Statistics Day - 2

Agenda

1. Create histograms
2. Measure of Centre Tendency
3. Measure of disperstion
4. Percentiles And Quartiles.
5. 5 number Summary (Box Plot).

1) Histogram

Ages = {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

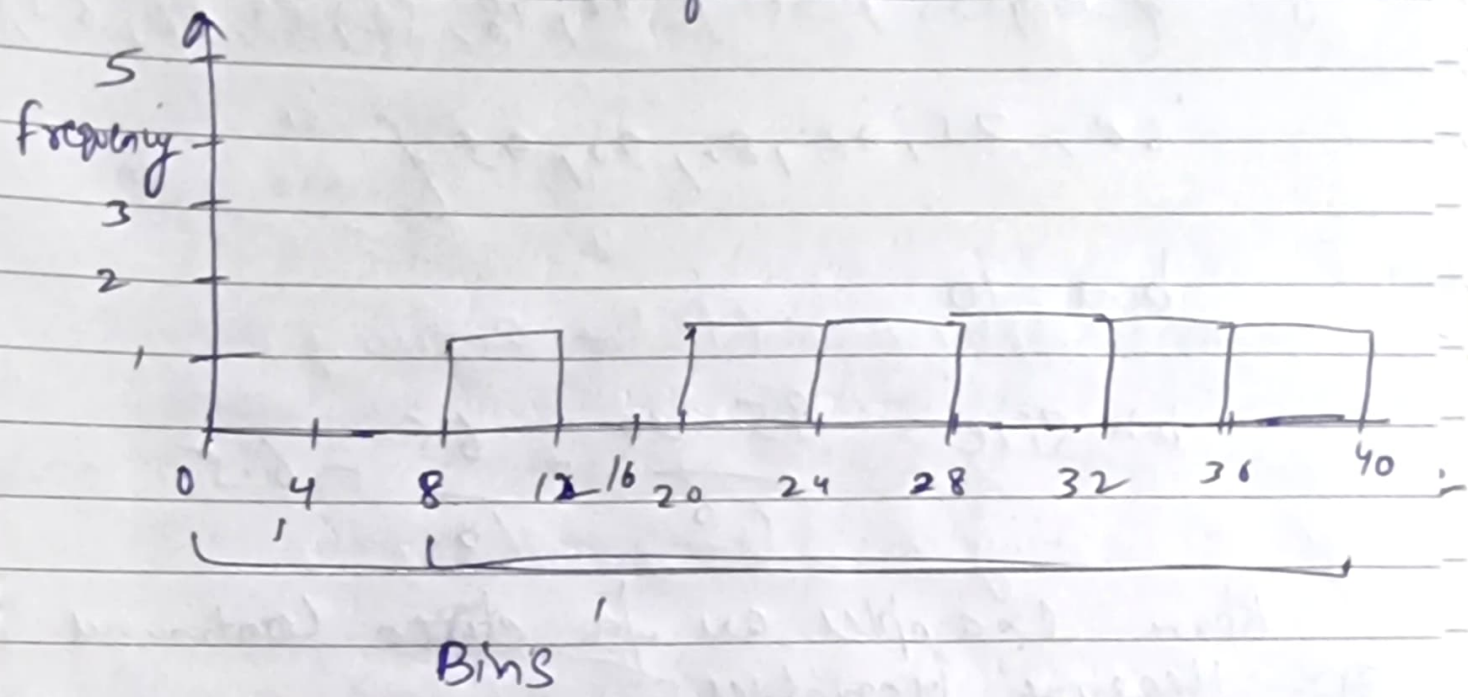
A - Sort the numbers.
B Bins - No. of groups
C Bin Size - Size of bins.

