Agenda:

1. Histograms
2. Measure of Central Tendency.
3. Measure of Dispersion.
4. Percentiles & Quartiles.
5. 5 Number Summary (Box Plot).

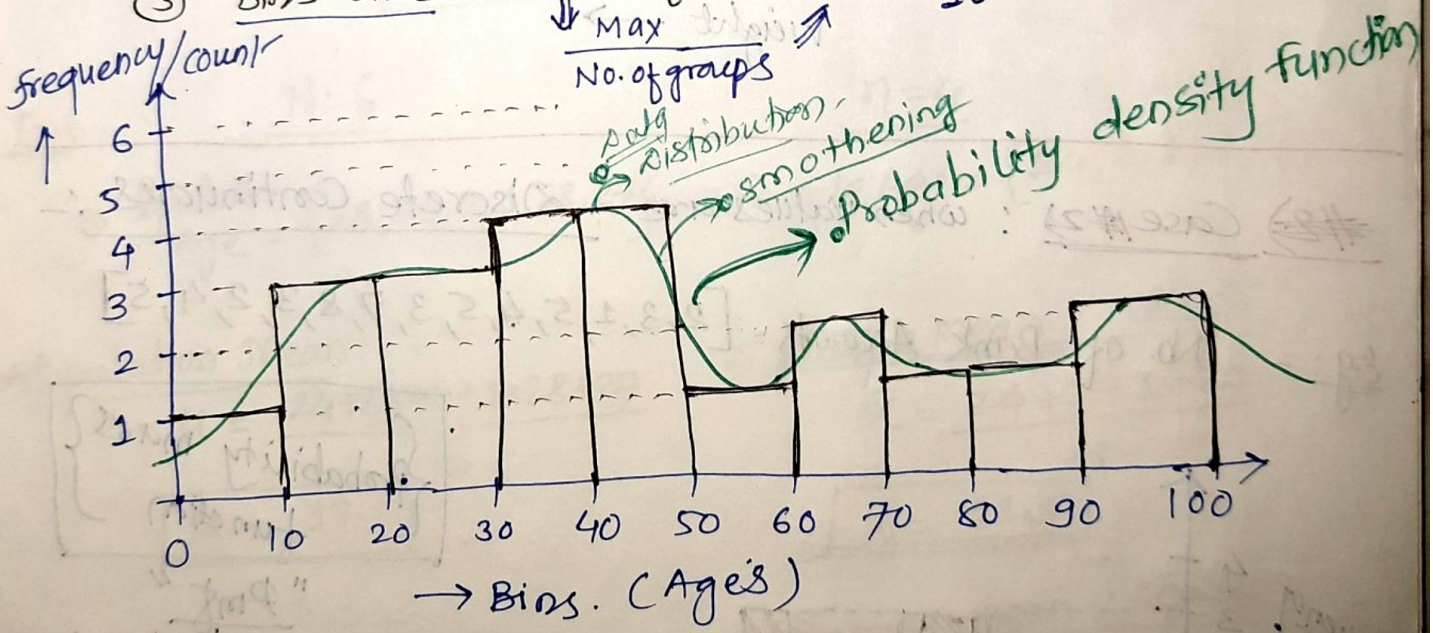
#1) Histograms :-

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 }

Min = 10
Max = 100

Steps: ① sort the Numbers.
② Define Bins = No. of groups. = 10.
③ Bins size = Size of Bins. = $\frac{100}{10}$ = 10.

