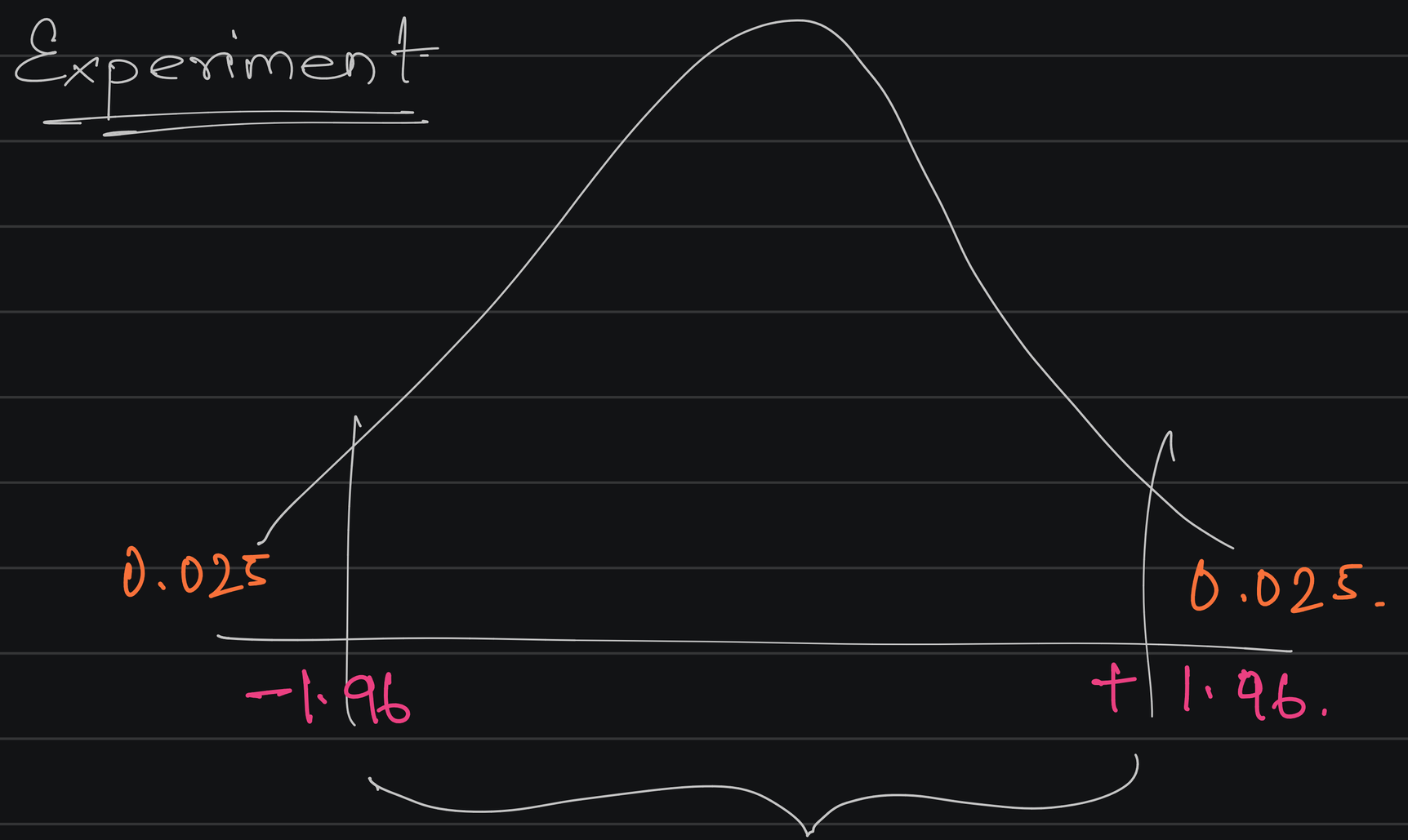
$s = 2.5$.

Z-test reqd.

* $n \geq 30$ or population SD (σ) given. } Z-test

* $n < 30$ and Sample SD (s) given } t-test

Experiment



Decision boundary.

Calculate test statistics. (z-test)

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ → Standard error.

$$= \frac{78 - 80}{2.5/\sqrt{40}}$$

$$= -5.05$$

Conclusion :-

Decision rule : If $z = -5.05$ is less than -1.96 or greater than $+1.96$, reject null hypothesis.

So, we reject the null hypothesis.

Conclusion : Machine is not working properly.

Q A complain was registered, the boys in a govt. school are underfed. Average weight of the boys of age 10 is 32 kgs with SD = 9 kgs. A sample of 25 boys were selected from the school, and the average weight was found to be 29.7 kgs. With C.I = 95%, check if it is true or false.

Ans. $\sigma = 9$ kg., $n = 25$.

$\mu = 32$ kg.

→ Z-test.

Step 1. Null Hypothesis : Boys are not underfed.

Alternate hypothesis : Boys are underfed

Step 2 : C.I. = 0.95, $\alpha = 0.05$.

Step 3 :

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{29.7 - 32}{9/\sqrt{25}}$$

$$= -\frac{2.3 \times 5}{9} = -1.27$$

within decision boundary.

$\therefore$ Null hypothesis is accepted

Conclusion :- Boys in the government School are not underfed.

Statistics

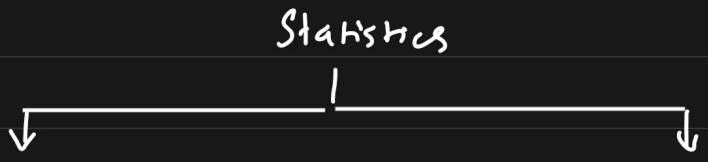
Statistics is the science of collecting, organizing and analyzing data.

Data : "facts or pieces of information"

Eg : Height of students in a classroom
→ {175cm, 152cm, 140cm, 130cm, 155cm}

Eg : Intelligence Quotient (IQ) of 5 randomly selected individuals (109, 89, 129, 101, 105, 106) → Data.

Two Types



① Descriptive Stats

- It consists of organizing and summarizing of data.
- Eg : Pdf, Histogram, Box plot, Bar chart, Pie charts

② Inferential Stats

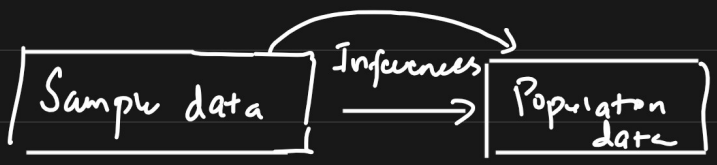
- It consists of using that you've measured to form conclusions
- Eg : Hypothesis Testing, p value, Z test, t-test, Anova, Chisquare

Eg : Lets say there are 20 maths classes at your university and you've collected the ages of students in one class.
Ages {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}

Descriptive stats : What is the average age of student in min = mode

your maths class?

Inferential question : Are the ages of students in this maths classroom similar to what you would expect in a normal maths class at this university?



Population And Sample Data. → Inferential Statistics → Results

Exit Polls ←

Elections : Punjab {AAP, Congress}

↙ Population of Punjab

1000 people ⇓

Eg: 2015 → Data Scientist

Jackets → Size ✗

10K, 40K → Christmas

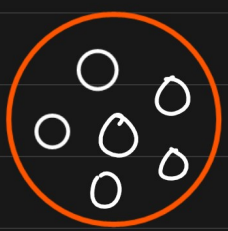
1-2% ⇒ Jackets

Population (N) ✓ Sample (n) ✓

Sampling Techniques

⇓

① Simple Random Sampling : Every member of the population (N) has an equal chance of being selected for your Sample (n)



② Stratified Sampling

Strata → Layers ⇓ Non overlapping groups

Clusters ⇑ groups

Gender → Male, Female

Age groups → 0-18, 18-35, 35-60

Blood Groups

Tax Slabs

Courses

Education qualification

Thanos

Australia

↳ Snap

Customs

③ Systematic Sampling

(N) → Select every nth individual

6th

3-4th

Stratified

Eg: Survey → Mail

(SBI credit card)

④ Convenience Sampling : Only those people who are interested will only be participating.

Eg: Data Science → AI

Youtube survey →

Healthcare Disease

{ Blind people }

{ Economics } → DATA Science

→ RBI → House hold Survey → Female ← UPSC

Exit Poll : Stratified + Random Sampling

Variable

A variable is a property that can take on any value

Eg: Height = 182
150
145
160

{ 182, 150, 145, 160 }
⇓
No

Two kinds of variable

① Quantitative Variable → Measured Numerically { Add, Substract, x, ÷ }

② Qualitative Variable.

↳ Eg: Gender → Male, Female { Based on some {characteristics} we can derive categorical variables }

{ Quantitative ⟶ Qualitative Variable. }

Eg: IQ

0-10 ⇓ Less IQ

10-50 ⇓ Medium IQ

50-100 ⇓ Good IQ

Quantitative

- Discrete variable
  - Eg: Whole number
  - Eg: No. of Bank accounts
  - {2, 3, 4, 5, 6, 7} 2.5, x 2.75 x
  - Eg: Total No. of children in a family
- Continuous Variable.
  - Eg: Height = 172.5, 162.5cm, 163.5cm.
  - Rainfall: 1.35, 1.25, 1.75, 2.25cm
  - Weight
  - Temp

Stock price.

Eg: 2, 3, 4, 5

2.5, 2.75

Eg: Total no. of Employees in a Company Eg: 10K,

Ass:

① What kind of variable Marital Status is? Categorical

② What kind of variable Nile River length is? Continuous quantitative

③ What kind of " Movie duration is? " "

④ What kind of variable IQ is? " "

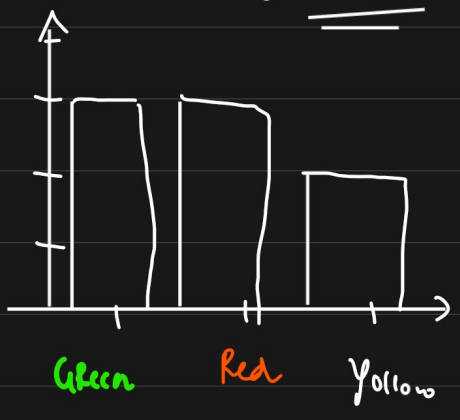
Frequency Distribution

Sample Dataset : Green, Red, Yellow, Green, Red, Yellow, Green, Red

Colors	Frequency
Green	3
Red	3
Yellow	2

① BAR GRAPH Frequency

Bar chart



⑧ Variable Measurement Scales

4 types of Measured Variable.

① Nominal data {Categorical data} → Ranking is not that important

Eg: Colors, Gender, Types of flowers

② Ordinal data

Percentiles

Student (Marks)	Rank
→ 100	1
96	2
57	4
85	3
44	5

Ordinal Data.

PHd → ⇓ ↙ {NLP}

Degree		Salary
PHd	1	✓
B.G	3	✓
Master	2	✓
BCA	4	✓
12	5	✓

③ Interval data ✓⇓

Assignment ⇓

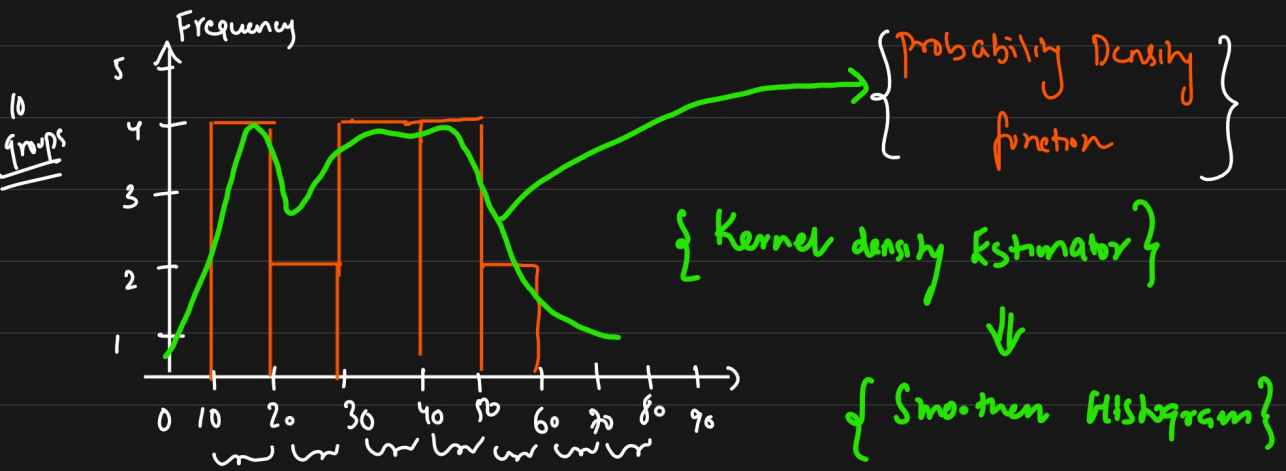
④ Ratio data ✓

⑧ Histograms ÷ Continuous

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

Histogram → Bins = 10

Mean, Median, Mode.



