DATA SCIENCE
DAY-2 Statistics

AGENDA

1. Histogram
2. Measure of central tendency
3. Measure of Dispersion
4. Percentiles & Quartiles
5. 5 number summary (Box Plot)

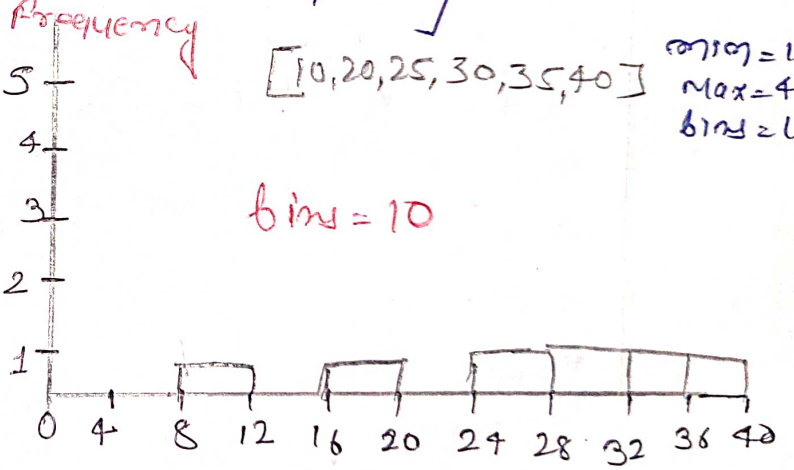
Mr

* Histogram

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

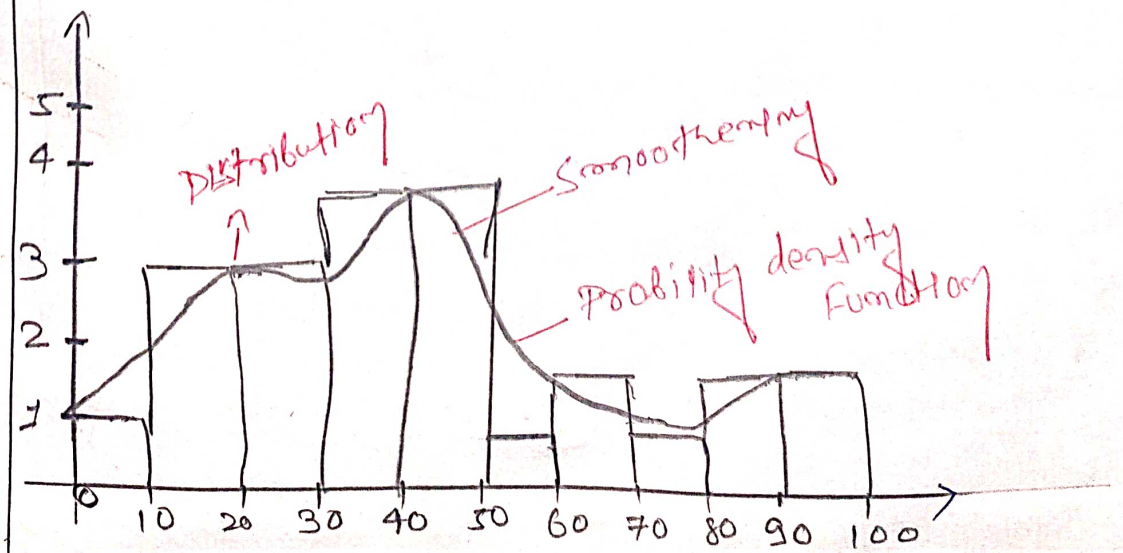
* Sort the numbers
Bins :→ No. of groups
Bin size :→ Size of Bins

? How to choose bins :- 9s totally your choice what you want to choose bin like 10, 5 or 20



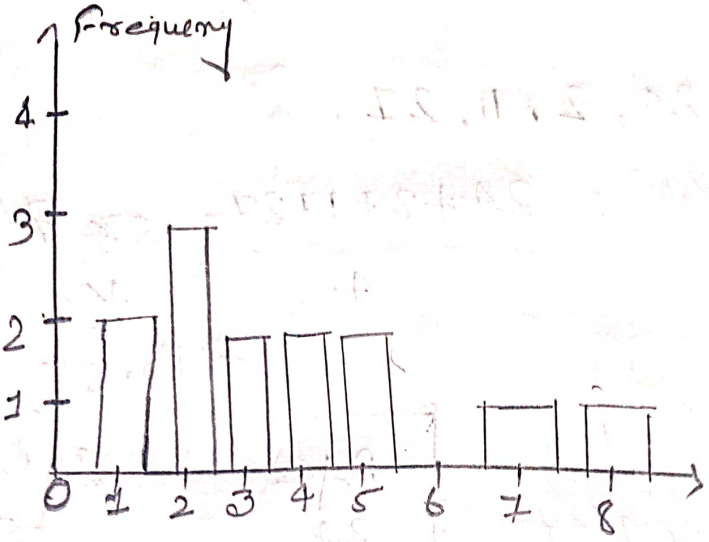
[10, 20, 25, 30, 35, 40]
min = 10
Max = 40
bins = 10