$\Downarrow$ Max / No. of groups



Eg. ② weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

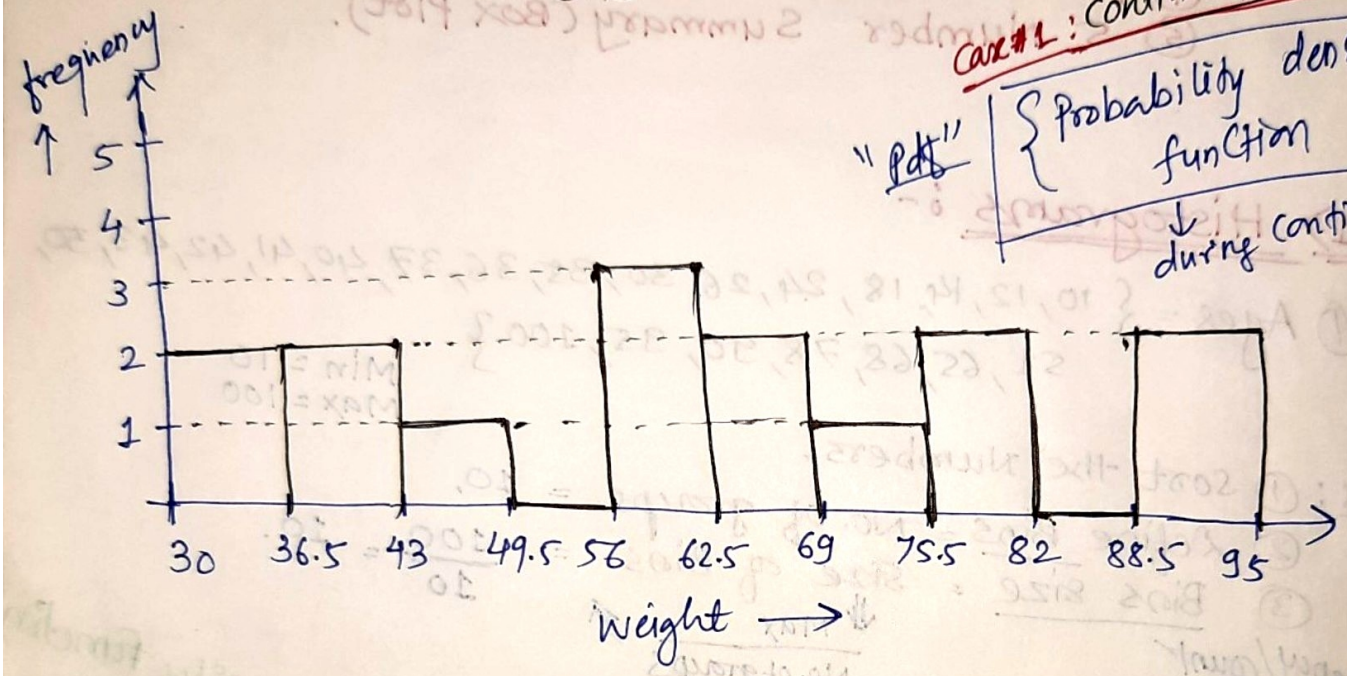
Max - 95
Min - 30

Sol^{n} → → Bins = 10 (since here we are starting from 30)

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

Case #1 : Continuous value

"Pdf" { Probability density function } ↓ during continuous value

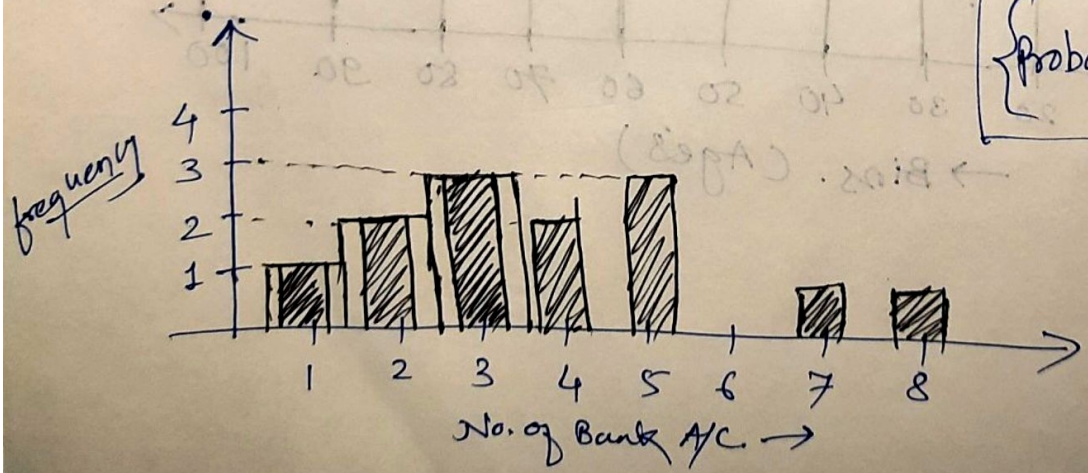


#2) Case #2) : when values are Discrete Continuous :-

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

{ Probability mass function }

"Pmf" (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, identifying the central position.