0 - 80 → 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35

Assignment

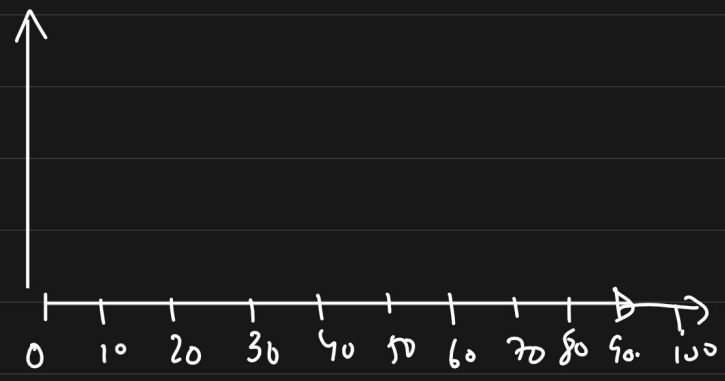
Eg: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

bins = 10

0-10 10-20 20-30 30-40

40-50 50-60 60-70

70-80 80-90 90-100

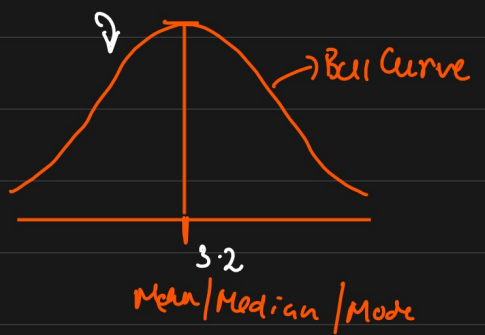


Intermediate Stats

1. Measure of Central Tendency
2. Measure of Dispersion
3. Gaussian Distribution
4. z - Score
5. Standard Normal Distribution
6. Central Limit Theorem

1. Measure of Central Tendency → Central position of the dataset

1. Mean ✓
2. Median ✓ {EDA & Feature Eng}.
3. Mode ✓



Population (N)

$X = \{1,1,2,2,3,3,4,5,5,6\}$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ ↓ Population mean

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$$

$$= \frac{32}{10} = 3.2$$

Sample (n)

Sample → $\bar{x}$ ↓ Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median

1, 2, 2, 3, 4, 5

$\bar{x} = \frac{1+2+2+3+4+5}{6} = \frac{17}{6} = 2.83$

1, 2, 2, 3, 4, 5, 100

$\bar{x} = \frac{1+2+2+3+4+5+100}{7} = \frac{117}{7} = 16.71$

Median ✓

1, 2, 2, 3, 4, 5, 100

$\bar{x} = 16.71$ //

Median = 3

1, 2, 2, 3, 4, 5 → odd or even

1, 2, 2, 3, 4, 5

$\frac{2+3}{2} = 2.5$

2.5 ≈ 2.83 //

Median

1, 2, 2, 3, 3, 4, 4, 5, 6

Error

[2, 3, 4]

Mode ÷ Highest frequency.

1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7

3

EDA

Feature Engineering

↳ NAN values ⇒

Continuous Values + Outlier

↳ Mean

Median

⇒ Categorical Variable.

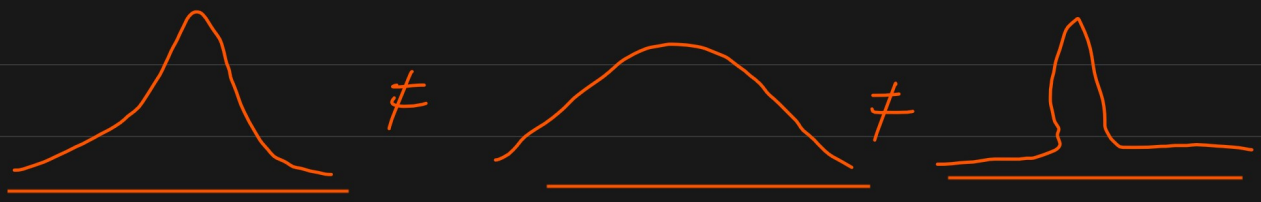
Mode

Agri — Libby, Sunflower, Rose, ..., MAN, MAN

Measure of Dispersion → {Dispersion}

① Variance
② Standard deviation

} ⇓ Spread ⇒ How the data is spread



① Variance { Bessels Correction, Degree of freedom }

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

(μ: Population mean)

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

($\bar{x}$: Sample mean; $n-1$: Bessel's correction)

Eg:

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

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71

$n = 6$

$$\frac{10.84}{5} = 2.168$$

($n-1$)

$\mu = 2.83$ {let consider as an example} 10.84

Variance ↑↑

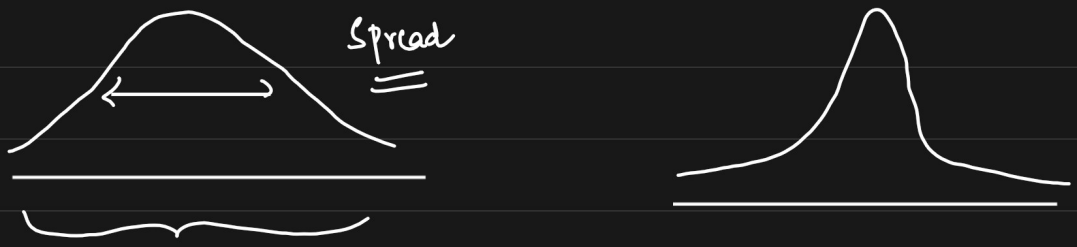
σ^2 ⇓ ✓

Variance = 6.42 Spread ↑↑

σ^2

Variance = 2.168

Spread ↑↑



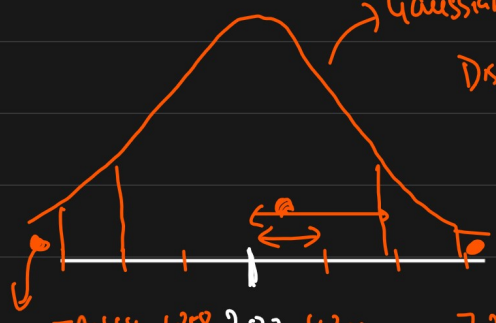
Standard deviation

$\sigma = \sqrt{\text{Variance}} = \sqrt{2.168} = \{1.472\}$

Variance ⇓

Spreadness ↓

1, 2, 2, 3, 4, 5



2.83 1 0
− 1.472
1.358

1.358
1.472
0.114

1 1
2.83
1.472
4.302
1.472
5.774
1.472
7.246

Percentiles And Quartiles

Percentages : 1, 2, 3, 4, 5

% of the numbers that are odd?

% of odd = $\frac{3}{5}$ = 60%

Percentiles : {CAT, GATE, SAT} ⇒ 99%

Defn : A percentile is a value below which a certain percentage of observation lie

99 percentile mean the person has got better marks than 99% of the Students.

Dataset : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

Average = 5

What is the percentile ranking of 10 ? n = 20

Percentile Rank of x = $\frac{\#\text{ of values below } x}{n} \times 100$

$= \frac{16}{20} \times 100 = 80$ percentile

$= \frac{17}{20} = 85$

② What value exists at percentile ranking of 25% ?

$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$ Doubt??

$= \frac{25}{100} \times (21) = 5.25 \rightarrow$ Index

Value = 5

Quartiles (25%)

Five Number Summary

1. Minimum
2. First Quartile (25%) Q1
3. Median
4. Third Quartile (75%) Q3
5. Maximum

Removing the outliers

Inter Quartile Range = (75% - 25%)
Q3 - Q1

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~29~~}

[Lower Fence ⟷ Higher Fence]

Lower Fence = Q1 - 1.5(IQR)

Higher Fence = Q3 + 1.5(IQR)

(25%) $Q1 = \frac{25}{100} \times (20) = 5^{th}$ index

Q1 = 3

IQR = Q3 - Q1 = 7 - 3 = 4

(75%) $Q3 = \frac{75}{100} \times 20 = 15^{th}$ index

Q3 = 7

Lower Fence = 3 - 1.5(4) = 3 - 6 = -3

Higher Fence = 7 + 1.5(4) = 7 + 6 = 13

[-3 ⟷ 13]

[-3 ⟷ 13]

-ve 3

Remaining

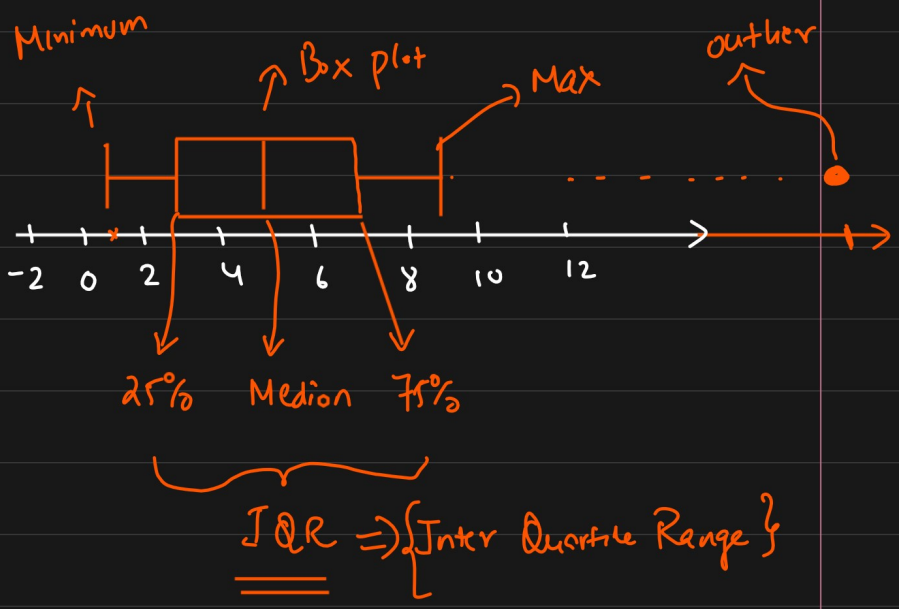
$\frac{5+5}{2} = 5$

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~20~~

5 Number Summary

Minimum = 1
Q1 = 3
Median = 5
Q3 = 7
Max = 9

Box plot → Outliers

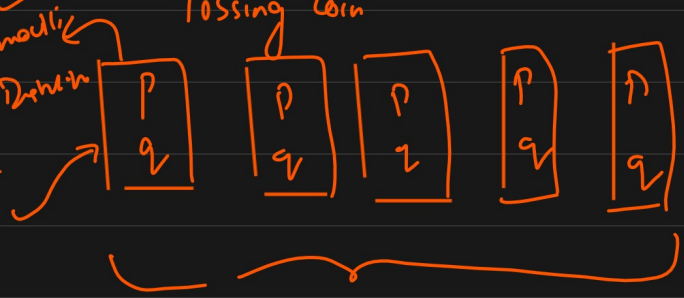


IQR ⇒ {Inter Quartile Range}

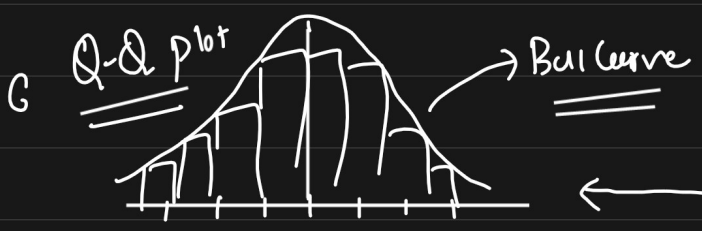
Use of Box plot

Distributions

1. Normal / Gaussian Distribution ✓
2. Standard Normal Distribution ✓
3. Z - Score ✓
4. Log Normal Distr ✓
5. Bernaullis Distribution ✓
6. Binomial Distribution }



1) Gaussian / Normal Distribution



Properties { Power Law }

1) Empirical Rule of Gaussian Distribution ⇓ 80-20%

↳ DATASET → IRIS Dataset } → Petal, Sepal length

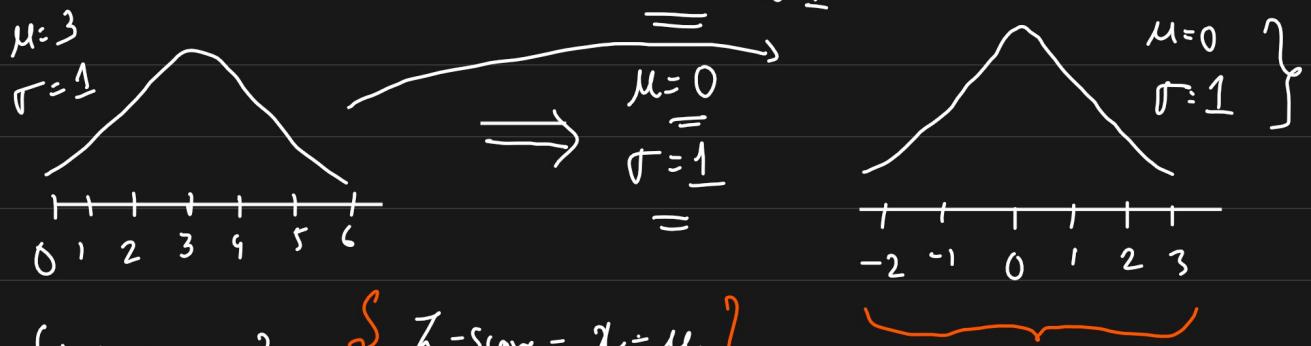
2) Weight of human beign 3) Height → Doctor ⇓ Domain expertise }

68.2, - 95.4 - 99.7 ↙ Outliers

Standard Normal Distribution

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

$\sigma = 1.414 \approx 1$



{1,2,3,4,5} $\left\{ Z\text{-score} = \frac{x - \mu}{\sigma} \right\}$

$\frac{2-3}{1}$ $\frac{3-3}{1} = 0$ $= \frac{1-3}{1}$

$= -2$

Why ?? $\mu = 0$, $\sigma = 1$

Standardization Vs Normalization

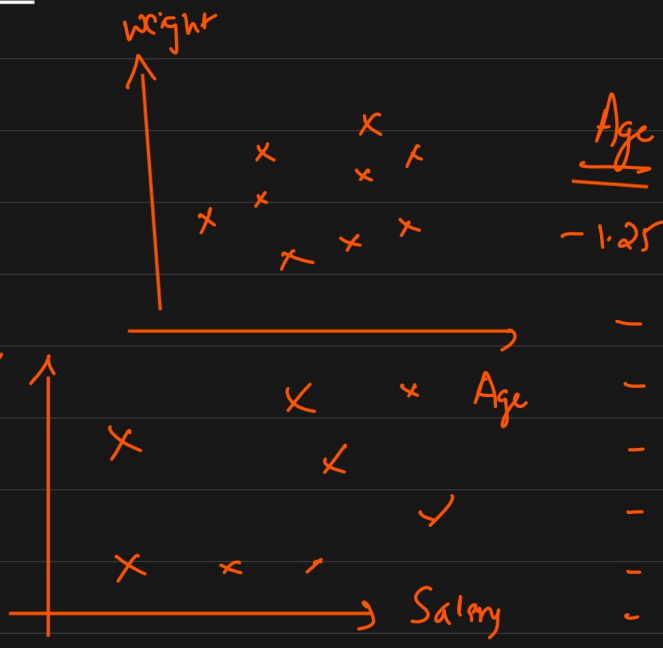
Age (years)	Weight (kg)	Salary (INR)
25	75	25K
26	80	30K
28	85	40K
30	60	80K
32	70	

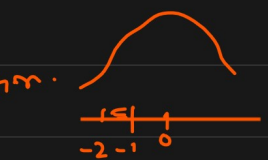
Different unit

Same unit scale ??

