Day-2 Statistics

4 - Sep - 2022

Topics

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, 18, 24, 26, 30, 35, 36, ~~37~~, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 102

① Sort the no

② Bins → no. of groups

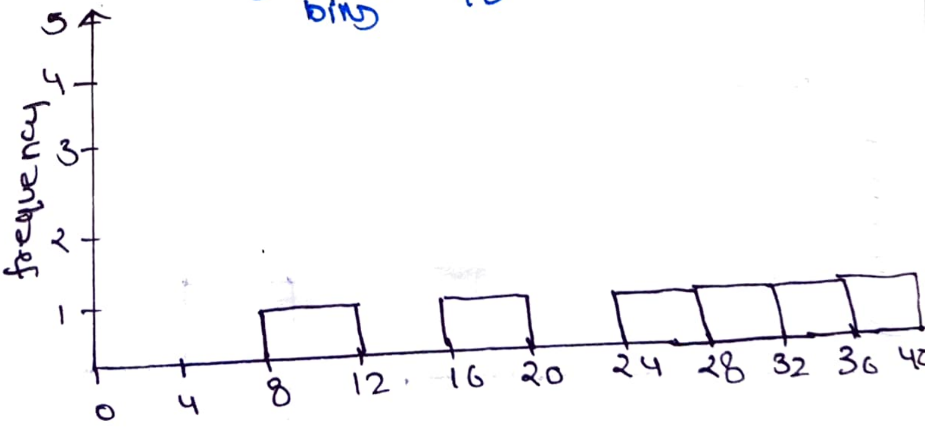
③ Bins size → size of Bins

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

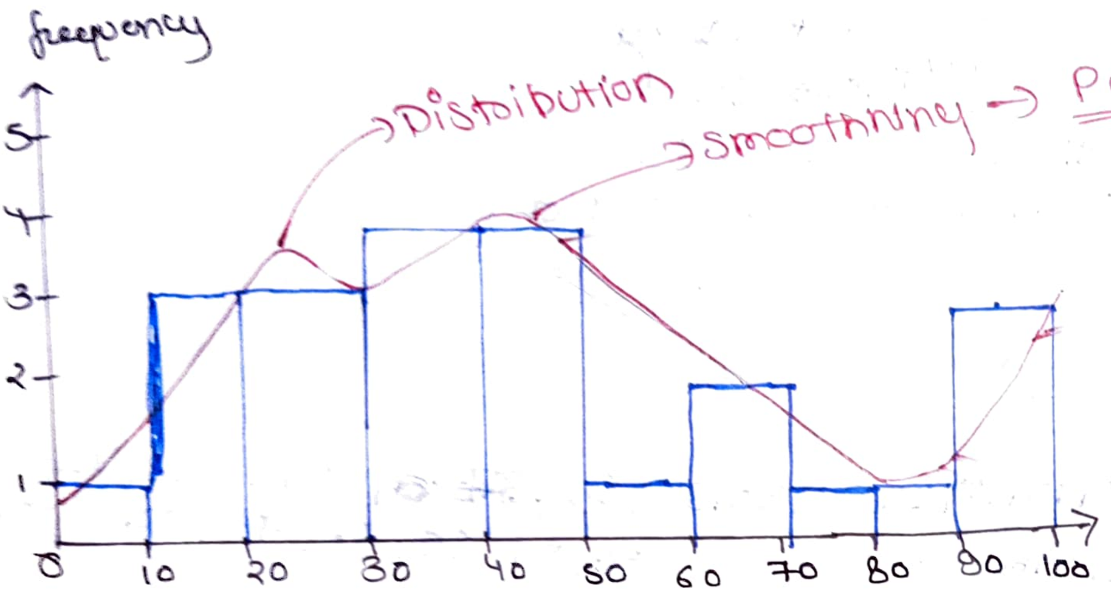
min = 10   max = 40

bins = 10 → means 10 group b/w 0 to 40

$= \frac{max}{bins} = \frac{40}{10} = 4$

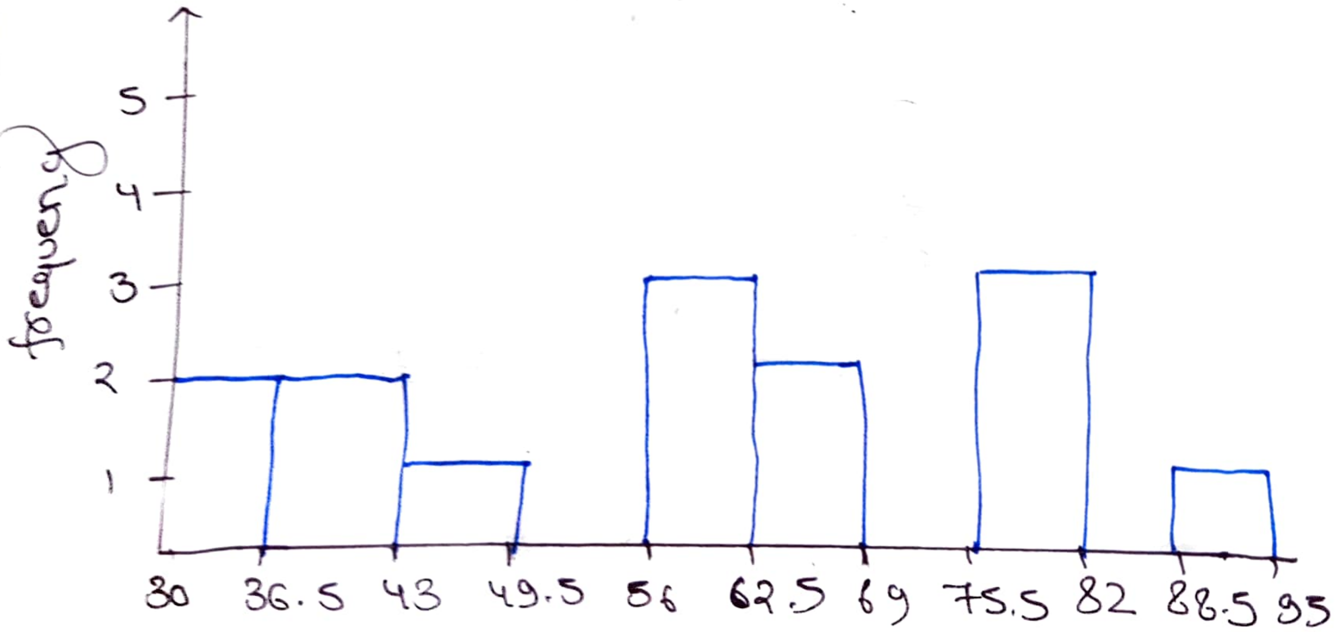


bins = 10 bin size = $\frac{100}{10}$ = 10 → 10 group b/w 0 to 100



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

