

① Agenda:

- ① Histograms
- ② measure of central Tendency.
- ③ measure of Dispersion.
- ④ Percentiles & Quartiles.
- ⑤ 5 Number Summary (Box Plot).

# 1) Histograms :-

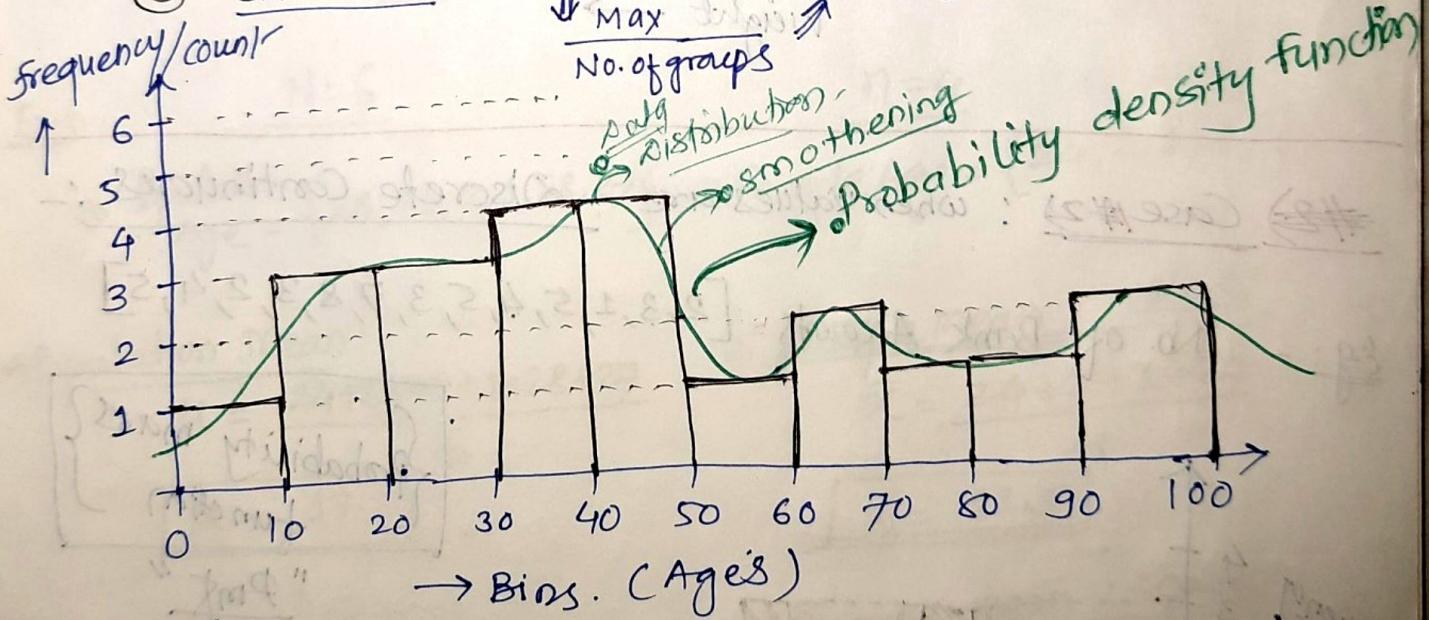
Ex:- ① Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

$$\begin{aligned} \text{Min} &= 10 \\ \text{Max} &= 100 \end{aligned}$$

Steps: ① sort the numbers.

② Define Bins = No. of groups. = 10.

③ Bins size = size of bins =  $\frac{100}{10} = 10$ .