$\frac{25 - 28.2}{2 \text{ to } 6}$

Maths → Scale





ML Scaler

Normalization [Min Max Scaler] { Standardization }

&

0 to 4
0 to 1 } { Min Max Scaler }

Normalization

⇓

CNN ← Convolutional Neural N/w.

(0-1)

Normalization

$$\left\{ x_{Norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \right\}$$

⇓

Min Max Scaler = 0 to 1

f1
2
5
6
8 ✓
1 ✓

f1'
0.14
0.571
0.71
1
0

$\frac{8-1}{8-1} = 1$

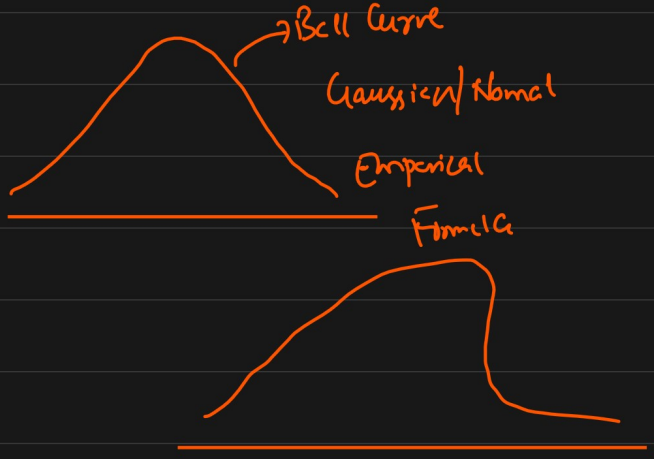
$\frac{1-1}{8-1} = 0$

$\frac{6-1}{8-1} = \frac{5}{7}$

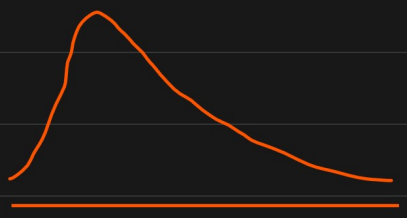
$= \frac{2-1}{8-1} = \frac{1}{7} = 0.142$

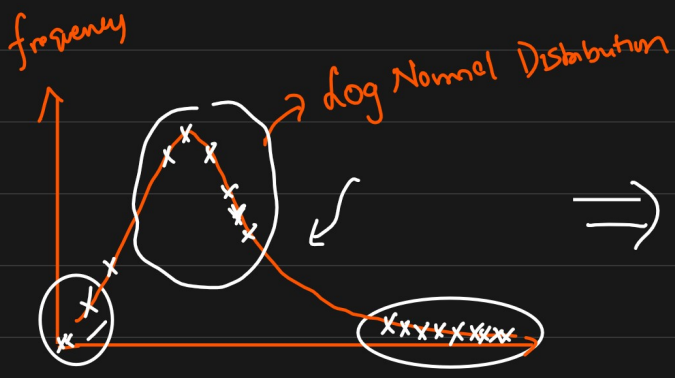
$\frac{5-1}{8-1} = \frac{4}{7} = 0.571$

Log Normal Distribution



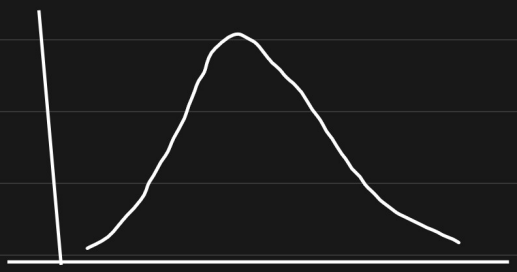
Skewed Curve

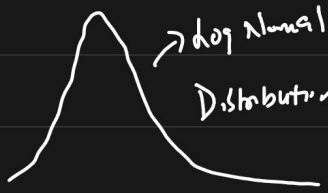




Gaussian Distribution

Normal Distribution



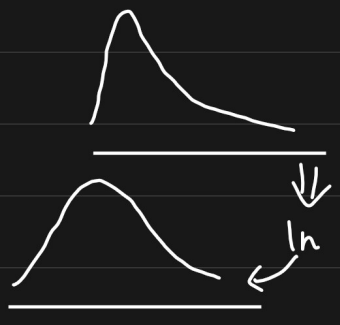


X = Log Normal Distributed

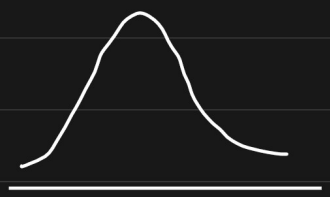
$\{Y = \ln(x)\}$ Gaussian Distribution

$\Downarrow \Downarrow \Downarrow$

$\{X = \exp(y)\} \rightarrow e^{y}$



X	$Y = \ln(x)$
25	=
30	=
40	=
35	=



(*) Bernoulli's Distribution.

Day 2 - Stats

① Z - Score $= \frac{x_i - \mu}{\sigma}$

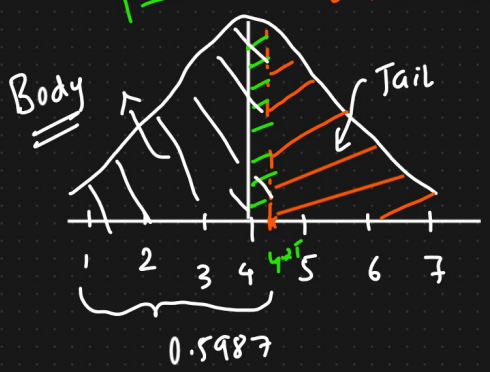
Stats Interview Question

{1, 2, 3, 4, 5, 6, 7}

$\mu = 4$

$\sigma = 1$

How many Standard deviation 4.25 fall from the mean??



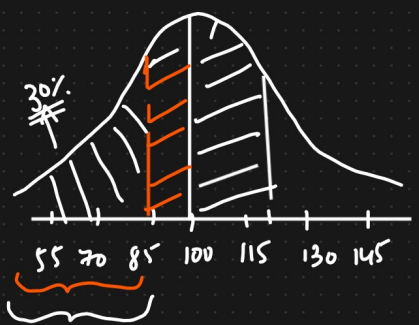
$$Z\text{-score} = \frac{x_i - \mu}{\sigma} = \frac{4.25 - 4}{1} = 0.25$$

Question : What percentage of scores fall above 4.25?

$1 - 0.59871 = 0.4013 \Rightarrow 40.13\%$

② In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have an IQ lower than 85?

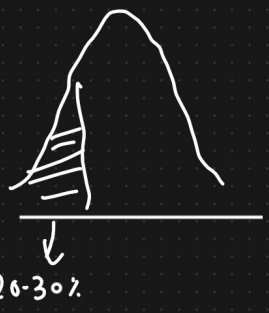
Ans)



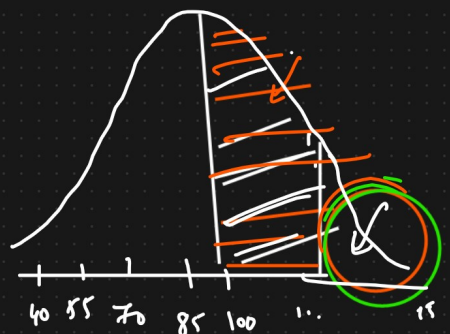
$$Z\text{-score} = \frac{85 - 100}{15} = \frac{-15}{15} = -1$$

① Area under this curve

$0.5 - 0.15866 = 0.34143 \Rightarrow 34.134\%$



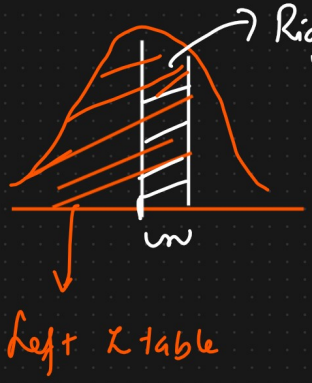
{ Greater than or equal 100 less than or equal 125 }



$$Z\text{score} = \frac{125 - 100}{15} = \frac{25}{15} = 1.66$$

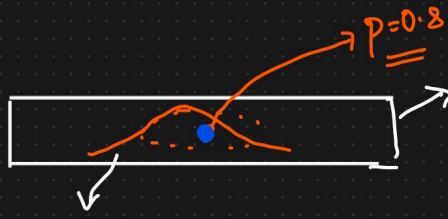
Ans = 0.4515 => 45.15%.

0.5 − 0.4515 = 0.0485 => 4.8%



P value, Hypothesis Testing, Confidence Interval

Out of all 100 touches, the no. of touches is 80



Out of all 100 touches, the no of times 40 times.

Hypothesis Testing, C.I, Significance value Together

Fair Coin

Coin ⟶ Test whether the coin is a fair coin or not by performing 100 tosses

P(H) = 0.5 ↓ P(T) = 0.5 ↓

Hypothesis Testing

Criminal is → Court SIMOLAY P(H) = 100% P(T) = 0%

1. Null Hypothesis — Coin is fair → (H_0)
2. Alternate Hypothesis — Coin is not fair → (H_1)
3. Experiments
4. Reject or Accept the Null Hypothesis

100 tosses →
- 50 times Head
- 50 times Tails

} Coin is fair

60 times Head
40 times Tail

} Coin is fair

30 times Head
70 times Tail

Domain Expertise

Confidence Interval, Significance Values

Health Care

↓

Covid vaccine Test

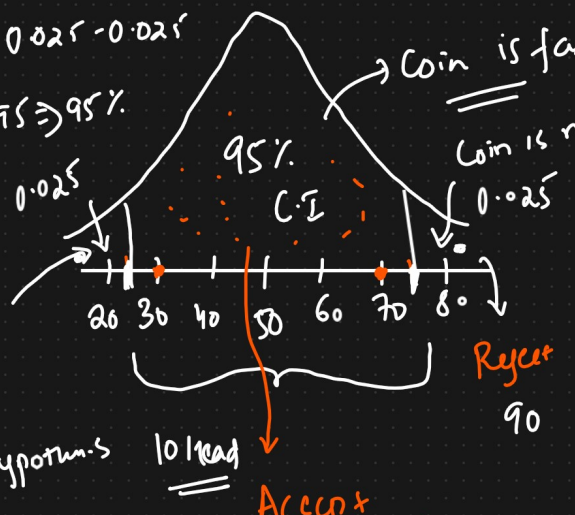
Error ↓↓

↓

Significance value = 0.05

C.I = 1 − 0.025 − 0.025

= 0.95 ⇒ 95%



Null Hypothesis

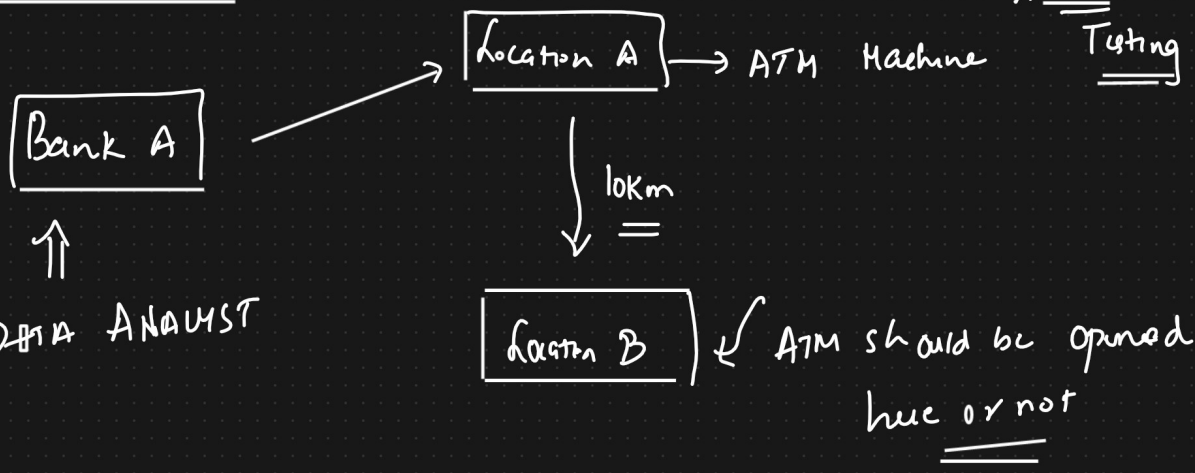
$\alpha = 0.45$

Medical $\alpha \uparrow\uparrow$.