bins = 10



(1) Discrete Continuous

Rank = {1, 2, 3, 5, 7, 8, 3, 2, 4, 5, 1, 2, 4}



{ Probability Mass function

(2) Measure of Central tendency

(1) Mean
(2) Median
(3) Mode

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

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

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

Population (N) $\boxed{N >> n}$ Sample (n)

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

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

Population age = { 24, 23, 2, 1, 28, 27 }

$N = 6$

Population Mean (μ) = $\frac{24+23+2+1+28+27}{6} = 17.5$

Sample age = { 24, 2, 1, 27 }

Sample mean ($\bar{x}$) = $\frac{24+2+1+27}{4} = \frac{54}{4} = 13.5$

$\Rightarrow 13.5$

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

Practical implementation

Age	Salary	Family size
—	—	—
—	NaN	—
NAN	—	—
—	—	—
—	—	NaN

(Replace NAN with mean, Median, Mode According to situation)

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

Mean of Age = $\frac{24+28+29+31+36}{5} = 29.6$

Mean of Salary = $\frac{45+50+60+70+80}{5} = 62$

Median

{1, 2, 3, 4, 5}

$\bar{x} = 3$

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

$\bar{x} = 19.16$

Steps to find out the Median:-

Sort the no.

Find Central Number

→ if the no. of element are even find avg. of two central elements

→ if the no of elements are we simply choose the central Number.

Eg:- for even

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

Mean = 25.6

Median = $\frac{5+6}{2} = \frac{11}{2} = 5.5$

* When there is no outliers → Mean
* When there is outliers → Median

* Mode

→ Most Frequent occuring element

Eg:- {1, 2, 2, 2, 3, 3, 3, 3, 4, 5} or {1, 2, 2, 2, 3, 3, 3, 4, 5}

Mode = 3

Mode = 2, 3

Practical Application

Types of flowers

Lily
Sunflower
Rose
NAN
Rose
Sunflower
Rose

? Replace NAN with Rose ?

⇒ use Mode with categorical variable

Measure of Dispersion

* Variance (σ^2) → Spread of Data
* Standard deviation (σ)

Variance

Population variance (σ^2)

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

Sample variance (s^2)

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

Population Variance (σ^2)

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

{1, 2, 3, 4, 5}

$\mu = 3$

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

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

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

Sample Variance (s^2)

