20 Jan/25

Day-2

Page No.
Date

→ 1) Measure of Central Tendency
→ 2) Measure of Dispersion
→ 3) Gaussian Distribution
→ 4) Z score
→ 5) Standard Normal Distribution

⇒

1. Arithmetic mean for Population & Sample :

i) mean (Average)

Population (N)

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Formula = $\bar{x} = \frac{\sum x_i}{N}$

$= \frac{1+1+2+2+3+3+4+5+5+6}{10}$

$= 3.2$

Sample (n)

$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$

$= 3.2$

* Central tendency
  i) mean ii) median iii) mode

→ What is Central Tendency?:
  ⇒ Refers to the measures used to determine the centre of the destribution of data.
  → Center of the data

$X = 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100$

mean = 12

median =
  i) Sort the numbers
  odd number = 11 (central Element)
  median = 3
  median = 3 //

Page No.
Date

Median - {1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100, 112}

$\frac{3^{th} + 4^{th}}{2} = \frac{3+4}{2} = 3.5$

median
mode = 3.5

→ Median works with well outliers.

* Mode :-

{1, 2, 2, 3, 4, 5, 6, 6, 6, 7, 8, 100, 200}

mode = 6 {most frequent Element}

↳ Measure of Central Tendency

Ex:- Dataset :-

Dataset

Type of flower	Petal length	petal width	rows
Rose			10% missing
lilly			
Sun-Flower			

{Missing value → most frequent occuring element}

⇓

Categorical variable

Ages of Students

Asc
25
26
...
...
32
36

mean ? ✓✓✓
median ?
mode ?

Page No.
Date

Measure of dispersion

Dispersion means Spread

1. Variance
2. Standard deviation

=>

1 Variance

Case: {1,1,2,2,4} = 10/5 = 2

Avg {2,2,2,2,2} = 10/5 = 2

both are same but distribution is different. To identify how to distribute are diff on that time may use variance & sd

Population variance

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

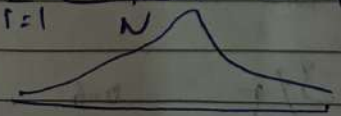
Sample variance

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

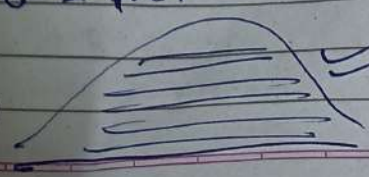
X	M	X − M	$(X-M)^2$
1	2.83	-1.83	3.3489
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.0289
4	2.83	1.17	1.3689
5	2.83	2.17	4.7089
			10.84

M = 14/6 = 2.83

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{N} = 10.84/6 = 1.81$$



i



ii

variance

which is maximum variance

ii → variance more

Page No.
Date

Variance $= \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$

Sample variance $S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$

Sample variance Example.

4, 8, 6.

$\bar{x} = 4+8+6 / 3 = 6$

$(4-6)^2 + (8-6)^2 + (6-6)^2$

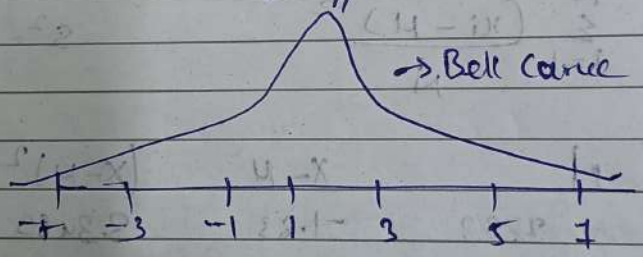
$-2^2 + 2^2 + 0^2$

$4+4 = 8$

$S^2 = \frac{8}{3-1} = 4$ //

Standard deviation $= \sqrt{4}$

$= 2$ //



→ Bell curve ie 2 3 ± 2.

* Percentiles and Quartile { first step to find outliers }

Data - 1, 2, 3, 4, 5

% of the numbers that are odd

% = No of numbers that are odd / Total Numbers

$= 3/5 = 0.6 = 60$ %

Page No.
Date

Percentile :- Definition :-

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

Eg:-

Dataset :- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 12, 12

Q) what is the percentile ranking of 10 ?

Percentile rank of 10.

$x = \frac{\text{No of values below } x}{n} \times 100$

$n = 20$

$= \frac{16}{20} \times 100$. So percentile

Q) What is percentile rank of 11

$x = \frac{\text{No of values below } x}{n} \times 100$

$= 17/20 \times 100 = 85\%$.

Q) What value ends at percentile ranking of 25th ?

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

$= \frac{25}{100} \times (20+1)$

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

This 5.25 is the index position

Page No.
Date

We take 5th and 6th Index
we do avg

$5+5/2 = 10/2 = 5$

5 → 25% (percentile)

(Q) what is 75%. ?

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

$\frac{75}{100}(21) = 15.75$ (Index value

Then we take 15 and 16 Term

$9+9/2 = 18/2 = 9$ //

* ⇒ Five Number Summary :-

i) Minimum
ii) First quartile (Q_1)
iii) median
iv) Third quantile (Q_3)
v) maximum

⇒⇒ Removing the outliers :-

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

27 is the outliers //

{Lower Fence ⟷ Higher Fence}

IQR : Inter Quartile

Page No.
Date

Lower Fence = $Q_1 - 1.5\,(IQR)$
Upper Fence = $Q_3 + 1.5\,(IQR)$

$IQR = Q_3 - Q_1$

Eg:- $Q_3 = 75\%$ → $75/100 \times (19+1) = 15$ (Index) = 7th
$Q_1 = 25\%$ → $25/100 \times (19+1) = 5$ (Index) = 3rd

$IQR = Q_3 - Q_1$
$= 7 - 3$
$= 4$

Now,
Lower Fence = $Q_1 - 1.5\,(IQR)$
$= 3 - 1.5\,(4)$
$= -3$
Higher Fence = $Q_3 + 1.5\,(IQR)$
$= 7 + 1.5(4)$
$= 7 + 6$
$= 13$

[Lower Fence ⟷ Higher Fence]
[-3 ⟷ 13]

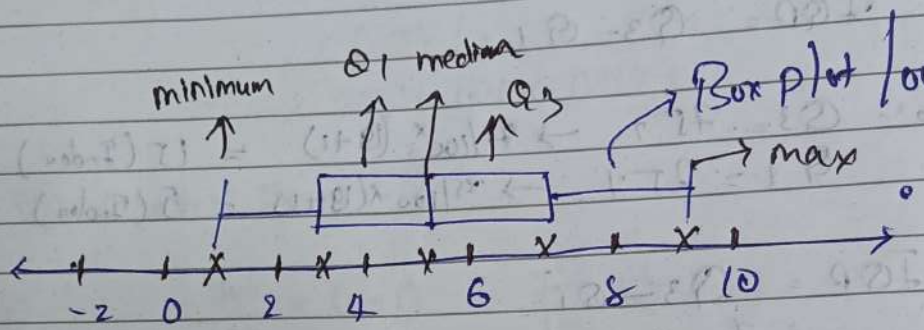
Outlier 13 & 27 value
27 outlier

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

Minimum = 1
Q_1 = 3
median = 5
Q_3 = 7
max = 9

5 Number Summary

Box plot



⇒ this use for data visualization

⇒ Variance :

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