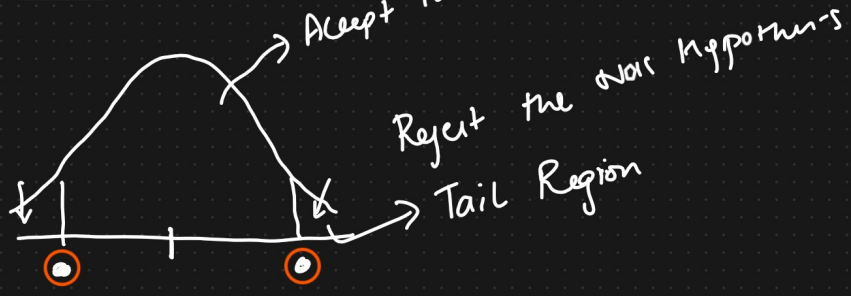
$\frac{0.45}{2} = 0.225$

Real World Project

Hypothesis Testing

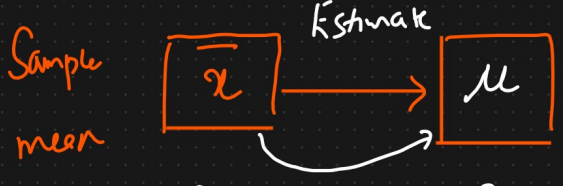


① Confidence Interval



Point Estimate

{ The value of any statistics that estimates the value of a parameter is called Point Estimate.



Influential

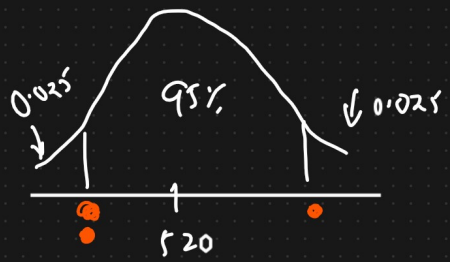
{ Sample Mean → Population }

Confidence Interval

t test Point Estimate ± Margin of Error ⇒ Population.

Q) On the quant test of CAT Exam, the (population) standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct a 80 (⇓) 95% C.I about the mean?

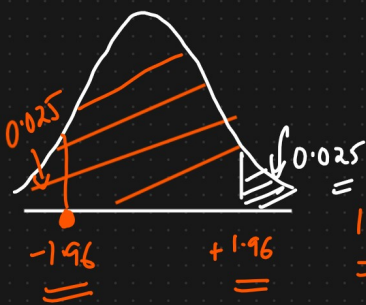
Ans) $\sigma = 100$ $n = 25$ $\bar{x} = 520$ C.I = 95% $\alpha = 0.05$ (1 − 0.95)



① Population Std is given {Z score} → Z table

Point Estimate ± Margin of Error → C.I

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$ → Standard Error



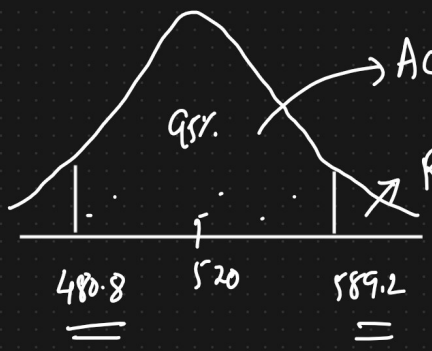
1 − 0.025 = 0.9750

Lower fence C.I = $\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ → $Z_{\frac{0.05}{2}}$ ⇒ $Z_{0.025} = 1.96$

Higher fence C.I = $\bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

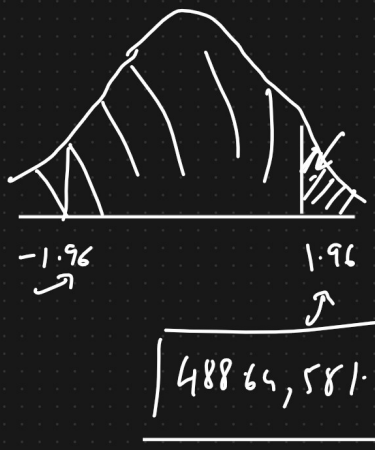
Lower fence = $520 - (1.96) * \frac{100}{\sqrt{25}} = 520 - (1.96) \times 20 = 480.8$

Higher fence = $520 + (1.96) \times 20 = 559.2$.



$$\frac{\alpha}{2} = 0.025$$

$$1 - 0.025 = 0.9750$$



488 to 552

(4) On the quant test of CAT exam, a sample of 25 test takers has a mean of 520 with a sample standard deviation of 80. Construct 95% C.I about the mean??

Ans) $\bar{x} = 520 \quad S = 80 \quad \alpha = 0.05 \quad n = 25$

t-test ⇒ t-table

{Because population sd is not given}

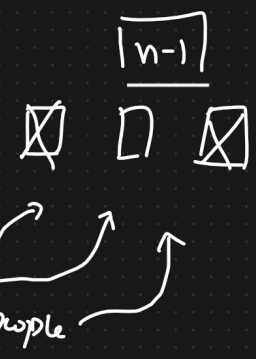
$$\bar{x} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \longrightarrow \text{Standard Error}$$

$t_{0.025}$

t-test

① Degree of freedom = $n - 1 = 25 - 1 = 24$

3 | n-1



$$\bar{x} \pm 2.064\left(\frac{80}{5}\right) \Rightarrow 486.976 \leftrightarrow 553.024$$

① Type 1 and Type 2 Error.
② One Tailed vs 2 Tailed Test

Type 1 and Type 2 Error

$H_0 \Rightarrow$ Coin is Fair
$H_1 \Rightarrow$ Coin is not Fair

Outcome 1:

We reject the Null Hypothesis in reality if it is false ⟶ Yes ✓

Outcome 2:

We reject the Null Hypothesis when in reality it is true ⇒ No ⇒ Type 1 Error ✗

Outcome 3

We accept the Null Hypothesis, when in reality it is false ⇒ Type 2 Error ✗

Outcome 4: We accept the Null Hypothesis when in reality it is True ✓

[Cancer → Not cancer]
↓ Fire

Reality Check

① Null Hypothesis is True or Null Hypothesis is False

Decision of Experiments?

Null Hypothesis is True or False.

Null Hypothesis
$H_0 \rightarrow$ The Criminial is not guilty
$H_1 \rightarrow$ " " is guilty

{→ Stock market is going to crash?

Confusion Matrix

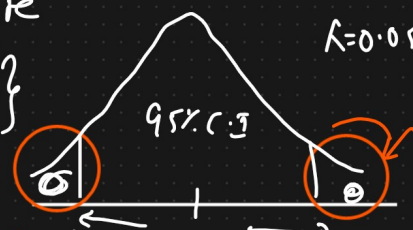
② 1 Tail and 2 Tail Test

Eg: Colleges is Karnataka has an 85% placement rate. A new college was recently opened and it was found that a sample of 150 students had a placement rate of 88%. with a standard deviation of 4%. Does this College has a diffucnt placement rate?

$\alpha = 0.05 = 95\%$ C.I

↳ 85%

{ placement rate less than 85% } ⇓ 1 tail



{ placement rate greater than 85% } 2 Tail Test

1 tail Test

Saturday

① Z test Hypothesis Testing
② T Test Hypothesis Testing
③ Significance value & P value.
④ ANOVA TEST
⑤ CHI SQUARE TEST
⑥ Practical

10 min probability

Sunday

EDA → 3-4 projects
FE →

→ Machine Learning

① Central Limit Theorem

② Infcuential Statistics

*) Z test {Z table} [5-6 problems]

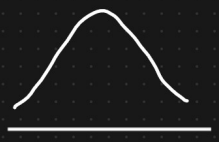
*) t test {t table}

*) Z test propotion population.

*) Chi square (categorical Tcst)

*) ANNOVA (F Test)

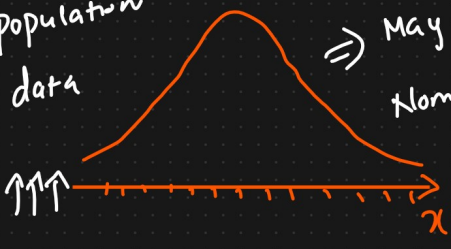
Infcuntial



① Central limit Theorem

n>30 ✓ Sample mean distribution

{ Population data ⇒ May be Gaussian Normal Distr



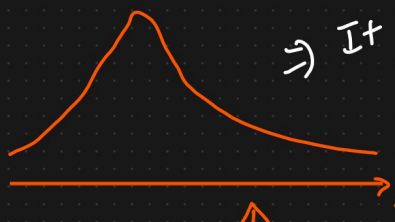
$\rightarrow [x_1, x_2, x_3, x_4, x_5 \cdots x_{30}] \rightarrow \bar{x}_1$ Sample 1

Sample 2 $[x_1, x_2, x_3\ x_4, x_5 \cdots x_{30}] \rightarrow \bar{x}_2$

$\rightarrow \bar{x}_3$

$\rightarrow \bar{x}_m$

⇒ It may not



Sample m

10,

n>30

Sample mean distribution

Gaussian Distribution

Normal Distribution

② Influential Statistics {Data Analyst, Data Scientist}

① 100K ⇒ T-shirt ⇒ No ⇒ Sample data ⇒ XL, L, Small

② iNeuron → Meetup → Hitesh ⇒ 300-400 people → T-shirts

{ 20% L, 10% XL, 10% XXL, 60% Medium } 500 t-shirts

Ordered: L, M, XL, XXL

Next Event : Z test $\bar{x}$ $(\mu, 5)$ → C.I [ ]

③ ATM ④ Measure the size of entire sharks

⑤ Amazon delivery { Percentile, Quartiles } ⇒

⑥ Hypothesis Testing

① A factory has a machine that fills 80ml of baby medicine in a bottle. An employee believes the average amount of baby medicine is not 80ml. Using 40 samples, he measures the average amount dispersed by the machine to be 78ml with a standard deviation of 2.5

(a) State Null and Alternate Hypothesis

(b) At a 95% C.I, is there enough evidence to support machine is not working properly.

Ans) Step 1

$n = 40$ $\bar{x} = 78$ $S = 2.5$

$H_0 : \mu = 80$ {Null Hypothesis}

$H_1 : \mu \neq 80$ {Alternate hypothesis}

Why Z test?

Step 2:

$\alpha = 0.05$ C.I = 95%

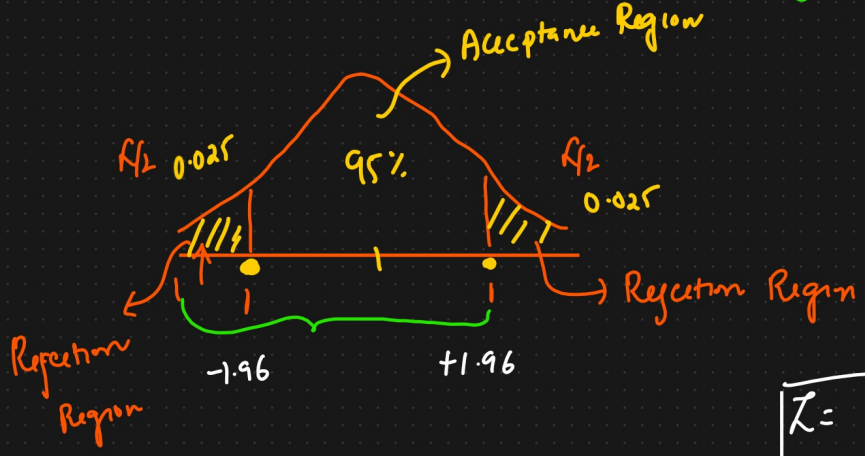
$n < 30$ { ① $n > 30$ $n \leq 30$ ② population std or sample std }

Step 3: Decision Boundary

$1 - 0.025 = 0.9750$

Why t test

{ ① Sample std = ② $n < 30$ }



$n = 1$ }

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

④ Calculate Test Statistics

$$Z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$s/\sqrt{n}$ ⇒ Standard Error

Sample standard deviation

$$= \frac{78 - 80}{2.5 / \sqrt{40}} = \frac{-2 \times \sqrt{40}}{2.5} = \frac{-2}{2.5} \times 6.32 = -5.05$$

⑤ State the Results

Decision Rule: If $Z = -5.05$ is less than -1.96 or greater than 1.96, then reject the null hypothesis with 95% C.I

Reject Ho Null hypothesis {There is some fault in the machine}

(2) In the population the average IQ is 100 with a standard deviation of 15. A team of scientists wants to test a new medication to see if it has a +ve or -ve effect, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect Intelligence? { 95% ⇓ C.I }

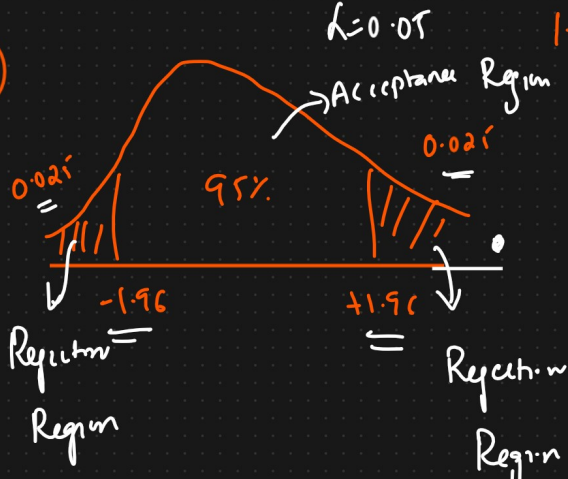
Ans) $\sigma = 15 \quad n = 30 \quad \bar{x} = 140$

(1) $H_0 : \mu = 100$

$H_1 : \mu \neq 100$

(2) $\alpha = 0.05 \quad C.I = 95\%$

(3)



If Z is less than -1.96 or greater than 1.96, reject the NULL Hypothesis

(4) $$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{140 - 100}{15/\sqrt{30}} = 14.60$$

$14.60 > 1.96$ Reject the NULL Hypothesis

(*) A Complain was registered, the boys in the Municipal Primary School are underfed. Average weight of boys of age 10 is 32kgs with S.D = 9kgs. A sample of 25 boys was selected from the municipal school and the average weight was found to be 29.5 kgs? With C.I = 95% Check whether it is True or False?