[10, 20, 25, 30, 35, 40] min = 10
max = 40

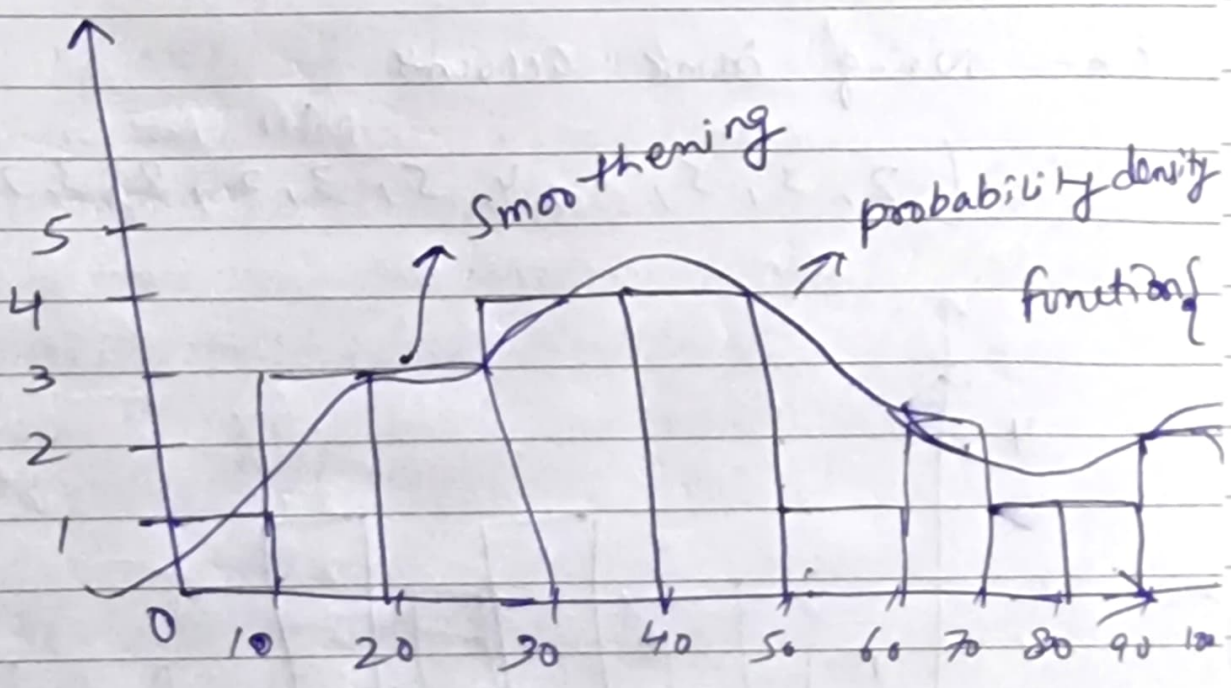
$$\text{Bin Size} = \frac{\text{Max value}}{\text{Bin}}$$

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

Bin Size = 4 for above list then



If Bin Size = $\frac{100}{10} = 10$.



Outliers - w/c looks completely different outside the distribution.

* Median

outlier

{1, 2, 3, 4, 5} → {1, 2, 3, 4, 5, (100)}

$\bar{u} = 3$ $\bar{u} = 19.16$

* Steps to find out median

(1) Sort the numbers
(2) Find the central number

[if the no. of element are even, we find the average of central Elements]

[If the no. of element are odd, we find the central elements.

Ex - Even → {1, 2, 3, 4, 5, 6, 7, 8, 100, 120}

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

odd: {0, 1, 2, 3, 4, 5, 7, 8, 100, 120}

median: 5

Sample age: $\bar{u} = \frac{\sum_{i=1}^{n}}{} = \frac{24+28+29+28}{4}$

$\bar{u} = 13.5$

Practical Application (Feature Engineering)

↓ Age ↓ Salary ↓ Family size

-	-	-
-	-	-
NAN	-	← loss of info
-	-	-
-	-	.
-	Nan	-
-	-	Nan
-	Nan	-
Nan	-	-

NAN - Not a number

Age	Salary
24	45
28	50
29	Nan
Nan	60
31	75
38	80
Nan	Nan

Age: 29.6

Salary: 62

★(2) Measure of Central Tendency

(A) Mean
(B) Median
(C) Mode

Measure of central tendency is the ~~[illegible]~~ single value that attempts to describe a set of data identifying the central position.

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

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

(Center of all Elements)

Population (N) Sample (n)

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

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

Population Age $= \{24, 23, 2, 1, 28, 77\}$

$N = 6$

Population Mean $= (\mu) = \frac{24+23+2+1+28+77}{6} = 17.5$

Weight = {30, 35, 38, 42, 48, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

bins = 10

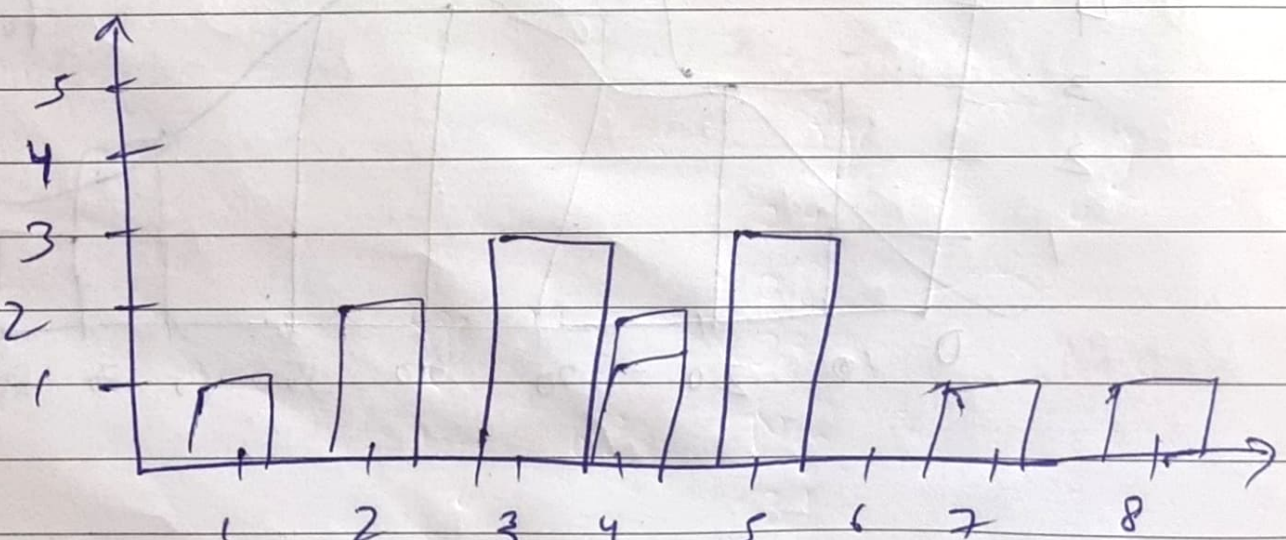
bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$

Above examples are ~~for~~ ~~discrete~~ continuous ~~Measure~~ variables.