bins = 10 bin size = $\frac{65}{10}$ = 6.5



② Discreate & Continous

Banks = {1,2,3,4,5,6}

No. of Banks accounts = [2,3,5,1,4,5,3,7,8 3,2,4,5]

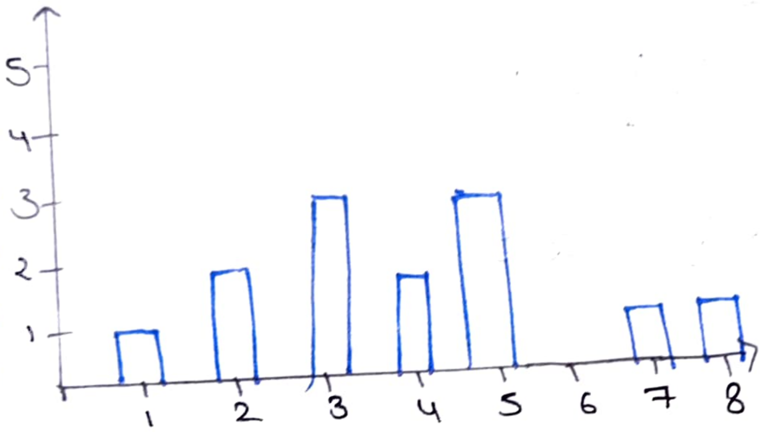
Discrete values

↓

Smoothing

↓ using

Probability mass function



Pdf = Probability density func → Continous

pmf = Probability mass func → discreate

→ Measure of Central Tendency

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

① Mean

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

Avg/mean $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$ → Specifying the central position

Population (N)

$$\boxed{\text{Population mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N}}$$