Ans) $\mu = 32\text{kgs}$ $\sigma = 9\text{kg}$ $n = 25$ $\bar{x} = 29.5$ $\alpha = 0.05$

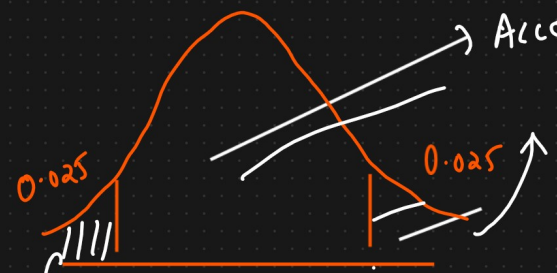
1) $H_0 = \mu = 32$ }

$H_1 = \mu < 32$

② $\alpha = 0.05$

$1 - 0.95 = 0.05$

③



④ $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

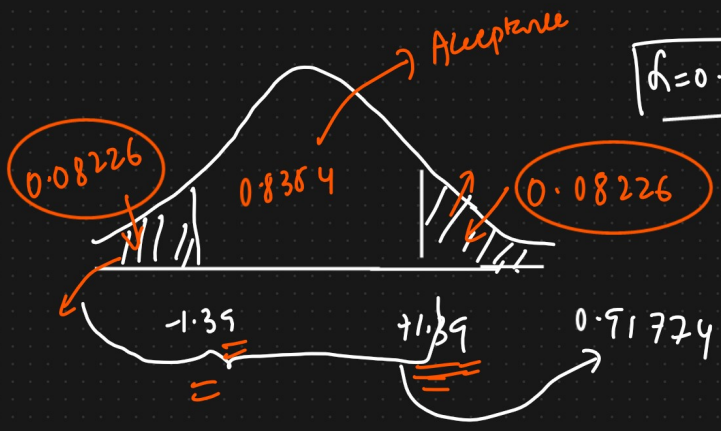
$= \dfrac{29.5 - 32}{9/\sqrt{25}} = -1.39$

Conclusion : $-1.39 > -1.96$ therefore we accept the Null Hypothesis

So, the boys are not underfed.

Significance value. ⇒

P value



$\alpha = 0.05$ ⇒ p value

$1 - 0.08226 - 0.08226 =$

$1 - 0.91774 = 0.08226$

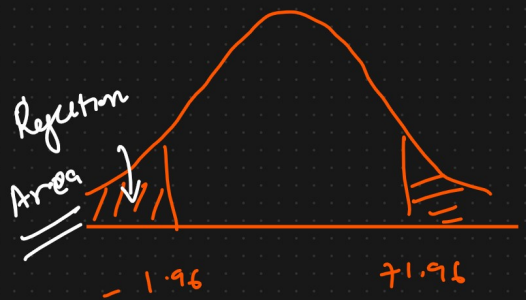
p value

$0.08226 + 0.08226 = 0.1645$

Significance value

$0.1645 > 0.05$

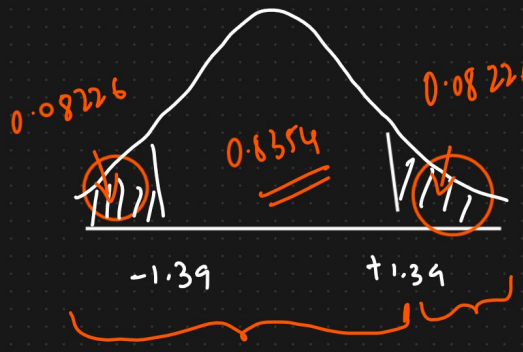
$p \geq 0.05 \Rightarrow$ Accept the Null Hypothesis



Z test

⇓

p value



New p value $= 0.08226 + 0.08226$

$= 0.16$

$1 - 0.08226 - 0.08226$

Domain

⇓

$0.1645 >$ Significance value

⇓

Accept the Null Hypothesis.

$1 - 0.91774 = 0.08226$

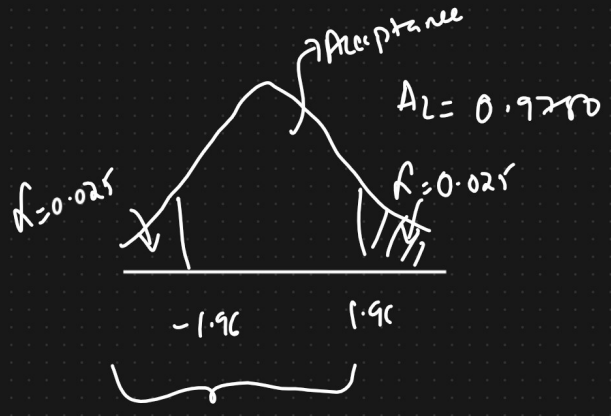
(#) The average weight of all residents in town XYZ is 168 lbs. A nutritionist believes the true mean to be different. She measured the weight of 36 individuals and found the mean to be 169.5 lbs with a standard deviation of 3.9.

(#) Assignment

(a) At 95% C.I is there enough evidence to discard the Null Hypothesis??

Ans) $H_0 : \mu = 168$ $\quad n = 36$ $\quad \bar{x} = 169.5$ $\quad s = 3.9$

$H_1 : \mu \neq 168$ $\quad C = 0.95$ $\quad \alpha = 1 - C.I = 0.05$



Z-test , t-test

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = 2.31$$

$2.31 > 1.96$ Reject the Null Hypothesis

(#) A company manufactures bike batteries with an average life span of 2 or more years. An Engineer believes this value to be less. Using 10 samples, he measures the average life span to be 1.8 years. with a standard deviation of 0.15.

a) State the Null and Alternate Hypothesis

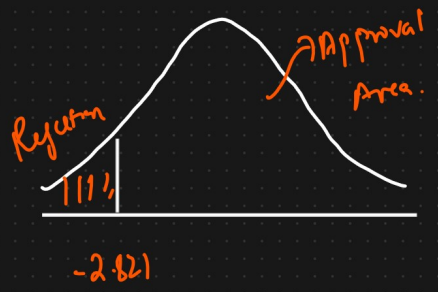
b) At a 99% C.I, is there enough evidence to discard the H_0?

Ans) $H_0 : \mu \geq 2$ $\quad n = 10$ $\quad \bar{x} = 1.8$ $\quad s = 0.15$ {Sample std is given}

$H_1 : \mu < 2$ $\quad n < 30$ $\quad$ t-test ??

② $\alpha = 0.01$ $\quad \alpha = 1 - C.I = 1 - 0.99 = 0.01$

③



Degree of freedom: $n - 1 = 9$

④ Calculate t-test statistics :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.8 - 2}{0.15/\sqrt{10}} = \frac{-0.2}{0.15/3.1622} = -4.216$$

⑤ Conclusion

$-4.216 < -2.821$ $\quad$ Reject the Null Hypothesis.

Z test with proportions

⑥ A tech company believes that the percentage of residents in town XYZ that owns a cell phone is 70%. A marketing manager believes that this value to be different. He conducts a survey of 200 individuals and found that 130 responded yes to

Owning a cell phone

(a) State the Null and Alternate Hypothesis?

(b) At a 95% C.I, is there enough evidence to reject the Null Hypothesis?

$n = 200$ $\quad x = 130$

(a) $H_0 \quad p_0 = 0.70.$

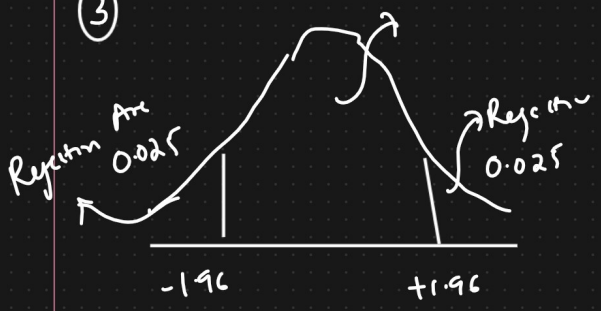
$H_1 \quad p_0 \neq 0.70$

$$\hat{p} = \frac{x}{n} = \frac{130}{200} = \frac{13}{20} = 0.65$$

$q_0 = 1 - p_0$

② $\alpha = 0.05 \quad C.I = 95\%$

③

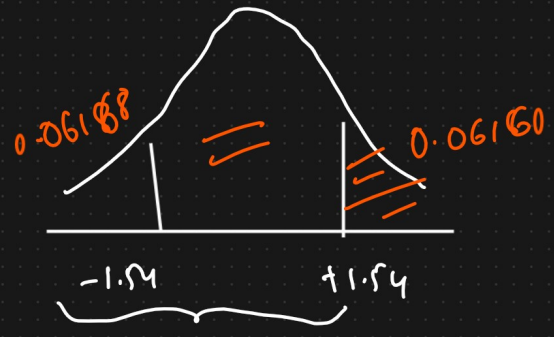


$$Z_{test} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$$= \frac{0.65 - 0.70}{\sqrt{\frac{0.7 \times 0.3}{200}}} \approx -1.54$$

At 95% C.I there is

$-1.54 > -1.96$, So we accept the Null Hypothesis



pvalue

$2 \times 0.06168 > 0.05$

Accept Null Hypothesis.

$1 - 0.93832 = 0.06168$

(4) A Car company believes that the percentage of residents in City ABC that owns a vehicle is 60% or less. A Sales manager disagrees with this. He conducts a hypothesis testing surveying 250 residents and found that 170 responded yes. to owning a vehicle.

(a) State the Null & Alternate Hypothesis

(b) At 10% significance level, is there enough evidence to support the idea that vehicle ownership in city ABC is 60% or less?

pvalue = .0.014

Statistics

{ 11:30 - 12pm }

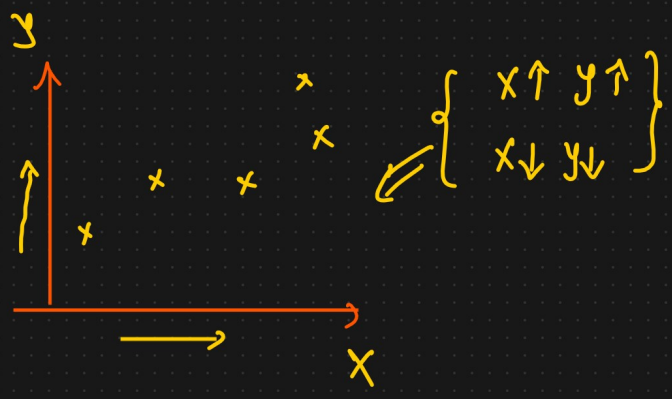
1. Covariance
2. Pearson Correlation Coefficient
3. Spearman Rank Correlation Coefficient.
4. CHI SQUARE TEST
5. ANNOVA (F-Test)

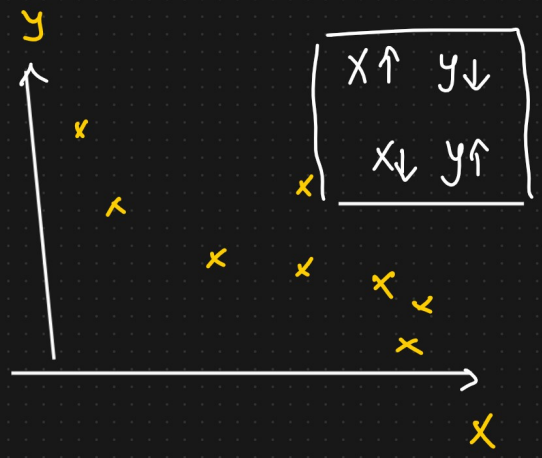
} practicals

Covariance

{ quantify the relationship between X & Y }

X	Y
X↑	Y↑
X↑	Y↓
X↓	Y↑
X↓	Y↓





$$Cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1} \Leftrightarrow Var(X) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$\Uparrow$ Cov(X, Y) $\Downarrow$

$$Cov(x,x) = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$= \frac{\sum (x_i - \bar{x}) \times (x_i - \bar{x})}{N-1}$$

$$Var(x) = \sum_{i=1}^{n} \frac{(x - \bar{x})^2}{n-1} \Rightarrow \sum_{i=1}^{n} \frac{(x-\bar{x}) * (x-\bar{x})}{n-1}$$

$$\Downarrow$$

$$Cov(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x}) * (x-\bar{x})}{n-1}$$

+ve

X ↑	Y ↑
X ↓	Y ↓

$\Rightarrow$ Positively Correlation

X	Y
2	3
4	5
6	7
$\bar{x} = 4$	$\bar{y} = 5$

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1}$$

$$= \frac{(2-4)(3-5) + (4-4)(5-5) + (6-4) * (7-5)}{2}$$

$$= \frac{(-2)(-2) + 0 + (2) \times (2)}{2} = \frac{8}{2} = 4$$

X ↑	Y ↓
X ↓	Y ↑

$\Rightarrow$ -ve Correlation $\Rightarrow$ -ve value.

