Page No.
Date

Session-33

Descriptive Stats-2.

- Quantiles
  - we divide our data (eg age column) into equal size bucket.
  - They can be used to detect outliers.

- Types of quantiles -

i) Quartile → Divide data into 4 equal parts
  → 25th percentile ---- 100 per-

ii) Deciles → Divide data into 10 equal parts.

iii) Percentiles → Divide data into 100 "

iv) Quintile → " " " 5 "

Things to remember -
→ Data should be sorted from low to high
→ 25th or any ~~per~~ tile may not be present in date

- Percentile -
  - 75 per- means 75% people are behind you.

formula → $PL = \frac{P}{100}(N+1)$

PL → desired percentile locatn
P → Percentile rank
N → total no. of observatn

Q. calculate 75^{th} percentile
78, 82, 84, 88, 91, 93, 94, 96, 98, 99

Sol
Sort data = 78, 82, 84, 88, 91, 93, 94, 96, 98, 99
1 2 3 4 5 6 7 8 9 10

$PL = \frac{75}{100}(10+1) = 8.25$

means b/w 8 & 9

$\Rightarrow 96 + 0.25(98-96) = 96.5$

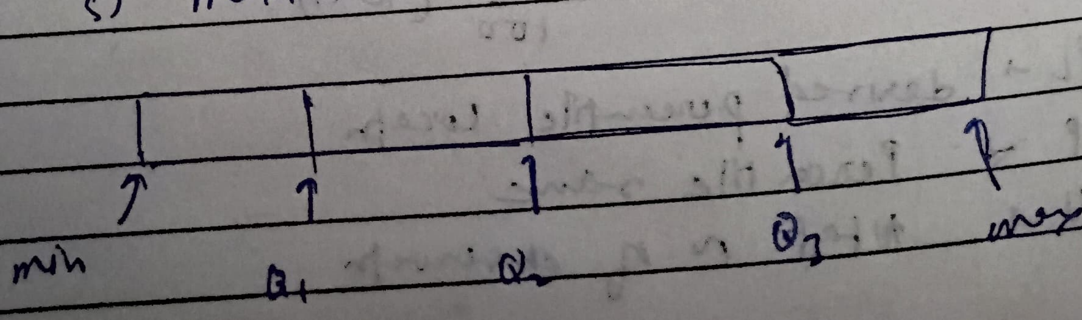
- Percentile of a value -

$$= \frac{X + 0.5Y}{N}$$

X = no. of values below given values
Y = " " equal to " "
N = total no. of values

Q. 88 $\Rightarrow \frac{3 + 0.5 \times 1}{10} = 0.35$
= 35 percentile

- 5 number Summary -
1) min. value
2) first quartile (Q1)
3) median (Q2)
4) Third quartile (Q3)
5) maximum value

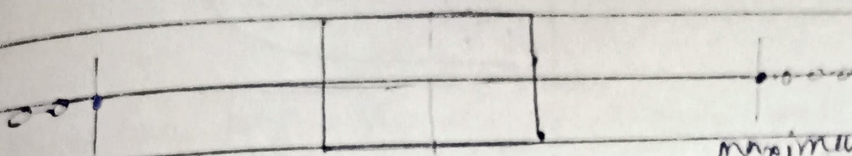


· Interquartile Range (IQR)

→ $Q_3 - Q_1$ (middle 50 %)

· Boxplots ·

IQR



min. Q_1 mean (Q_2) Q_3 maximum

$(Q_1 - 1.5 * IQR)$ $(Q_3 + 1.5 * IQR)$

6	213	241	260	261	290	314	321	350	1800
1	2	3	4	5	6	7	8	9	10

$Q_2 = \frac{50}{100} \times 11 = 5.5 = 285.5$

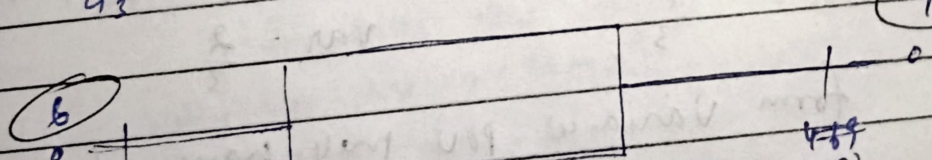
$Q_1 = \frac{25}{100} \times 11 = 2.75 = 213 + 0.75(241 - 213)$