Sample (n)

$$\boxed{\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}}$$

$\boxed{N >> n}$ → Population always greater than Sample eg

Population Age = $\{24, 23, 2, 1, 28, 27\}$ $\boxed{N=6}$

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

$$= 17.6 \Rightarrow \boxed{\mu = 17.6}$$

Sample Age = $\{24, 2, 1, 27\}$ $\boxed{n=4}$

$$\bar{x} = \frac{24+2+1+27}{4} = 13.5$$

$$\boxed{\bar{x} = 13.5}$$

* $\left.\begin{array}{l}\mu \geq \bar{x} \\ \bar{x} \geq \mu\end{array}\right\}$ these Situation can happen

→ Practical Application (feature Engg)

eg

Age	salary	family size
—	—	—
—	—	—
NAN	—	—
—	NAN	—
—	—	NAN
—	—	—
—	—	—
NAN		

→ Droping Row having NAN

↓

loss of Info

↓ instead

NAN are replaced by Mean

Eg

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

Age Mean = 29.6   Salary mean = 62

new Data → 80   200 $\xrightarrow[\text{mean}]{\text{new}}$ Age Mean = 38   Salary Mean = 85

→ Outliers

Outliers :- an observation that lies at abnormal distance from other values in a random sample from a population

② Median

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

↑ outlier

$\bar{x} = 3 \longrightarrow \bar{x} = 19.16$

→ Steps to find out median

① Sort the numbers

② find the central ~~Index~~ Number

① if no of Elements are even we find the avg. of central Elements

② odd no. of Element → find central elements eg {1, 2, 3, 4, 5, 6, 7, 8, 100, 120}

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

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

Median = 5

* no outlier → Mean
with outlier → Median

③ **Mode** → most frequent occuring elements

eg {1, 2, 2, 3, 3, 3, 4, 5}

mode = 3

eg {1, 2, 2, 2, 3, 3, 3, 4, 5}

mode = 2, 3

eg Types of flowers
Lily, Sunflower, Rose, NAN, Rose, Sunflower, Rose, NA

~~* mode~~
* mode is mostly use with categorical variable

* Measoure of Dispersion

① variance (σ^2) ← spread of data

② Standard deviation (σ)

① Variance

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

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

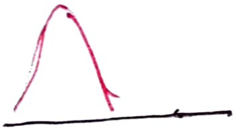
- $x_i - \mu$ → deviation from mean

eg

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

$\mu = 3$

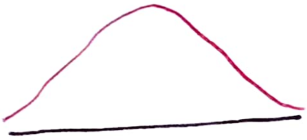
$\sigma^2 = 2$



$\{1, 2, 3, 4, 5, 6, 80\}$

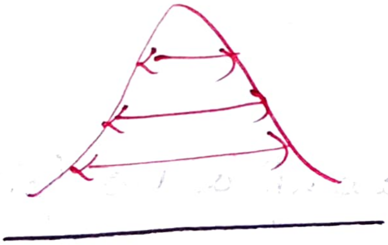
$\mu = 14.4$

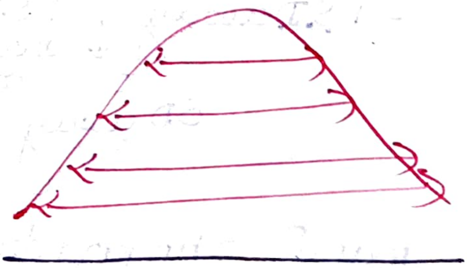
$\sigma^2 = 719.10$



- As variance keeps on inc spread inc

*





σ^2 < σ^2

② Standand dwiation ($-\sqrt{\sigma^2}$)