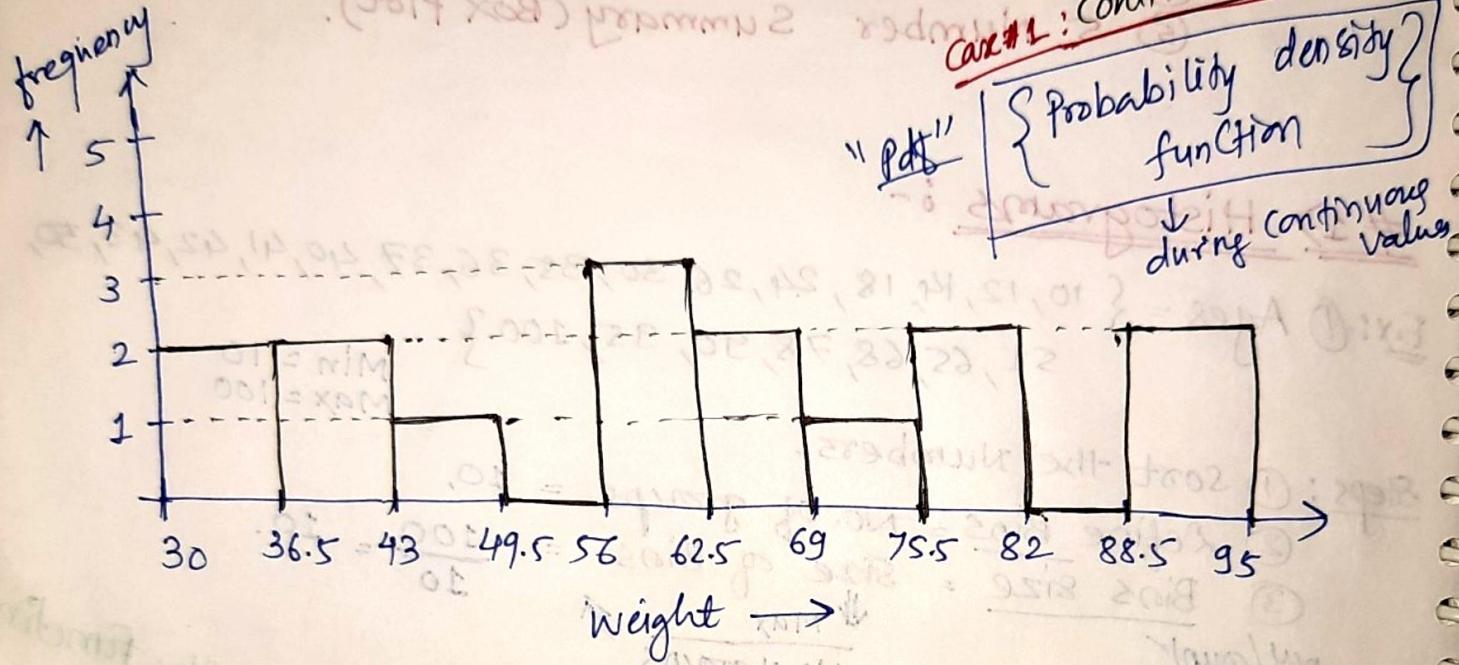
Eg. ②

$$\text{weight} = \{30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95\}$$

Sol: → Bins = 10

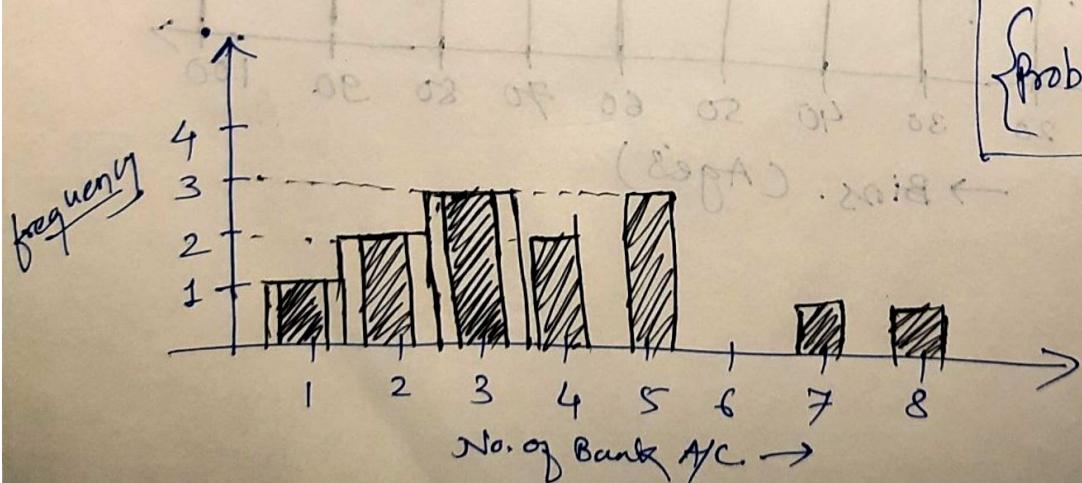
(since here we are starting from 30)

$$\rightarrow \text{Bin Size} = \frac{\text{Max} - \text{Min}}{\text{Bins}} = \frac{95 - 30}{10} = \frac{65}{10} = 6.5.$$