#Now - for Discrete Continuous

Ex- No. of bank accounts

= [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



For ~~Smoothing~~ Probability mass functions is used

will discuss later.

No Outliers → Mean
Outliers - Median.

*Mode = {Most frequent occuring elements}

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

Mode = 3

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

[2, 3]

Practical aspect :-

Data set - Types of Flowers

Lily
Sonflowers
Nan
Rose
Sunflowers
Rose
Nan

- In this we can replace Nan with Mode value i.e Rose.

* what is the value that consist of 25 percentile

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

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

(5) 5 number Summary

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

} We Remove the Outliers (Box Plot)

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

[lower Fence ⟷ Higher fence]

Q_1

lower fence = $Q_1 - 1.5(IQR)$

Higher Fence = $Q_3 + 1.5(IQR)$

③ Standard deviation $(\sqrt{\sigma^2}) = \sigma$

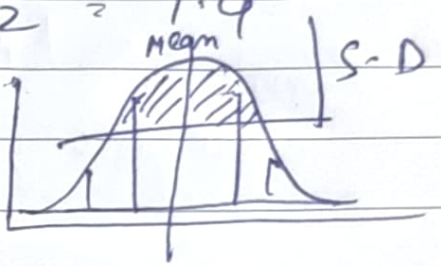
A Statistic that measure the dispersion of a dataset relative to its mean and is calculated by all the Square root of variance

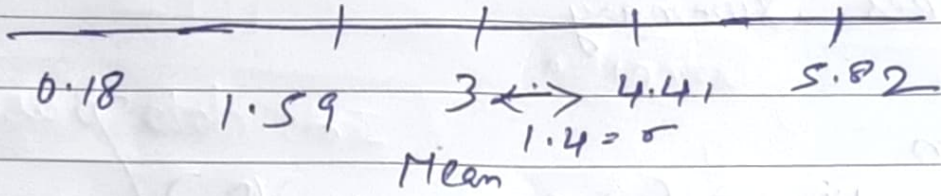
{ 1, 2, 3, 4, 5

$\sigma^2 = 2$

$\sigma = 1.4$

$\sigma = \sqrt{2} = 1.4$





④ Percentile & Quartile.

A percentile is a value below w/c a certain percentage of observation lie.

99 percentile : It means the person that got better marks than 99% of the entire Student

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

- What is the percentile rank of 16

Percentile rank of x = # $\frac{\text{No. of value below } x}{n} = \frac{16}{20} \times 100$

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

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

$$\sigma^2 = (1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + 8$$

$$\frac{(80-(4.4)^2}{7}$$

$$\sigma^2 = 719.10$$

- Dof
- Bessil Loorection.

(3) Measure of Dispersion

(A) Variance (σ^2) ← Spread of Data

(B) Standard deviation (σ) ←

Variance

Population Variance (σ^2)

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

Sample variable (S^2)

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

{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

{1, 2, 3, 4, 50, 60, 70, 100}

{1, 2, 3, 4, 5} → $\mu = 3$

{1, 2, 3, 4, 5, 6, 80} → $\mu = 14.4$

$Q_1 = \frac{25}{100} \times 21 = 5.25$, Index = 3

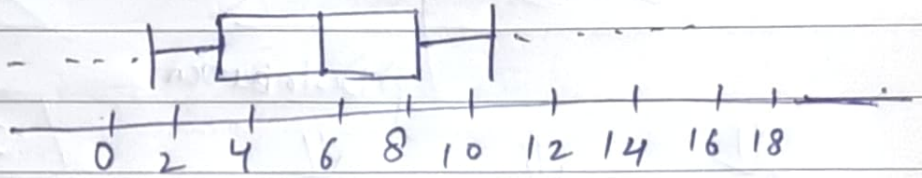
IQR = (Inter Quartile Range)

$= Q_3 - Q_1 = 75 - 25$

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

1. Minimum = 1
2. $Q_1 = 3$
3. Median = 5
4. $Q_3 = 7.5$
5. Max = 9

Box Plot



↓

Used to treat outliers

↓

5 Number Summary

To Calculate Q_3

$Q_3 = \frac{75}{100} \times 21 = 15.75$

Index = $\frac{8+7}{2} = 7.5$

$Q_1 = \frac{25}{100} \times 21 = 5.25$

Index = 3

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

Higher fence = $7.5 + (1.5)(4.5) = 14.25$