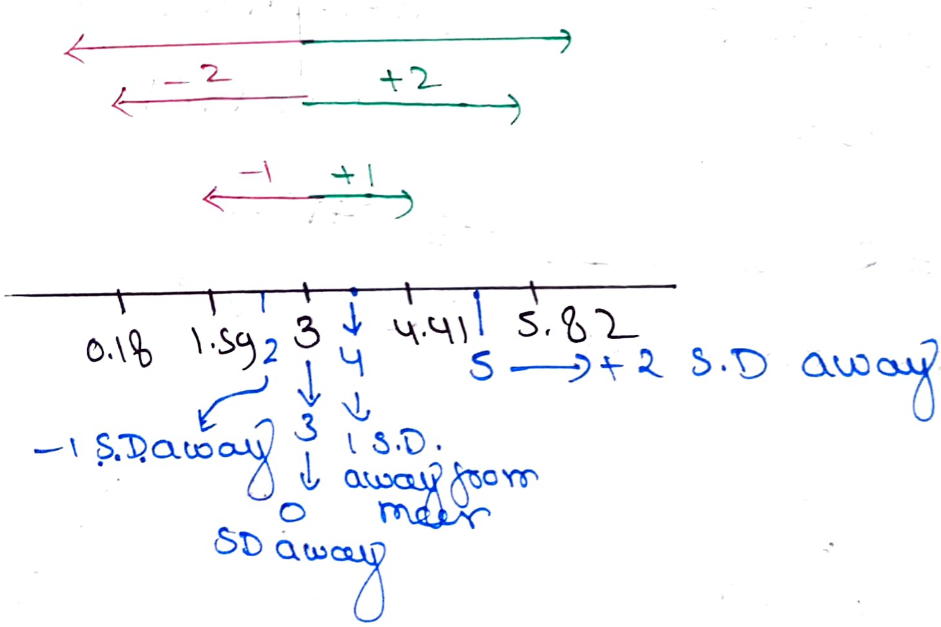
Eg 1, 2, 3, 4, 5

$$\mu = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

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

$$= \frac{2^2+1^2+0+1^2+2^2}{5} = \frac{4+2+4}{5} = \frac{10}{5}$$

$$\boxed{\sigma^2 = 2} \quad \boxed{\sigma = 1.41}$$



→ How many Standand deviation away a no falls from the mean.

* Percentile & Quartiles

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

$$\text{Percentage of even no.} = \frac{\text{no. of Even no}}{\text{total no. of no}} = \frac{4}{8} = 0.5 = 50\%$$

→ Percentile

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

eg

99 percentile → the person has got better marks than 99% of the entire students

eg

$$\text{Dataset} = \overset{\text{0th index}}{\overset{\uparrow}{2}}, 2, 3, 4, 5, \overset{\text{5th index}}{\overset{\uparrow}{\boxed{5}}}, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12$$

- what is the percentile rank of 10

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

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

- Percentile rank of 8

$$= \frac{9}{20} = 45 \text{ percentile}$$

↓

45% of obs are below 8

* 5 numben Summany

① Minimom
② first Quotile [25 percentile] [Q1]
③ Median
④ Third Quootile [75 percntitile] (Q3]
⑤ Maxium

Remove the outlien
⇓
Box plot

eg {1,2,2,2,[3,3]3,4,5,5,5,6,6,6,6,[7,8],8,9,27}

note:- a small no can also be a outlen

$$\boxed{\text{Lowenfence} \Rightarrow Q_1 - 1.5(IQR)}$$

$$\boxed{IQR = Q_3 - Q_1}$$

↓
Inter Quortile Range

$$\boxed{\text{Highen fence} = Q_3 + 1.5(IQR)}$$

[Lowen fence ⟷ Highen fence]

$$Q_1 = \frac{25}{100} * (n+1) = \frac{25}{100} \times 21 = 5.25 \Rightarrow \text{Index} = 3$$

$$Q_3 = \frac{75}{100} * 21 = 15.75 = \frac{8+7}{2} = 7.5$$

$$\text{Lowenfence} = 3 - 1.5(4.5) = -3.65$$

$$\text{Highen fence} = 7.5 - (1.5)(4.5) = 14.25$$

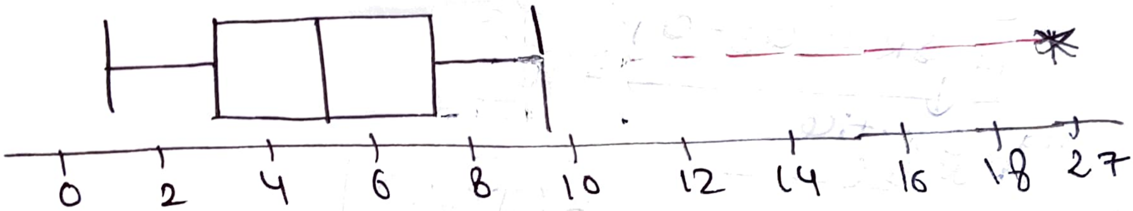
Lowerfence ⟷ Highrfence
= -3.65 14.25

$\underbrace{\hspace{10em}}$

all values should lie b/w
— this

→ 27 not lies b/w fence so remove it

① min = 1
② Q1 = 3
③ median = 5
④ Q3 = 7.5
(5) max = 9



Box plot
↓
— To treat outliers