#2) Case #2: when values are Discrete Continuous :-

Eg:- No. of Bank Accounts = [2, 3, 1, 5, 4, 5, 3, 7, 8, 3, 2, 4, 5]



{ Probability mass function }

"Pmt"  
(during discrete values)

# 2] Measure of Central Tendency :-

- ① Mean
- ② Median
- ③ Mode

} — A measure of CT is a single value that attempts to describe a set of data by identifying the central position.

### ① Mean :

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

$$\begin{aligned} \text{Average/Mean} &= \frac{1+2+3+4+5}{5} \\ &= \frac{15}{5} \\ &\rightarrow = 3 \end{aligned}$$

Population ( $N$ )

$N > n$

Sample ( $n$ )

• Population Mean  $(\mu) = \frac{\sum_{i=1}^N x_i}{N}$

• Sample Mean ( $\bar{x}$ )  $= \frac{\sum_{i=1}^n x_i}{n}$

Ex :-

$$\underline{N=6}$$

$$\Rightarrow \text{Population} = \{24, 23, 21, 28, 27\}$$

Age = {24, 23, 21, 28, 27}

$$\underline{n=4}$$

$$\Rightarrow \text{Sample Age} = \{24, 21, 27\}$$

Population Mean :

$$\mu = \frac{24+23+21+28+27}{6}$$

$$\boxed{\mu = 25.5}$$

$\Rightarrow$  Sample Mean : case ①

$$\bar{x} = \frac{24+21+27}{4}$$

$$\boxed{\bar{x} = 22.5}$$

$\Rightarrow$  Sample mean : case ②

$$\bar{x} = \frac{24+23+28+27}{4}$$

$$\boxed{\bar{x} = 25.5}$$

Hence

$\mu > \bar{x}$
$\mu \leq \bar{x}$

and

Both possible.

## • Practical Application of mean (Feature Engineering) :-

Eg:	Age	Salary	Family Size	To:
	-	-	-	
	-	-	-	
<u>Loss of info -</u>				
If we remove Row	NAN	-	-	→ If we remove the NAN row, then there will be loss of info/data.
	-	-	-	→ Hence to prevent loss of info, we replace NAN with mean value of particular column.
	NAN	NAN	NAN	
	(D) signs	(D) signs	(D) signs	
	NAN	NAN	NAN	

Eg:-	Age	Salary	Age Mean = $\frac{24+28+29+31+36}{5} = 29.6$
	24	45	
	28	50	
	29	NAN - 62	
	NAN 29.6	60	
	31	75	
	36	80	
	NAN 29.6	NAN - 62	

$$\therefore \text{Replace Age NAN} = 29.6$$

$$\& \text{Salary NAN} = 62.$$

## ② Median :-

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

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

$$\boxed{\bar{x} = 3}$$

$$X = \{1, 2, 3, 4, 5, \boxed{100}\}$$

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

$$\boxed{\bar{x} = 19.16}$$

→ when there is an outlier, we calculate median instead of mean.

### # Steps to find out median :

① Sort the numbers

② find the central number:

{Case-i :- if no. of elements are even we find the average of central elements,

case-ii :- if no. of elements are odd, we find the central elements. }

e.g.:  $X = \{1, 2, 3, 4, \boxed{5, 6}, 7, 8, 100, 120\}$

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

$$\left. \begin{array}{l} \text{Mean} = \frac{1+2+3+4+5+6+7+8+100+120}{10} \\ \boxed{\text{Mean} = 25.6} \end{array} \right\}$$

~~Hence, when~~

Hence, when :

i) No Outliers → Calculate Mean

ii) with Outliers → Calculate Median.