① Mean :

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

Average/Mean $= \frac{1+2+3+4+5}{5}$

$= \frac{15}{5}$

$= 3$

Population (N) | $N \geq n$ | Sample (n)

- Population Mean (Mu) $(\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$
- Sample Mean $(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$

Ex:

$N = 6$

⇒ Population Age $= \{24, 23, 2, 1, 28, 27\}$

⇒ Population Mean :

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

$\mu = 17.5$

$n = 4$

⇒ Sample Age $= \{24, 2, 1, 27\}$

⇒ Sample mean : case ①

$\bar{x} = \frac{24+2+1+27}{4}$

$\bar{x} = 13.5$

⇒ Sample mean : case ②

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

$\bar{x} = 25.5$

Hence $\mu \geq \bar{x}$ and $\mu \leq \bar{x}$ Both possible.

• Practical Application of mean (Feature Engineering) :- ④

Eg: Age Sallary family size

To :

Loss of info

NAN

If we remove Row

⇒ If we remove the NAN Row/Line, 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

NAN

Eg:

Age	Sallary
24	45
28	50
29	NAN - 62
NAN - 29.6	60
31	75
36	80
NAN - 29.6	NAN - 62

Sol ⇒ • Age mean = $\frac{24+28+29+31+36}{5}$

↳ = 29.6

• Sallary mean = $\frac{45+50+60+75+80}{5}$

↳ = 62

∴ Replace Age NAN = 29.6

& Sallary NAN = 62

(2) Median :-

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

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

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

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

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

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

⇒ 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. }

eg:- $X = \{1,2,3,4,\boxed{5,6},7,8,100,120\}$

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

$\text{Mean} = \frac{1+2+3+4+5+6+7+8+100+120}{10}$

$\boxed{\text{Mean} = 25.6}$

Hence, when :

(i) No Outliers → Calculate Mean

(ii) with Outliers → Calculate Median.

③ Mode :-

- "Most frequent occuring elements".

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

$\therefore$ Mode = 3 $\therefore$ Mode = 2, 3

eg: Dataset

Types of flowers ← {categorical variable}

Lily

Sunflower

Rose

NAN ← Rose

Rose

Sunflower

Rose

NAN ← Rose

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

$\therefore$ Replace NAN with Rose, as Rose occured most in column.

#3] Measure of Dispersion :-

① variance (σ^2) → "Study Spread of Data"

② Standard deviation (σ)

① Variance (σ^2)

(A) Population Variance (σ^2)

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

(B) Sample Variance (s^2)

