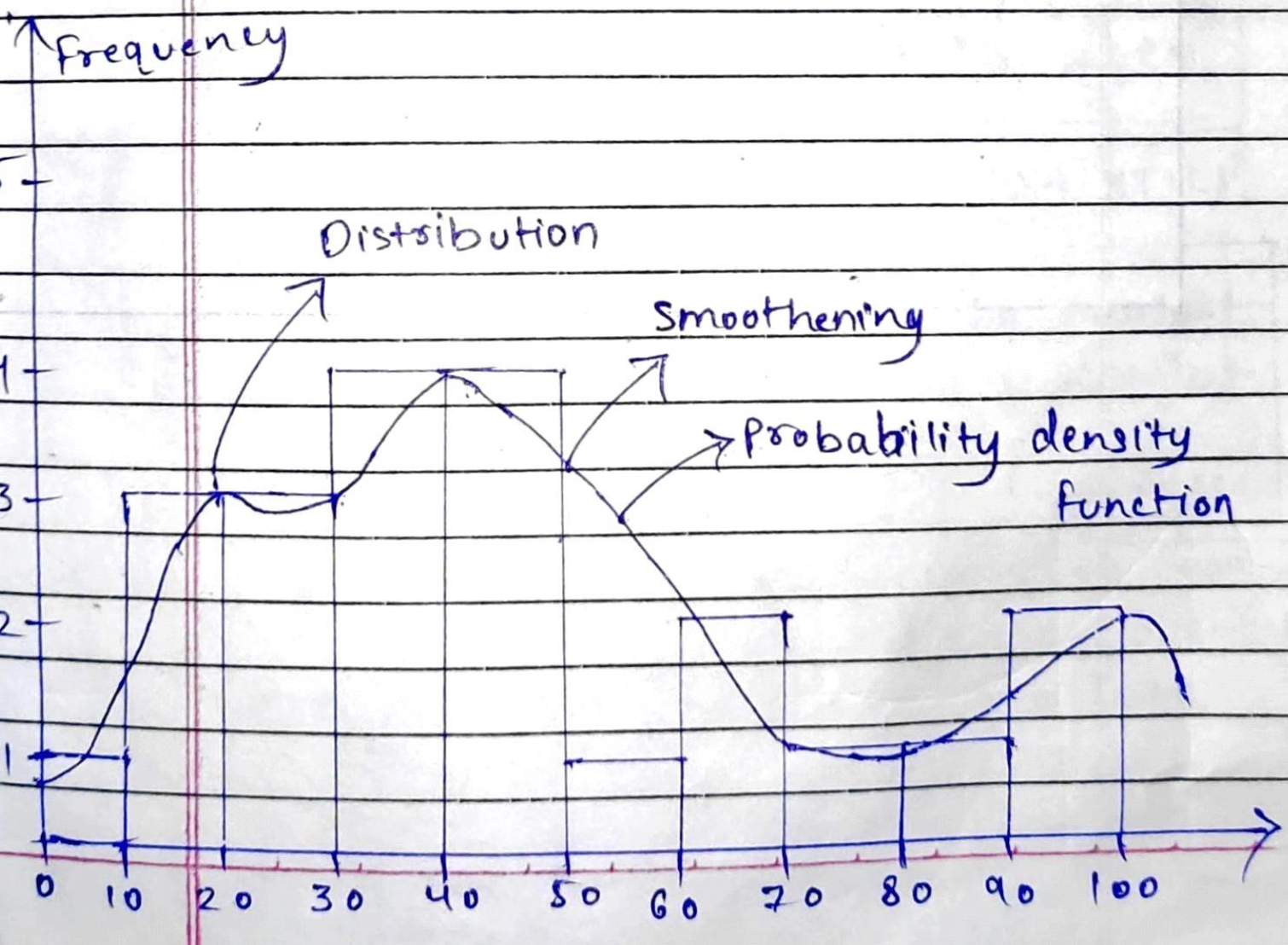
Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37,
40, 44, 42, 43, 50, 51, 65, 68, 78, 90, 95,
100}