### ③ Mode :-

- "Most frequent occurring elements"

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

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

$\therefore \boxed{\text{Mode} = 3}$

$$\therefore \boxed{\text{Mode} = 2, 3}$$

e.g. Dataset

types of flowers ← {categorical variable}

Lily

Sunflower

Rose

NAN

Rose

Sunflower

Rose

NAN

- If NAN had occurred most in column, then we need to remove NAN.

$\therefore$  replace NAN with Rose, as Rose occurred most in column.

$$\text{Ans} = \frac{3+2}{5} = \boxed{0.6}$$

### #3] Measure of Dispersion :-

- ① Variance ( $\sigma^2$ ) → "Study spread of Data"
- ② Standard deviation ( $\sigma$ )

#### ① Variance ( $\sigma^2$ )

##### A) Population Variance ( $\sigma^2$ )

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

##### B) Sample Variance ( $s^2$ )

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

~~Population Variance~~

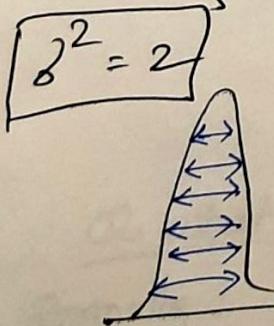
$$\text{Ex: } ① X = \{1, 2, 3, 4, 5\}$$

$$\mu = 3$$

$$\sigma^2 = \frac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$$

$$\sigma^2 = \frac{4+1+0+1+4}{5}$$

$$\sigma^2 = \frac{10}{5}$$

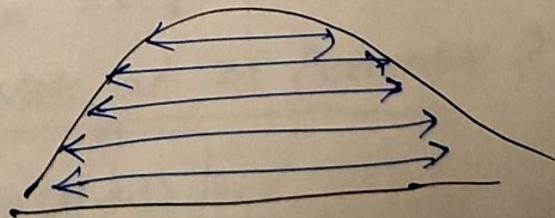


$$X = \{1, 2, 3, 4, 5, 6, 8, 0\}$$

$$\mu = 14.4$$

$$\sigma^2 = \frac{[(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2]}{7}$$

$$\sigma^2 = 719.10$$



∴ Variance  $\propto$  spread ↑↑

② standard deviation ( $\delta$ ) i.e.  $\delta = \sqrt{\sigma^2}$  ⑧  
 i.e.  $\sqrt{\text{variance}}$

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

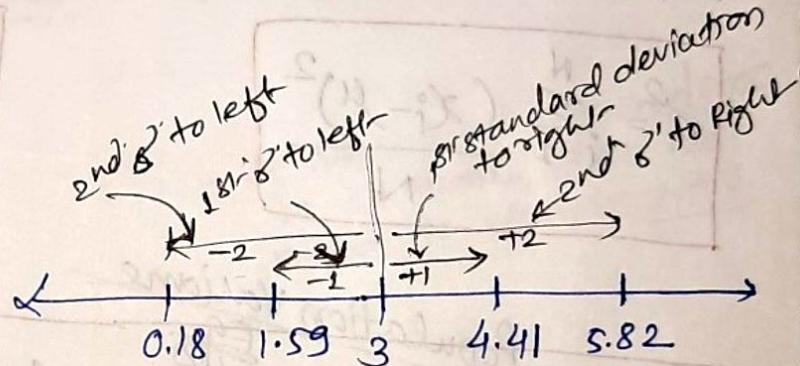
$$\boxed{\mu = 3}$$

$$\sigma^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

$$\begin{aligned} &= \frac{4+1+0+1+4}{5} \\ &= \frac{10}{5} \\ \boxed{\sigma^2 = 2} \end{aligned}$$

$$\therefore \delta = \sqrt{2} = 1.41$$

$$\boxed{\delta = 1.41}$$



$$0.1812 = \delta$$

$$\frac{1+1+0+1+1}{5} = \delta$$

$$\frac{01}{2} = \delta$$

$$\boxed{\delta = 0.8}$$

$\boxed{\text{Mean} \times \text{Median}}$

## #4] Percentiles & Quartiles :-

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

$$\cdot \% \text{ of even No.} = \frac{\text{No. of even No.}}{\text{Total No. of No.}} = \frac{4}{8} \times 100 = 50\%$$

### ② Percentiles :-

- A percentile is a value below which a certain percentage of observations lie.