{1, 2, 3, 4, 50, 60, 70, 80}

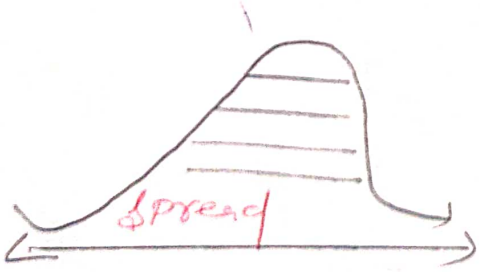
{1, 2, 3, 4, 5, 6, 80}

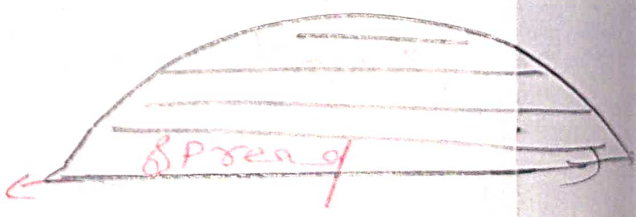
$\mu = 14.4$

$\sigma^2 = [(1-14.4)^2 + (2-14.4)^2 \ldots (80-14.4)^2]$

$$\boxed{\sigma^2 = 719.10}$$

Variance ↑ ⟷ Spread ↑





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

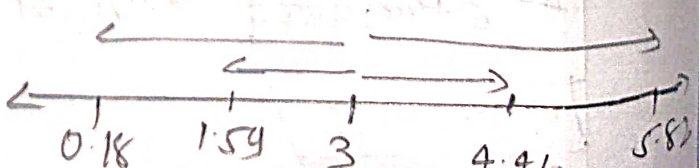
{1, 2, 3, 4, 5}

$\mu = 3$

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

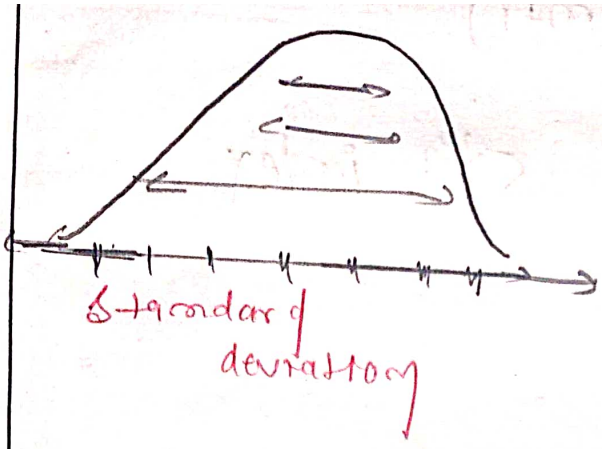
$\sigma^2 = 2$

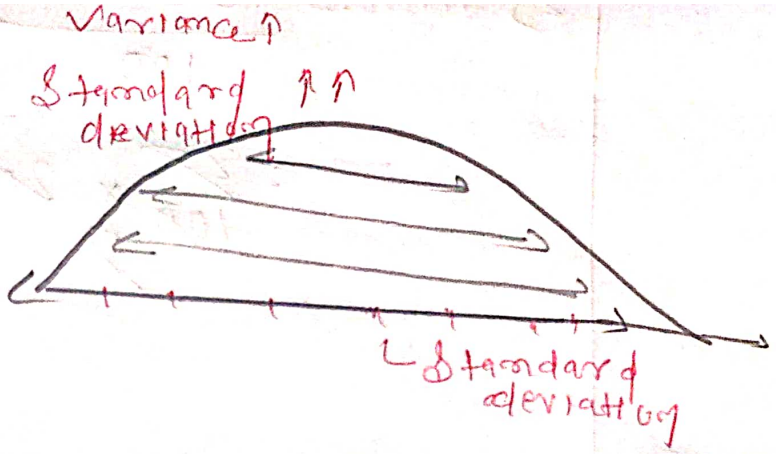
$\sigma = \sqrt{2} = 1.41$



$\mu + \sigma = 3 + 1.41 = 4.41$

$\mu - \sigma = 3 - 1.41 = 1.59$





(4) Percentile & Quartiles

Percentiles :- {1,2,3,4,5,6,7,8,9,10,11,12}

Percentile of even no :- $\frac{\text{No. of to even no.}}{\text{Total No}} = \frac{6}{12} = 50\%$

Defination

:- A Percentile is a value below with a certain Percentage of observation be

99 Percentile ⇒ 9+ means the person has got better marks than 99% of the entire students.

* Rank for same multiple value will be same

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 ?

10 = 10

Percentage Rank = $\frac{\text{No. of value below } x}{n} = \frac{16}{20} = 80\%$

② What is the value that exist at 25% Percentile

⇒ Value = $\frac{\text{Percentile}}{100} * (n+1)$

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

⑤ 5 number Summary

* Minimum
* First Quartile (Q1) 25%.
* Median
* Third Quartile (Q3) 75%.
* Maximum

⇒ Remove the outliers

Box Plot

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

How to Remove outliers

Lower_fence = $Q_1 - 1.5 \times (IQR)$

Higher_fence = $Q_3 + 1.5 (IQR)$

↑ Standard deviation

$IQR = Q_3 - Q_1$

Inter quartile Range

$Q_1 = \frac{25}{100} \times (n+1)$

$= \frac{25}{100} \times (20+1) \Rightarrow 5.25$ ← Index = 5

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

$= \frac{75}{100} \times (20+1) = 15.75$ ← Index = $\frac{8+7}{2} = 7.5$

Lower Fence = $3-(1.5)(4.5) \Rightarrow -3.65$

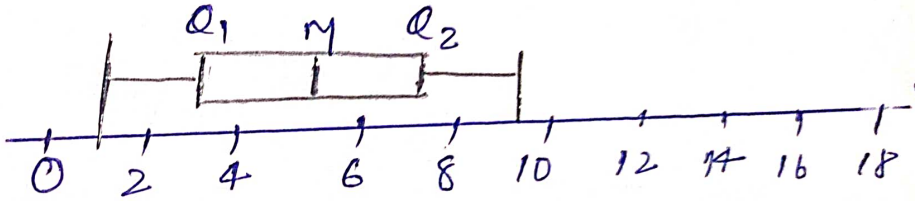
Higher Fence = $7.5+(15)(4.5) = 14.25$

$IQR = 7.5-3 = 4.5$

Box Plot

5 No. Summary

1. Minimum – 1
2. Q_1 – 3
3. Median – 5
4. Q_3 – 7.5
5. Maximum – 9



To treat Outliers