Disadvantage Covariance

$Cov(x,y)$ $\Rightarrow$ +ve value or -ve value $\Rightarrow$

$Cov(x,y) = 500$

$Cov(y,z) = 600$

Limit

+500 -300

-400 +1000

-400

∞

$$\Downarrow$$

Relationship [-1 to 1]

$$\Uparrow$$

② Pearson Correlation Coefficient

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

[-1 to 1]

The more the value towards 1 more +ve it is correlated

Dataset : 1000 features

Independent features

X | ~~X~~ | Z | A | B | C → O/P (Dependent feature)

+ve correlated

X, Y ⇒ 99%

90% 0.9

-ve Correlation ⇓ Keep it

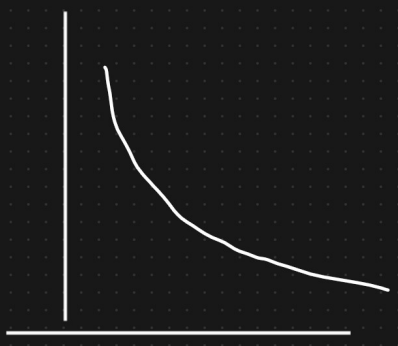
⑧ Spearman Rank Correlation

$$r_s = \frac{Cov(R(x), R(y))}{\sigma_{R(x)} \sigma_{R(y)}}$$

Marks

X	Y	R_x	R_y
1	2	4	4
3	4	3	3
7	5	2	2
0	7	5	1
8	1	1	5

Spearman Rank Corr = 1



-1

⑦ Chi Square

The Chi Square Test Claims about population proportions.

It is a non parametric test that is performed on categorical (nominal or ordinal) data.

Ex) In the 2000 U.S Census, the ages of individuals in a small town were found to be the following.

<18	18-35	>35
20%	30%	50%

In 2010, ages of n=500 individuals were sampled. Below are the results

<18	18-35	>35
121	288	91

Using $\alpha = 0.05$, would you conclude the the population distribution of ages has changed in the last 10 years?

Ans)

	<18	18-35	>35
Expected	20%	30%	50%

95% C.I

n=500

Observed	121	288	91
Expected	100	150	250

① H_0 = the data meets the expected distribution

H_1 = the data donot meet the expected distr

② State Alpha : $\alpha = 0.05$

③ Calculate the degree of freedom

$df = n - 1 = 3 - 1 = 2$ ⇒ 3 Categories.

④ Decision

Chi Square Table.

If χ^2 is greater ⇓ 5.99 than, Reject H_0

⑤ Calculate Chi square Test

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe} = \frac{(121-100)^2}{100} + \frac{(288-150)^2}{150} + \frac{(91-250)^2}{250}$$

$$\chi^2 = 232.494$$

$232.494 > 5.991$ Reject the Null Hypothesis.

⑥ A school principal would like to know which days of the week students are most likely to be absent. The principal expect the students will be absent equally during the 5-day school week. The principal selects a random sample of 100 teachers asking them which day of the week they had the highest number of

Student absenses. The Observed and expected results are shown in the table below. Based on these results, do the days for the highest number of absences occurs with equal frequencies (use. 95% C.I)

	Monday	Tuesday	Wednes	Thursday	FRIDAY
Observed	23	16	14	19	28
Expected	20	20	20	20	20.

Ans = 6.3 {Accept the Null Hypothesis}

{Practicals + EDA + Feature Engineering}

Statistics

① ANOVA (F-Test) → 1 hour }
② KDA → { Solve Some Examples. }

ANOVA : { Analysis of Variance }

ANOVA Is a statistical method used to compare the means of 2 or more group

ANOVA :

① Factors (variables) Medicine
② Levels { Dosage }

Anxiety reducing { Gender }

0mg	50mg	100mg	M	F
9	7	4		
8	6	3		
7	6	2		
8	7	3		
6	8			

factor : Dosage

Levels : 0mg, 50mg, 100mg

Types of ANOVA

One Way ANOVA : One factor with atleast 2 levels, levels are independent.

② Repeated Measures ANOVA – One factor with atleast 2 levels, but levels are { dependent }

Factor Levels	Running kms	KARTIK Day 1	Day 2	Day 3	Day 4
1		6	10	2	10 kms
2		–	–	–	–

Eg. Study hours of KARTIK

1 Factor: Day 1 ⟷ Day 2 ⟷ Day 3

10 hrs | 5 hrs | 2 hrs

Gym

Day 1 ⟷ Day 2 ⟷ Day 3

2 hrs (5 Eggs) | 1.5 hr (2 Eggs) | 1 hr

④ Factorial ANOVA

⑤ Factorial ANOVA : Two or more factor (each of which with atleast 2 levels), levels can be either independent, dependent or both (mixed)

Eg

factor ↓	Day 1	Day 2	Day 3
Men	9	7	4
	8	6	3
	7	5	2
Women	8	7	3
	8	8	4
	9	7	3

(factor → Day 1, Day 2, Day 3)

One Way ANOVA (F-test) ⇒ Inferential Stats

⇓

Comparing means of 2 or more groups

1) Researchers want to test a new anxiety medication. They split participants into 3 conditions (0mg, 50mg, 100mg), then ask them to rate their anxiety level on scale of 1-10. Are there any differences between the 3 conditions using $\alpha = 0.05$?

0mg	50mg	100mg
9	7	4
8	6	3
7	6	2
8	7	3
8	8	4
9	7	3
8	6	2

① $H_0 = \mu_{0mg} = \mu_{50mg} = \mu_{100mg}$

H_1 = not all μ's are equal

② State α and C.I

$\alpha = 0.05$ C.I = 95%

③ Calculate the Degree of freedom

Statistics

$N = 21$ ⇓ $n = 7$

$a = 3$ → {No. of levels}

→ $df_{Between} = a - 1 = 3 - 1 = 2$

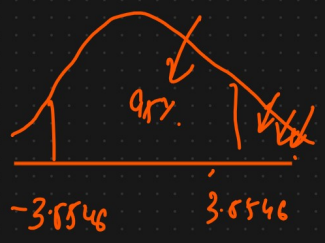
→ $df_{within} = N - a = 21 - 3 = 18$

→ $df_{Total} = N - 1 = 21 - 1 = 20$

④ State Decision Rule

$df_{Between} = a - 1 = 3 - 1 = 2$ $\{(2, 18)\}$

$df_{within} = N - a = 21 - 3 = 18$



If F test is greater than 3.5546, Reject the Null Hypothesis

If F test is less than -3.5546 " " " "

⑤ Calculate F Test Statistics

$F_{Test} = \frac{MS_{between}}{MS_{within}} = \frac{49.34}{0.57} =$

	SS	df	MS	F Test
Between	98.67 →	2	49.34	86.56
Within	10.29	18	0.57	
Total	108.96	20		

$N = 21 \quad n = 7$

$SS_{between} = \left[\frac{\sum(\sum a_i)^2}{n}\right] - \frac{T^2}{N}$

$T^2 = [57 + 47 + 21]^2$

$= (125)^2$

$\sum(\sum a_i)^2 = (9+8+7+8+8+9+8)^2 + (7+6+6+7+8+7+6)^2 + (4+3+2+3+4+3+2)^2$

$= 57^2 + 47^2 + 21^2$

$SS_{Between} = \frac{57^2 + 47^2 + 21^2}{7} - \frac{125^2}{21} = 98.67$

② $SS_{within} = \sum y^2 - \frac{\sum(\sum a_i)^2}{n}$

$p\ 0.78$

$\alpha = 0.05$

$= \sum y^2 - \left[\frac{57^2 + 47^2 + 21^2}{7}\right] = 10.29$

$\sum y^2 \downarrow 853$

$\sum y^2 = 9^2 + 8^2 + 7^2 + 8^2 + 8^2 + 9^2 + \ldots + 2^2 = 853$

$0.78 > 0.05$

⇓

Accept

Final Conclusion

$86.56 > 3.5546$ so we reject the Null hypothesis

$H_0 = \mu = \text{Some value}$

$H_1 : \mu \neq \text{Some value}$ → 95% C.I

Virginica

Petal width

→ $H_0 = \mu_{virgin} = \mu_{stressed} = \mu_{\dots}$

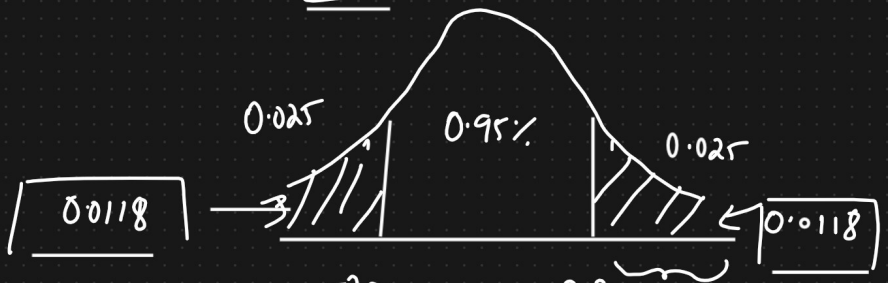
$H_1 =$ $\neq$ p value $\neq$ ⇒ Reject the Null Hypothesis

0.0118
0.0118
0.0228

0.0228 < 0.05 $\quad 1 - 0.025 = 0.975$

Z test $\quad \alpha =$

Z test statistics



$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$Z = 2.304$

$2.304 > 1.96$ Reject

Stats Interviews Questions

1. Question: What is the Central Limit Theorem and why is it important?

Answer: The Central Limit Theorem (CLT) states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal (or Gaussian) distribution, regardless of the original distribution of the variables. It's crucial in statistics because it allows us to make inferences about populations using the normal distribution, which has well-understood properties.

---

2. Question: Explain Type I and Type II errors.

Answer:

- **Type I Error (False Positive, or Alpha error):** It's when you incorrectly reject a true null hypothesis.
- **Type II Error (False Negative, or Beta error):** It's when you fail to reject a false null hypothesis. The significance level (usually denoted by α) is the probability of making a Type I error. The power of a test is 1 minus the probability of making a Type II error (β).

---

3. Question: What is R-squared in linear regression?

Answer: R-squared, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. An R-squared value of 1 indicates that the regression predictions perfectly fit the data. Values of R-squared close to 1 indicate a large proportion of the variance in the dependent variable is explained by the regression model, while values close to 0 indicate the opposite.

---

4. Question: What is the difference between correlation and causation?

Answer: Correlation indicates a mutual relationship or association between two variables. When one variable changes, there's a tendency for the other variable to change in a specific direction. However, correlation does not imply causation. Causation means that a change in one variable is responsible for a change in another. For example, even if there's a strong correlation between ice cream sales and the number of drowning incidents, this doesn't mean buying more ice cream causes more drownings. A lurking variable, like temperature, can be influencing both.

---

5. Question: What is the difference between a parametric and a non-parametric test?

Answer: Parametric tests make assumptions about the parameters of the population distribution from which the sample is drawn, such as assuming that the population has a normal distribution. Examples include t-tests and ANOVA. Non-parametric tests, on the other hand, do not make strong assumptions about the population's distribution. Examples include the Mann-Whitney U test and Kruskal-Wallis test.

---

6. Question: Explain p-value.

Answer: The p-value is a measure used to help determine the significance of the results in hypothesis testing. It represents the probability of observing the current data, or something more extreme, given that the null hypothesis is true. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so we reject it. A larger p-value suggests weaker evidence against the null hypothesis, so we fail to reject it.

---

7. Question: Describe the difference between cross-validation and bootstrapping.

Answer: Cross-validation is a technique for evaluating the performance of a statistical model by partitioning the original sample into a training set to train the model, and a test set to evaluate it. One common method is k-fold cross-validation. Bootstrapping, on the other hand, is a resampling technique used to estimate the distribution of a statistic (like the mean or variance) by sampling with replacement from the data. It helps assess the variability of a sample statistic and construct confidence intervals.

These are just a few potential questions. Depending on the role, interviewers might go deeper into specific topics or might also incorporate more practical, hands-on problems.

Some More Difficult Questions

1. Question: Can you explain the different measures of central tendency?

Answer: The three main measures of central tendency are the mean, median, and mode:

- **Mean:** It is the average of all the numbers in a dataset.
- **Median:** It is the middle value in a dataset when the numbers are arranged in order.
- **Mode:** It is the number that appears most frequently in a dataset.

2. Question: What is the difference between population and sample?

Answer: A population includes all members of a specified group, while a sample is a subset of the population. Statistics calculated on a population are called parameters, while those calculated on a sample are called statistics.