e.g. - 99 percentiles  $\Rightarrow$  It means the person has got better marks than 99% of the entire students.  
(data should be in ascending order always)

- Dataset : {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12, ?}

Q. What is the percentile rank of 10?

Formula:  $\boxed{\text{Percentile Rank of } x = \frac{\text{No. of value below } x}{n}}$

$$\therefore \text{percentile rank of } 10 = \frac{16}{20} = 80 \text{ Percentile.}$$

$$\therefore 8 = \frac{9}{20} = 45 \text{ percentile.}$$

$$\therefore 6 = \frac{7}{20} = 35 \text{ percentile.}$$

Q. What is the value that exists at 25 percentile?

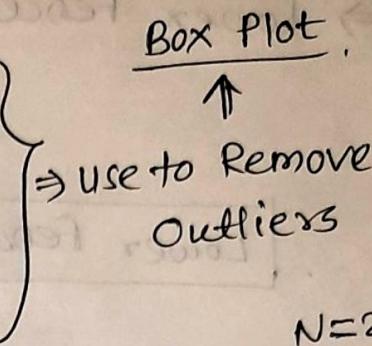
Formula:  $\boxed{\text{value} = \frac{\text{Percentile} \times (n+1)}{100}}$

$$\therefore \text{Value} = \frac{25 \times 20}{100} = 5^{\text{th}} \text{ index}$$

$$\boxed{\text{value} = 5.}$$

## #5] 5 Number Summary :-

- ① Minimum
- ② First Quartile (25 Percentile) ( $Q_1$ )
- ③ Median
- ④ Third Quartile (75 Percentile) ( $Q_3$ )
- ⑤ Maximum

Box Plot   
⇒ Use to Remove Outliers

N=20

Eg:-  $\{1, 2, 2, 2, \boxed{3}, \boxed{3}, \boxed{3}, 4, 5, 5, 5, 6, 6, 6, \boxed{6}, \boxed{7}, \boxed{8}, 8, 9, \boxed{27}\}$  - outlier

formulae:- [ Lower fence  $\leftarrow$  Higher fence ]

$$\bullet \text{Lower fence} = Q_1 - 1.5(\text{IQR})$$

$$\bullet \text{Higher fence} = Q_3 + 1.5(\text{IQR})$$

$$\bullet \therefore \text{IQR} = Q_3 - Q_1$$

(Inter Quartile Range)

$$\text{Soln} \Rightarrow Q_1 = \frac{25}{100} \times (n+1) = \frac{25}{100} \times (20+1) = \frac{25}{100} \times 21.$$

$Q_1 = \underline{5.25 \text{ index}}$  (here we will take Avg of 5<sup>th</sup> & 6<sup>th</sup> index)

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

$$\therefore Q_1 = 3$$

$$\Rightarrow Q_3 = \frac{75}{100} \times (n+1) = \frac{75}{100} \times (20+1) = \frac{75}{100} \times 21$$

$Q_3 = \underline{15.75 \text{ index}}$  (here we will take Avg of 15<sup>th</sup> & 16<sup>th</sup> index)

$$= \frac{7+8}{2}$$

$$\therefore Q_3 = 7.5$$

$$\therefore \text{IQR} = Q_3 - Q_1 = 7.5 - 3$$

$$\therefore \text{IQR} = 4.5$$

$$\Rightarrow \text{Lower Fence} = Q_1 - 1.5(\text{IQR})$$

$$= 3 - 1.5(4.5)$$

$$= 3 - 6.75$$

**Lower Fence = -3.75**

$$\Rightarrow \text{Higher Fence} = Q_3 + 1.5(\text{IQR})$$

$$= 7.5 + 1.5(4.5)$$

**Higher Fence = 14.25**

$$\Rightarrow \therefore [\text{Lower Fence} \leftrightarrow \text{Higher Fence}]$$

$$[-3.75 \leftrightarrow 14.25]$$

**Range**

$\therefore \{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$  } **outlier will be removed for further calculation**

final values will be :-

① Minimum = 1

②  $Q_1 = 3$

③ Median = 5

④  $Q_3 = 7.5$

⑤ Maximum = 9

Box Plot

$\Rightarrow$  To treat outliers