$= 234$

$Q_3 = 328.25$

1800

6



~~469~~ 350

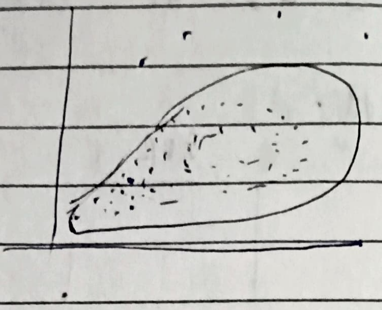
213 234 (Q_1) 328.25 (Q_3)

min = ~~$Q_1 - IQR$~~ $= Q_1 - 1.5 \times IQR = 93$

max = $Q_3 + 1.5 \times IQR = 469$

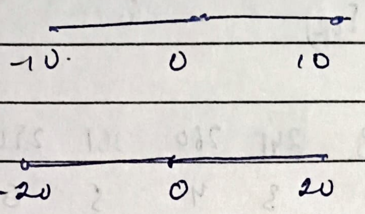
since 93 is not present so we will stop at 213 (acc. to data)

· Scatter plot



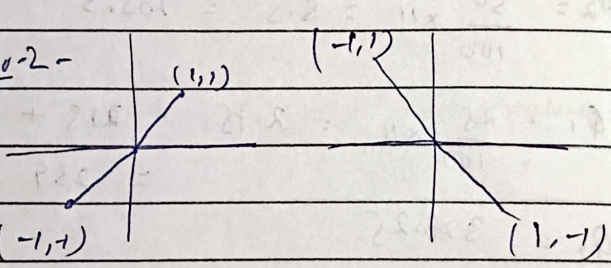
· Covariance -

Scenario-1 :-



form mean POV they have not any diff.

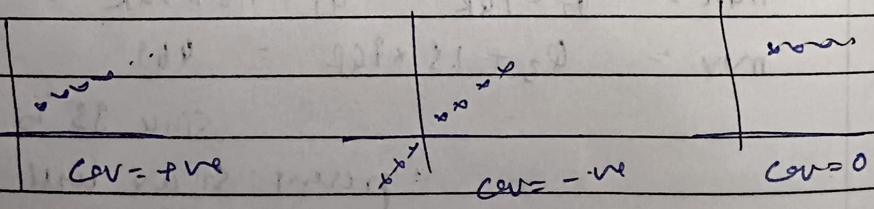
Scenario-2 -



var = $\frac{2}{3}$ var = $\frac{2}{3}$

form Variance POV they have not any diff.

· Covariance => The type of relationship b/w two numerical column



one col ↑ then another col ↑

one col ↑ then another col ↓

No relationship

Page No.
Date

- formula →

Population : $\dfrac{\Sigma (X-\mu_x)(Y-\mu_y)}{N}$

Sample : $\dfrac{\Sigma (X-\bar{X})(Y-\bar{Y})}{n-1}$

- Disadvantage cov → It doesn't tell about the strength of relation

⇒ Covariance of a variable with itself is variance

$$cov = \frac{\Sigma (X_i-\bar{X})(Y_i-\bar{Y})}{n-1} = \frac{\Sigma (X_i-\bar{X})(X_i-\bar{X})}{n-1}$$
$$\Rightarrow \frac{\Sigma (X_i-X)^2}{n-1}$$

- Correlation –

It tells the degree to which two variables are related.

→ ranges [-1, 1]

- 1 ⇒ perfectly +ve relation
- more towards 1 means more +vely related

$$Corr = \frac{cov(x,y)}{\sigma_x * \sigma_y}$$

→ Correlation doesn't imply causation.