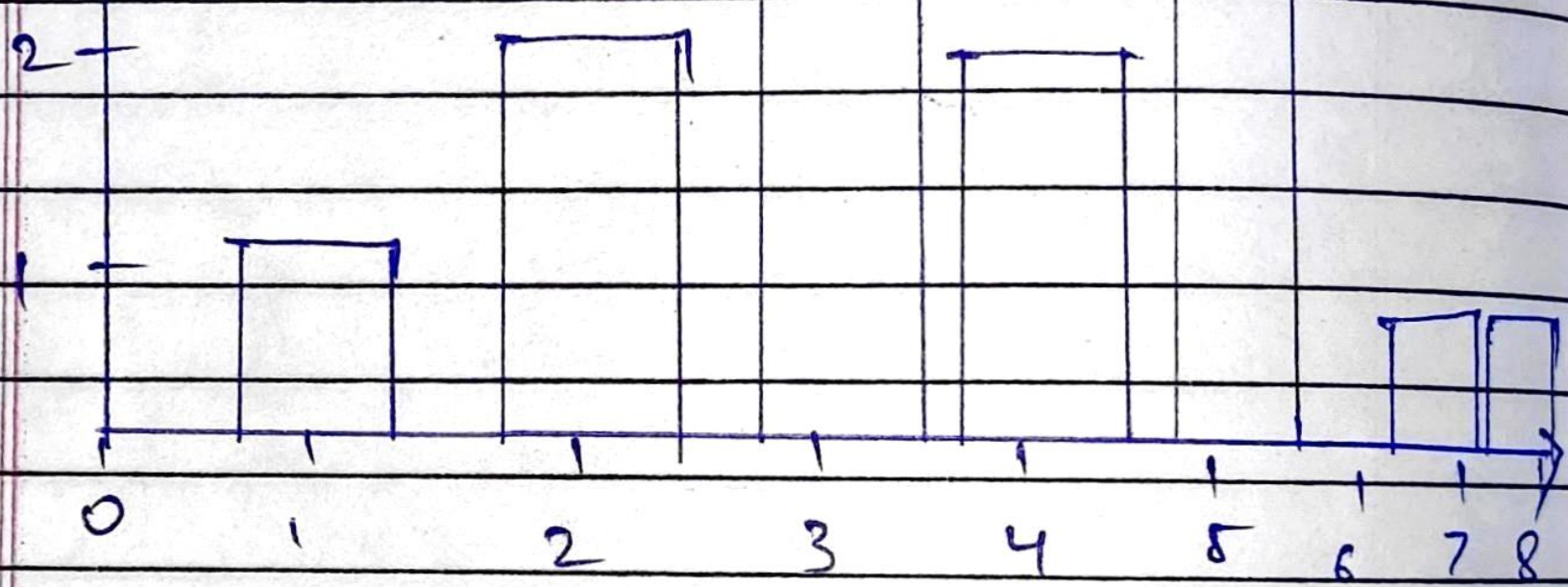
max = 100
min = 10

Bin Size = $\frac{100}{20} = 5$

Bin Size = $\frac{max}{bins}$

Group = bin size
Group size = bin size





Pdf : Probality density function

Pmf : Probability mass function

Continuous
(Pdf)

discrete
(Pmf)

1. Mean
2. Median
3. Mode

{ A measure of CT is a single values that attempts to describe a set of data identifying the central position.

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

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

Population (N) $N >> n$ sample (n)

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

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

$N = 6$ $N > n$

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

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

~~$\bar{x}$ & M~~

$\bar{x}$ & M

Average

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

$\bar{x} = 13.5$

~~$M > \bar{x}$~~ $M > \bar{x}$

~~$\bar{x} > M$~~ $\bar{x} > M$

np.nan ⟹ Null values.

Practical Application (Feature Engineering)

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

NAN → Not a Number

Age	salary (K)
24	45
28	50
29	NAN
NAN	60
31	75
36	80
NAN	NAN

Average (mean)

Age = 29.6 salary = 62.

outliers ⇐ 80 ⇐ 200

(Age) mean = 38 salary = 85

(2) * Median

outliers

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

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

$\frac{1+2+3+4+5+100}{6}$

$= \frac{115}{6} = 19.16$

Steps to find Median:-

(1) Sort the Numbers
(2) Find the central number :-

(i) if the no. of elements are even we find the average of central elements.
(ii) if the no. of elements are odd we find the central elements.

Examples :-

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

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

no impact in median

Mean = 25.6

* if there are no outliers we use mean
* if there is outliers then we use median.

median = 5

(3) Mode :- { Most frequent occuring elements }

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

3

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

Page No.
Date

Dataset { categorical variable }

Types of Flowers:

Lily

Rose ← Most frequent elements.

sunflower NAN = Rose.

NAN ← Rose

Rose

sunflower

Rose

NAN ← Rose

* Measure of Dispersion:

(1) Variance (σ^2)

(2) standard deviation. (σ)

Variance

Population Variance (σ^2)

Example :

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

$\Downarrow$

μ = Variance (Population mean) $= \frac{\sum}{N} = 3$

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

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

$$+ (80 - 14.5)^2 \Big]$$

$$7$$

$$\sigma^2 = 719.10$$

$$\boxed{(\sigma^2) \quad (\sigma^2) \\ 2 \quad < \quad 719.10}$$

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

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

$$\sigma^2 = 2$$

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

No. of even Numbers

Total no. of Numbers

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

Percentiles = Grate, CAT, JEE, SAT

⇓

Percentiles.

Percentile Rank of x = # No. of value below x / n

$\frac{8}{20}$ = 45 percentile

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

$\frac{14}{20}$ = 70 percentile.

* what is the value that consists at 25 percentile.