3. Question: How do you handle missing data?

Answer: Handling missing data can involve various techniques:

- **Deletion:** Remove records with missing values.
- **Imputation:** Fill missing values with estimated ones, e.g., using the mean, median, or mode of the known values, or using more complex algorithms or models to predict the missing value.
- **Analysis:** Use statistical techniques designed to handle missing values, such as multiple imputation or full information maximum likelihood estimation.

4. Question: What is the interquartile range (IQR) and why is it useful?

Answer: The IQR is a measure of statistical dispersion and is calculated as the difference between the upper (Q3) and lower (Q1) quartiles in a dataset. It is useful for understanding the spread of the data and for identifying outliers, as it is not affected by extremely large or small values.

5. Question: Explain the concept of skewness in statistics.

Answer: Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable. A negative skew indicates that the left tail of the distribution is longer, while a positive skew indicates that the right tail is longer. A skewness of zero indicates a perfectly symmetrical distribution.

6. Question: Can you describe what a box plot represents?

Answer: A box plot, or box-and-whisker plot, visually displays the distribution of a dataset, including its central tendency and variability. The box represents the interquartile range (IQR, Q3-Q1), the line inside the box shows the median, and the whiskers extend to the smallest and largest observations in the dataset.

7. Question: What is the difference between variance and standard deviation?

Answer: Variance and standard deviation are both measures of dispersion or spread in a dataset. Variance is the average of the squared differences from the mean, while the standard deviation is the square root of the variance. The standard deviation is more commonly used because it is in the same units as the data.

8. Question: What is a z-score and what is it used for?

Answer: A z-score is a statistical measurement that describes a value's relation to the mean of a group of values. It is measured in terms of standard deviations from the mean. A z-score is used to determine how unusual a value is, and it's commonly used for hypothesis testing, outlier detection, and comparison of scores from different datasets.

9. Question: Can you explain what covariance and correlation are?

Answer:

- **Covariance:** It is a measure of the joint variability of two random variables. A positive covariance indicates that the variables tend to increase and decrease together, whereas a negative covariance indicates that as one variable increases, the other tends to decrease.
- **Correlation:** It is the normalization of covariance to have values between -1 and 1, providing a measure of the strength and direction of the linear relationship between the two variables. A correlation of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear correlation.

These questions can help interviewers evaluate a candidate's understanding and knowledge of descriptive statistics concepts.

Lets Increase The Complexity

1. Question: How does the presence of outliers affect the mean and median of a dataset?

Answer: Outliers can greatly affect the mean because the mean considers all values in its calculation. An extreme outlier can pull the mean up or down, making it less representative of the central location of the data. The median, however, is more resistant to outliers since it depends only on the middle value(s) of an ordered dataset. In datasets with outliers, the median can often be a better representation of central tendency.

2. Question: Describe the concept of kurtosis. How is it different from skewness?

Answer: Kurtosis measures the "tailedness" of a probability distribution. High kurtosis indicates a distribution with tails heavier or more extreme than the normal distribution, and low kurtosis indicates a distribution with tails lighter than the normal distribution. While skewness deals with the asymmetry and direction of skew (left or right), kurtosis deals with the extremities (or outliers) in the distribution tails.

3. Question: How do you interpret the value of a Pearson correlation coefficient?

Answer: The Pearson correlation coefficient, often denoted as rr, measures the strength and direction of a linear relationship between two variables. Its values range between -1 and 1.

- $r = 1$r = 1: Perfect positive linear relationship.
- $r = -1$r = −1: Perfect negative linear relationship.
- $r = 0$r = 0: No linear correlation. The closer rr is to 1 or -1, the stronger the linear relationship. However, a strong correlation does not imply causation.

4. Question: Explain Simpson's Paradox and its implications in descriptive statistics.

Answer: Simpson's Paradox occurs when a trend or relationship between two variables reverses or disappears when they are examined in the context of a third variable. This can happen due to confounding factors. It emphasizes the importance of considering all relevant factors when interpreting statistical relationships.

5. Question: In a given dataset, what are the differences and relationships between the range, variance, and standard deviation?

Answer:

- **Range:** The difference between the maximum and minimum values in the dataset.
- **Variance:** The average of the squared differences from the mean.
- **Standard Deviation:** The square root of the variance.

The range provides a sense of the full spread of the data but is sensitive to outliers. The variance gives a measure of how data points differ from the mean, but it's in squared units of the data. Standard deviation, being the square root of variance, gives dispersion in the original units of the data and is commonly used because of this.

6. Question: How would you decide between using the mean vs. median as a measure of central tendency?

Answer: The decision often depends on the shape of the data distribution and the presence of outliers:

- For a symmetric distribution without outliers, the mean and median will be close, and either could be used.
- For skewed distributions or distributions with outliers, the median is usually a better representation because it is less affected by extreme values.

7. Question: Why might standard deviation be a misleading measure of spread in some situations?

Answer: Standard deviation can be misleading, especially when the data contains outliers, since it considers all deviations from the mean in its calculation. Extreme values can inflate the standard deviation, making it seem as though the data is more spread out than it actually is. In such cases, other measures like the interquartile range might be more appropriate.

Lets Try Some Usecases On Descriptive Stats

Usecase: A company sells products in three regions: North, South, and West. The sales team wants to understand the sales performance across these regions to allocate resources more efficiently.

Dataset:

```
java
| Region | Monthly Sales (in thousands) |
|--------|------------------------------|
| North  | 12, 15, 14, 13, 17, 19, 20  |
| South  | 22, 21, 20, 23, 25, 26, 28  |
| West   | 32, 30, 31, 29, 30, 33, 35  |
```

Question 1: Which region has the highest average monthly sales?

Process:

1. Calculate the mean (average) for each region.

Formula for Mean: $\mu = \frac{\sum x}{n}$μ = n∑ x

- μμ = mean
- $\sum x$∑ x = sum of all observations
- nn = number of observations

2. Compare the means to determine the region with the highest average sales.

Answer:

- **North:** $\mu = \frac{12 + 15 + 14 + 13 + 17 + 19 + 20}{7} = 15.7$μ = 712+15+14+13+17+19+20 = 15.7
- **South:** $\mu = \frac{22 + 21 + 20 + 23 + 25 + 26 + 28}{7} = 23.6$μ = 722+21+20+23+25+26+28 = 23.6
- **West:** $\mu = \frac{32 + 30 + 31 + 29 + 30 + 33 + 35}{7} = 31.4$μ = 732+30+31+29+30+33+35 = 31.4

The **West** region has the highest average monthly sales.

Question 2: Which region has the most consistent monthly sales (lowest variability)?

Process:

1. Calculate the standard deviation for each region to measure the spread of sales.

   Formula for Variance: $\sigma^2 = \frac{\sum (x - \mu)^2}{n}$σ2 = n∑(x−μ)2

   Formula for Standard Deviation: $\sigma = \sqrt{\sigma^2}$σ = σ2

   - σσ = standard deviation
   - μμ = mean
   - $\sum$∑ = sum of squared differences from the mean
   - nn = number of observations

2. Compare the standard deviations. The region with the lowest standard deviation is the most consistent.

Answer:

- **North Variance:** $\sigma^2 = \frac{(12 - 15.7)^2 + \ldots + (20 - 15.7)^2}{7} = 8.96$σ2 = 7(12−15.7)2+...+(20−15.7)2 = 8.96

  North Standard Deviation: $\sigma = \sqrt{8.96} = 2.99$σ = 8.96 = 2.99

- **South Variance:** $\sigma^2 = \frac{(22 - 23.6)^2 + \ldots + (28 - 23.6)^2}{7} = 6.8$σ2 = 7(22−23.6)2+...+(28−23.6)2 = 6.8

  South Standard Deviation: $\sigma = \sqrt{6.8} = 2.61$σ = 6.8 = 2.61

- **West Variance:** $\sigma^2 = \frac{(32 - 31.4)^2 + \ldots + (35 - 31.4)^2}{7} = 4.67$σ2 = 7(32−31.4)2+...+(35−31.4)2 = 4.67

  West Standard Deviation: $\sigma = \sqrt{4.67} = 2.16$σ = 4.67 = 2.16

The **West** region has the most consistent monthly sales due to the lowest standard deviation.

Using these processes, you can provide insights into various data-driven questions by employing descriptive statistics. The core concept involves using measures of central tendency (like mean) and measures of spread (like standard deviation) to glean insights from datasets.

Lets Increase the Complexity On Usecases

Usecase: A hospital wants to analyze the recovery times of patients undergoing a specific surgery. The data for recovery times (in days) over a month is as follows:

```
css
| Patient Group | Recovery Times (days)                |
|---------------|--------------------------------------|
| A             | 5, 6, 4, 5, 7, 5, 6                  |
| B             | 7, 8, 7, 9, 8, 7, 9                  |
| C             | 5, 7, 6, 5, 6, 6, 5                  |
```

Question 1: Which patient group has the quickest median recovery time?

Process:

1. Sort the recovery times for each group in ascending order.
2. Find the median (middle value) for each group.
3. Compare the medians.

Answer:

- **Group A Median:** The middle value of the sorted list (4, 5, 5, 5, 6, 6, 7) is 5.
- **Group B Median:** The middle value of the sorted list (7, 7, 7, 8, 8, 9, 9) is 8.
- **Group C Median:** The middle value of the sorted list (5, 5, 5, 6, 6, 6, 7) is 6.

Patient Group A has the quickest median recovery time of 5 days.

Question 2: Which patient group has the least variation in recovery times?

Process:

1. Calculate the range (difference between maximum and minimum values) for each group.
2. The group with the smallest range has the least variation.

Answer:

- **Group A Range:** 7 - 4 = 3
- **Group B Range:** 9 - 7 = 2
- **Group C Range:** 7 - 5 = 2

Patient Groups B and C both have the least variation in recovery times with a range of 2 days.

Question 3: How do the interquartile ranges (IQR) of the groups compare?

Process:

1. Calculate the first quartile (Q1) and third quartile (Q3) for each group.

   IQR Formula: IQR = Q3 - Q1

2. Subtract Q1 from Q3 to get the IQR for each group.

3. Compare the IQRs.

Answer:

- **Group A IQR:** For the sorted list (4, 5, 5, 5, 6, 6, 7), Q1 = 5 and Q3 = 6. IQR = 6 - 5 = 1.
- **Group B IQR:** For the sorted list (7, 7, 7, 8, 8, 9, 9), Q1 = 7 and Q3 = 9. IQR = 9 - 7 = 2.
- **Group C IQR:** For the sorted list (5, 5, 5, 6, 6, 6, 7), Q1 = 5 and Q3 = 6. IQR = 6 - 5 = 1.

Patient Groups A and C have the same IQR of 1 day, which is less than Group B's IQR.

These use cases illustrate how to utilize various descriptive statistics measures to analyze and interpret real-world data. By understanding the distributions, central tendencies, and variations of datasets, decisions can be more data-driven and informed.

Usecases on Different type Of Distributions

Usecase: An e-commerce company analyzes its website's page load times in seconds over a month to optimize user experience. The data includes:

```
css
| Day  | Load Times (seconds)                                        |
|------|-------------------------------------------------------------|
| 1    | 3, 2.5, 2.8, 3.1, 15 (Outlier due to a server glitch)       |
```

```
| 2    | 2.6, 2.5, 2.7, 2.9, 2.8                                    |
| 3    | 2.7, 2.8, 2.6, 2.5, 3                                      |
| ...  | ...                                                         |
```

Question 1: What impact do outliers have on the average load time?

Process:

1. Calculate the mean with and without outliers.
2. Compare both means to gauge the effect of outliers.

Answer:

With the outlier: Mean = (3 + 2.5 + 2.8 + 3.1 + 15) / 5 = 5.28

Without the outlier: Mean = (3 + 2.5 + 2.8 + 3.1) / 4 = 2.85

The outlier significantly increases the average page load time by 2.43 seconds.

Question 2: How can we transform load times to normalize the data?

Process:

1. Use logarithmic transformation.
2. Compute the logarithm (base 10 or natural logarithm) of all page load times.

Answer: Log-transforming the data can help in dealing with skewed data or data with outliers. If the original load time was 3 seconds, the transformed value using a natural log would be ln(3) ≈ 1.0986.

Question 3: Describe the distribution of load times using histograms.

Process:

1. Divide the data into bins (e.g., 2-2.5 seconds, 2.5-3 seconds).
2. Count the number of observations within each bin.
3. Plot the frequency of observations vs. bins.

Answer: Using the histogram, you might find, for instance, that most page load times cluster around 2.5-3 seconds, indicating the mode of the distribution. Peaks would represent common load times, while troughs would show less frequent load times.

Question 4: What is the Probability Density Function (PDF) for day 2's load times?

Process:

1. Estimate the PDF from the data (often using kernel density estimation).
2. Plot the continuous curve, showing how densities of load times vary.

Answer: The PDF will be a continuous curve indicating the probability of the page taking a specific time to load. For instance, the peak around 2.7 seconds might have a higher value, indicating it's the most common load time for day 2.

Question 5: What is the Probability Mass Function (PMF) for load times on day 3?

Process:

1. For discrete data, compute the proportion of each unique load time.
2. Plot these proportions.

Answer: The PMF might show, for instance, that the probability of the page taking exactly 2.7 seconds to load is 0.2 (or 20%). It gives probabilities for discrete outcomes.

These analyses can be deepened using more data and more advanced statistical methods, but the use case provides an insight into how different techniques in descriptive statistics can be used in a practical scenario.