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

~~Population Variance~~

Eg: ① $X = \{1, 2, 3, 4, 5\}$

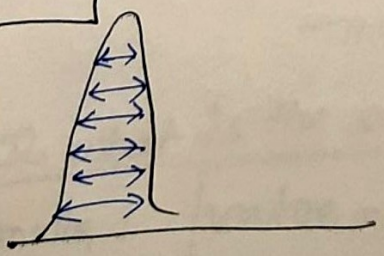
$\mu = 3$

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

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

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

$\sigma^2 = 2$

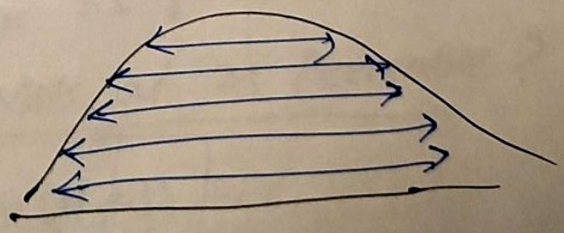


Eg: ② $X = \{1, 2, 3, 4, 5, 80\}$

$\mu = 14.4$

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

$\sigma^2 = 719.10$



∴ Variance ↑↑ ∝ spread ↑↑

(2) Standard deviation (σ) :- $= \sqrt{\sigma^2}$

i.e. $\sqrt{\text{variance}}$

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

$\mu = 3$

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

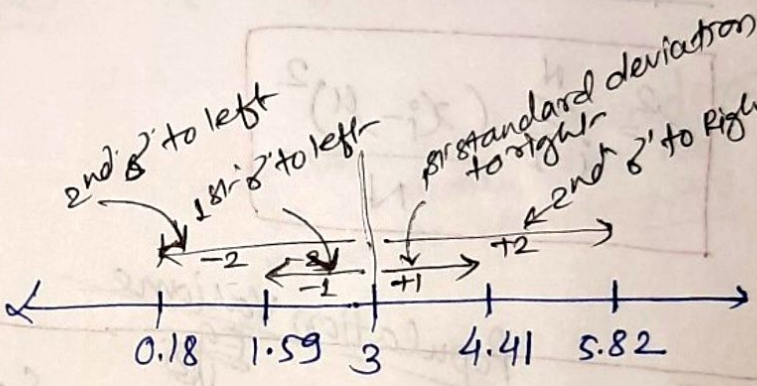
$$= \frac{4+1+0+1+4}{5}$$

$$= \frac{10}{5}$$

$\sigma^2 = 2$

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

$\sigma = 1.41$



#4] Percentiles & Quartiles :-

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

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

◉ Percentiles :-

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

eg. - 99 percentiles ⇒ 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, 9, 9, 10, 11, 11, 12.}

Q. what is the percentile rank of 10?

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

∴ percentile Rank of 10 = $\frac{16}{20}$ = 80 Percentile.

-11- 8 = $\frac{9}{20}$ = 45 percentile.

-11- 6 = $\frac{7}{20}$ = 35 percentile.

Q. what is the value that exists at 25 Percentile?

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

∴ value = $\frac{25}{100} \times 20 = 5^{th}$ index

value = 5.

#5] 5 Number Summary :-

Box Plot ↑ ⇒ use to Remove Outliers

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

N=20

Eg:- {1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27} - outlier
(4th, 5th, 6th; 14th, 15th, 16th)

Formulae :- [Lower fence ⟷ Higher fence]

- Lower fence = Q1 − 1.5 (IQR)
- Higher fence = Q3 + 1.5 (IQR)
- ∵ IQR = Q3 − Q1

(Inter Quartile Range)

Sol^n :- ⇒ $Q1 = \frac{25}{100} \times (n+1) = \frac{25}{100} \times (20+1) = \frac{25}{100} \times 21$

Q1 = 5.25 index (here we will take Avg of 5th & 6th index)

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

∴ Q1 = 3

⇒ $Q3 = \frac{75}{100} \times (n+1) = \frac{75}{100} \times (20+1) = \frac{75}{100} \times 21$

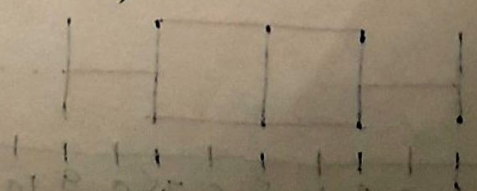
Q3 = 15.75 index (here we will take Avg of 15th & 16th index)

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

Q3 = 7.5

⇒ ∴ IQR = Q3 − Q1 = 7.5 − 3

∴ IQR = 4.5



P.T.O.

⇒ Lower Fence = $Q1 - 1.5(IQR)$

$= 3 - 1.5(4.5)$

$= 3 - 6.75$

Lower Fence = −3.75

⇒ Higher Fence = $Q3 + 1.5(IQR)$

$= 07.5 + 1.5(4.5)$

Higher Fence = 14.25

⇒ ∴ Range [Lower Fence ⟷ Higher Fence]

[−3.75 ⟷ 14.25]

Range

∴ {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 Q1 = 3

③ Median = 5

④ Q3 = 7.5

⑤ Maximum = 9

Box Plot ⇓

⇒ To treat outliers