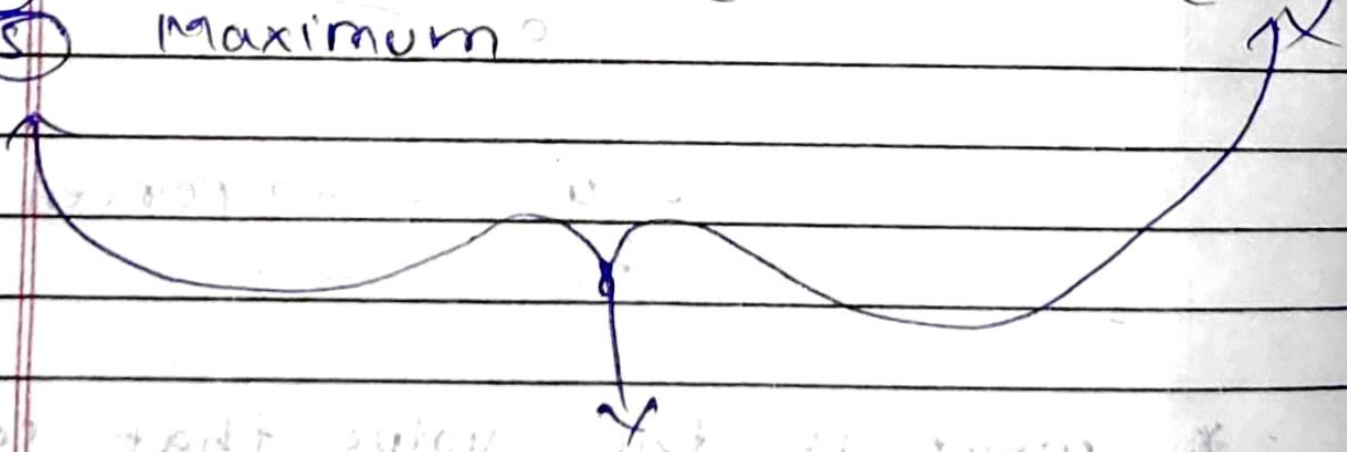
Value = Percentile / 100 * n

100

= 19.95

* 5 number summary :-

1. Minimum
2. First Quartile (25 percentile) (Q_1)
3. Median
4. Third Quartile (75 percentile (Q_3)
5. Maximum



remove the outliers.

⇓

Boxplot

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

Create a set :

lower fence ⟷ Higher fence

lower fence = $Q_1 - 1.5(IQR)$

Higher fence = $Q_3 + 1.5(IQR)$

IQR ⇒ Inter Quartile Range (IQR)

75 25

$IQR = Q_3 - Q_1$

$Q_1 = \frac{25}{100} * (n+1) = \frac{25}{100} * 21 = 5.25 \Rightarrow$ index

$= 3$

$Q_3 = \frac{75}{100} \times 21 = 15.75 \Rightarrow$ index $= \frac{8+7}{2}$

$= 7.5$

Lower fence $= 3 - 1.5(4) = 3 - 6 = -3$

lower fence = -3

Page No.
Date

lower fence = $3 - (1.5)(4.5) = -3.65$

$= \boxed{-3.68}$

Higher fence = $7.5 + (1.5)(4.5) =$

$\boxed{14.25}$

lowerfence $\longleftrightarrow$ Higher fence

$-3.65 \longleftrightarrow 14.25$

1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

Outliers

(1) Minimum = 1

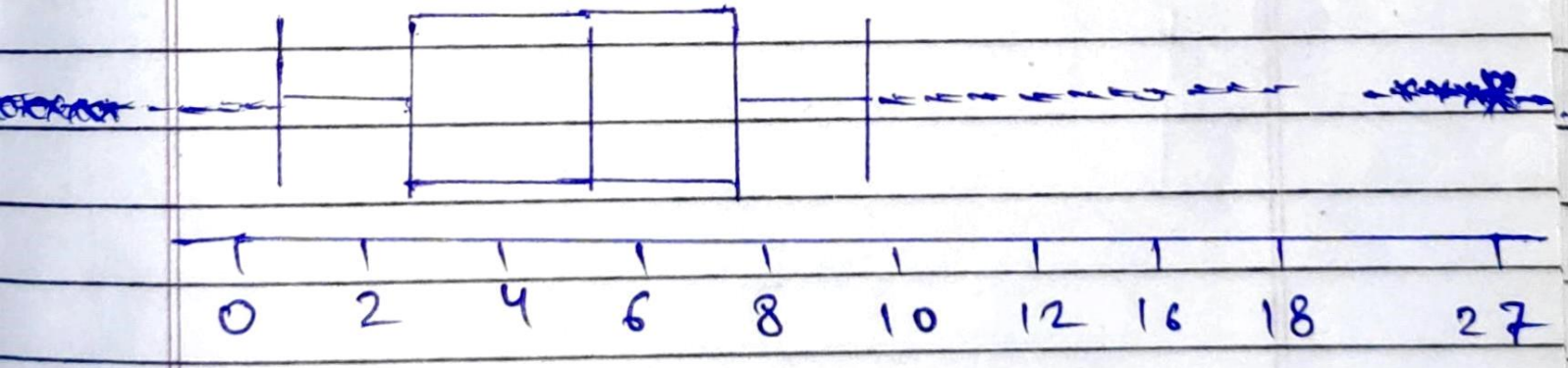
(2) $Q_1 = 3$

(3) Median = 5

(4) $Q_3 = 7.5$

(5) Maximum = 9.

Box plot



To treat Outliers