Try this By your Own

Usecase 1: A pharmaceutical company has developed a new drug. During clinical trials, they measured the time (in hours) it took for patients to show symptom relief. They're particularly interested in how quickly the drug works.

Dataset Sample:

```
sql
| Patient Number | Relief Time (hours) |
|----------------|---------------------|
| 1              | 3.5                 |
| 2              | 2.8                 |
| 3              | 4.1                 |
| ...            | ...                 |
```

Question 1: Do the relief times follow a normal distribution?

Process:

1. Plot a histogram of the relief times.
2. Overlay a normal distribution curve on the histogram.

Answer: If the histogram matches closely with the normal distribution curve, then the relief times likely follow a normal distribution.

Question 2: What percentage of patients experienced relief within 3 hours, assuming the data follows a normal distribution?

Process:

1. Calculate the z-score for 3 hours: $z = \frac{X - \mu}{\sigma}$z = σX−μ
2. Look up this z-score in a z-table to find the percentage of patients.

Answer: If the z-score is, for example, -0.5 and corresponds to 30% on the z-table, then 30% of patients experienced relief within 3 hours.

Usecase 2: A factory produces light bulbs. They have a dataset of the number of bulbs produced each day and the percentage of defective bulbs. They want to improve the quality control process.

Dataset Sample:

```
sql
| Day  | Defective Bulbs (%) |
|------|---------------------|
| 1    | 2                   |
| 2    | 1.5                 |
| 3    | 3                   |
| ...  | ...                 |
```

Question 1: Do the percentages of defective bulbs follow a Poisson distribution?

Process:

1. If the occurrence of defects is rare and random, the distribution might follow a Poisson distribution.
2. Plot the PMF of the observed defects and compare with the PMF of a Poisson distribution with the same mean.

Answer: If the observed PMF aligns closely with the Poisson PMF, it's likely that the defect rates follow a Poisson distribution.

Question 2: If the data follows a binomial distribution, what is the probability that more than 5% of the bulbs are defective on any given day?

Process:

1. Use the binomial probability formula: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$P(X = k) = (kn)pk(1 − p)n−k Where:

- nn is the total number of trials (bulbs produced)
- kk is the number of successes (defective bulbs)
- pp is the probability of success on a single trial.

2. Calculate the probability for 5%, 6%, 7%,... and sum these probabilities.

Answer: The sum of the probabilities gives the likelihood that more than 5% of the bulbs are defective on any given day.

These use cases illustrate the application of different types of distributions (normal, Poisson, binomial) in real-world scenarios. In practice, determining the fit of a distribution would require more rigorous statistical testing, but this gives an overview of the process.

Try Some Inferential stats Ussecases with Python Code

1. Z-test

Question: A national examination board believes that the students in state X score an average of 52 in mathematics. A state education official disputes this and collects a random sample of 100 student scores from the state. The sample has an average score of 54 with a standard deviation of 10. At the 0.05 significance level, is the official correct?

Solution:

- **Null Hypothesis (H0):** The students in state X have an average score of 52.
- **Alternative Hypothesis (Ha):** The students in state X do not have an average score of 52.

Step by Step Process:

1. Calculate the z-score: $z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$z = σ/ n
   $\sqrt{}$
   X¯−μWhere:

- $\bar{X}$X¯ = sample mean = 54
- μμ = population mean = 52
- σσ = sample standard deviation = 10
- nn = number of samples = 100

2. Compare the z-score to the critical z-value for a 0.05 significance level (two-tailed).
3. If |z| > z-critical, reject the null hypothesis.

```
python
import math
import scipy.stats as stats

X_bar = 54
mu = 52
sigma = 10
n = 100

z = (X_bar - mu) / (sigma/math.sqrt(n))
p = 1 - stats.norm.cdf(abs(z))

alpha = 0.05
if p < alpha:
    print("Reject the null hypothesis")
else:
    print("Do not reject the null hypothesis")
```

---

2. T-test

Question: A company claims its new energy drink increases stamina. 15 people were tested before and after consuming the drink. Test if the drink has a significant effect on stamina at the 0.05 significance level.

Solution:

Given that the measurements are paired (before and after for the same individual), use a paired t-test.

Step by Step Process:

1. Compute the difference in stamina for each individual.
2. Compute the mean and standard deviation of these differences.
3. Calculate the t-statistic.
4. Compare the t-statistic to the critical t-value for a 0.05 significance level.

```
python
import numpy as np

before = np.array([...])  # insert stamina values before drinking
after = np.array([...])   # insert stamina values after drinking

differences = after - before
t_stat, p_value = stats.ttest_rel(after, before)

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Do not reject the null hypothesis")
```

---

3. ANOVA

Question: A farmer tests three types of fertilizers to see which one produces the highest crop yield. Is there a significant difference in yield across the fertilizers?

Solution:

Step by Step Process:

1. Use one-way ANOVA to compare the means of crop yields from the three fertilizers.
2. If the p-value is below the significance level, there is a significant difference.

```
python
fertilizerA = np.array([...])  # insert yields for fertilizer A
fertilizerB = np.array([...])  # insert yields for fertilizer B
fertilizerC = np.array([...])  # insert yields for fertilizer C

f_stat, p_value = stats.f_oneway(fertilizerA, fertilizerB, fertilizerC)

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Do not reject the null hypothesis")
```

---

4. Chi-Square Test

Question: A company wants to know if there's a relationship between gender (male, female) and product preference (Product A, Product B). They survey 100 customers. Is product preference independent of gender?

Solution:

Step by Step Process:

1. Construct a contingency table of gender vs. product preference.
2. Compute the chi-square statistic and p-value.
3. If the p-value is below the significance level, they are not independent.

```
python
# Contingency table: rows = gender, columns = product preference
observed = np.array([[30, 20],  # males
                     [25, 25]]) # females

chi2_stat, p_value, _, _ = stats.chi2_contingency(observed)

alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Do not reject the null hypothesis")
```

---

5. Regression

Question: An e-commerce website wants to understand if the time spent on the website (in minutes) predicts the total amount spent (in dollars). They gather data from 100 users. Determine if there's a relationship.

Solution:

Step by Step Process:

1. Run a simple linear regression with time spent as the independent variable and amount spent as the dependent variable.
2. If the p-value for the slope is below the significance level, there's a significant relationship.

```
python
from statsmodels import api as sm

time_spent = np.array([...])    # insert time spent by users
amount_spent = np.array([...])  # insert amount spent by users

X = sm.add_constant(time_spent)  # adding a constant
model = sm.OLS(amount_spent, X).fit()

alpha = 0.05
if model.pvalues[1] < alpha:
    print("Reject the null hypothesis")
else:
    print("Do not reject the null hypothesis")
```

Remember, these are simplifications of what you might encounter in the real world. In practice, you would also check assumptions, consider additional covariates, and potentially apply more sophisticated statistical models.

This usecase definitely you should solve

Scenario: Online Retail Store A/B Testing

An online retail store has recently introduced a new webpage design to increase the amount of time users spend on the page and ultimately increase purchases. They have conducted A/B testing, where Group A is exposed to the old design, and Group B to the new design. They've collected data on the time spent on the webpage and whether a purchase was made.

Objective: Determine if the new webpage design leads to a significant increase in both time spent on the webpage and the likelihood of making a purchase.

Steps:

1. Define the Problem:

- **Null Hypothesis (H0):** The new webpage design does not significantly affect the time spent on the webpage and the likelihood of making a purchase.
- **Alternative Hypothesis (HA):** The new webpage design significantly affects the time spent on the webpage and the likelihood of making a purchase.

2. Data Collection:

- Collect data on time spent on the webpage and purchasing behavior for both groups.

3. Data Exploration and Preprocessing:

- Understand the basic statistics of the datasets.
- Handle missing values if any.
- Check and handle outliers.

4. Perform T-Test on Time Spent:

- Conduct an Independent Samples t-test to compare the mean time spent on the webpage by the two groups.

5. Perform Chi-Square Test on Purchase Behavior:

- Construct a contingency table of the groups and purchasing behavior.
- Conduct a Chi-Square test to check the independence of the group and purchasing behavior.

6. Decision Making:

- Based on the p-values from the t-test and Chi-Square test, reject or fail to reject the null hypothesis.
- Make recommendations for the business.

Python Code:

```
python
import numpy as np
import pandas as pd
import scipy.stats as stats

# Suppose `data` is the collected data with 'group', 'time_spent', and 'purchase' columns
# 'group' - 'A' for control group and 'B' for test group
# 'time_spent' - time spent by the user on the webpage
# 'purchase' - 1 if the user made a purchase, 0 otherwise

# Sample data creation
data = pd.DataFrame({
    'group': ['A', 'A', 'B', 'A', 'B', 'B', 'A', 'B'],
    'time_spent': [3, 5, 7, 4, 6, 8, 2, 7],
    'purchase': [0, 1, 1, 0, 1, 1, 0, 1]
})

# Step 4: Perform T-Test on Time Spent
group_A_time_spent = data[data['group'] == 'A']['time_spent']
group_B_time_spent = data[data['group'] == 'B']['time_spent']
```

```
t_stat, p_value_time = stats.ttest_ind(group_A_time_spent, group_B_time_spent)

# Step 5: Perform Chi-Square Test on Purchase Behavior
contingency_table = pd.crosstab(data['group'], data['purchase'])
chi2_stat, p_value_purchase, _, _ = stats.chi2_contingency(contingency_table)

# Step 6: Decision Making
alpha = 0.05

if p_value_time < alpha:
    print("There is a significant difference in time spent on the webpage between the two groups.")
else:
    print("There is no significant difference in time spent on the webpage between the two groups.")

if p_value_purchase < alpha:
    print("There is a significant difference in purchasing behavior between the two groups.")
else:
    print("There is no significant difference in purchasing behavior between the two groups.")
```

In this script:

- Data for two groups (A and B) is analyzed.
- A T-test is conducted to compare the mean time spent on the webpage for the two groups.
- A Chi-Square test is conducted to assess the association between the group and purchasing behavior.
- Decisions are made based on the p-values from the tests to determine whether there are significant differences in time spent on the webpage and purchasing behavior between the two groups.

Recommendation:

- If there is a significant increase in both time spent and likelihood of making a purchase, recommend the implementation of the new webpage design.
- If not, further analysis and perhaps more A/B testing may be needed to identify effective strategies for improving webpage performance and sales.

Lets Solve with each step by step

Scenario: Online Retail Store A/B Testing

Hypothetical Data:

Group A (Old Design):

- Time spent (minutes): [3, 5, 4, 6, 5, 5, 6, 4]
- Purchases: [0, 1, 0, 1, 1, 0, 0, 1]

Group B (New Design):

- Time spent (minutes): [6, 7, 7, 7, 8, 6, 7, 8]
- Purchases: [1, 1, 1, 1, 1, 1, 0, 1]

Step-by-Step Solution with Numerical Calculation:

1. Define the Problem:

As stated in the previous answer.

2. Data Collection:

Data has been hypothetically given.

3. Data Exploration and Preprocessing:

- **Calculate the means and standard deviations for both groups:**

Group A Time spent:

- Mean = (3 + 5 + 4 + 6 + 5 + 5 + 6 + 4) / 8 = 4.75 minutes
- Standard Deviation ≈ 1.16 minutes

Group B Time spent:

- Mean = (6 + 7 + 7 + 7 + 8 + 6 + 7 + 8) / 8 = 7 minutes
- Standard Deviation ≈ 0.76 minutes

4. Perform T-Test on Time Spent:

- **Compute the t-statistic using the formula:** $t = \frac{X1 - X2}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$t = n1s12+n2s22 X1⁻−X2⁻

Where:

- $X1$X1⁻ and $X2$X2⁻ are the sample means of group A and B.
- $s1$s1 and $s2$s2 are the sample standard deviations of group A and B.
- $n1$n1 and $n2$n2 are the sample sizes of group A and B.

Using the values: $t = \frac{4.75 - 7}{\sqrt{\frac{1.16^2}{8} + \frac{0.76^2}{8}}}$t = 81.162+80.762 4.75−7$t \approx -5.91$t ≈ −5.91

Consulting a t-distribution table for df=14 (since df = n1 + n2 - 2) and α=0.05 (two-tailed), the critical t-value is approximately ±2.145.

Since -5.91 < -2.145, we reject the null hypothesis for time spent.

5. Perform Chi-Square Test on Purchase Behavior:

Construct a 2x2 contingency table:

	Purchase=0	Purchase=1	Total
Group A	4	4	8
Group B	1	7	8

- **Calculate the expected frequencies for each cell:**
  For example, for Group A and Purchase=0: Expected frequency = (row total * column total) / grand total = (8 * 5) / 16 = 2.5
- **Compute the chi-square statistic using the formula:**
  $\chi^2 = \sum \frac{(observed - expected)^2}{expected}$χ2 = ∑ expected(observed−expected)2

Using the observed and expected frequencies: $\chi^2 \approx 3.6$χ2 ≈ 3.6

For a 2x2 table with α=0.05, the critical value from the chi-square distribution is approximately 3.841.

Since 3.6 < 3.841, we fail to reject the null hypothesis for purchase behavior.

6. Decision Making:

- The t-test indicates a significant difference in time spent on the webpage between the two designs.
- The chi-square test suggests that purchase behavior isn't significantly different between the groups, although it's close to the threshold.

Recommendation:

- The new webpage design seems to engage users for longer periods.
- While purchase behavior hasn't shown a statistically significant change, it's close, and with a larger sample, it might.

Further A/B testing or possibly combining this new design with other strategies could be beneficial for sales.
