Day-4

ML with Python

Statistics

Types
- Descriptive
- Inferential

example- Screw manufacturing
1M
500 → quality (sample)
which party will win the election:
500 to 1000 → Sample
1cr → population

Descriptive → 1000 row or 1M → 20
Inferential sample → expt → infer about population

Descriptive ⇒
1. measures of central tendency
2. Measures of dispersion
3. Covariance correlation

3m's (mean, median, mode)
→ Range, quartiles, std, variance.

किफायती
Date
Pages
Rituraj

Mean

① 5 4 2 1 3 → 15/5 = 3

② 500 4 2 5 3 1 → 515/6 = 85.83 → 500 ⇒ outlier (extreme)

③ 1 2 3 → 6/3 = 2

M F M F ⇒ ? mean

Data → Missing values (Blank, ?, 0)

Numerical ⇒ missing values → mean
if there are outliers

categorical ⇒

Remove outliers → mean
Median → outliers

Median

rules ⇒ ① Sort it
② n (no. of items)
③ n is odd ⇒ median = $\left(\frac{n+1}{2}\right)^{th}$ term
④ n is even ⇒ median = $\frac{1}{2}\left(\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\right)$ term

5 3 1 4 2

1 2 3 4 5

n = 5 (odd)

$\frac{5+1}{2}$ = 3rd term

median = 3

500 4 2 1 3 5 = 1 2 3 4 5 Sorted

n = 6 (even)

$\frac{1}{2}\left[\left(\frac{6}{2}\right)^{th} + \left(\frac{6}{2}+1\right)^{th}\right)$

$\frac{1}{2}(3^{rd} + 4^{th})$

$\frac{3+4}{2} = 3.1$

Mode

G O O R R O G G G mode = G

1 2 3 3 3 2 3 mode = 3

Date
Pages
Rituraj

* Measures of Dispresion

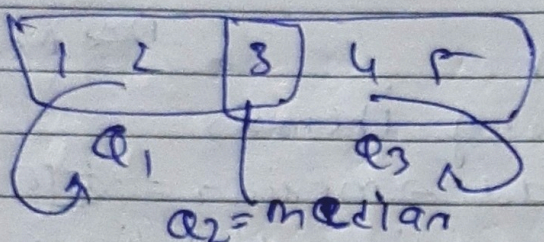
* Range → Max - Min

Age = 1, 2, 3 . . . 20

20 - 1 = 19

* 2, 5, 1, 3, 4

Quartiles



① Short it
② Median
③ Q_2 = median
④ Exclude

$Q_1 = 2$
$Q_2 = 3$
$Q_3 = 4$

* Outliers ⇒

$IQR = Q_3 - Q_1$

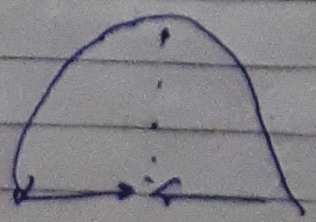
$< Q_1 - 1.5 * iqr$
$> Q_3 + 1.5 * iqr$

$IQR = Q_3 - Q_1 = 4 - 2 = 2$

$< 2 - 1.5 * 2 = -1$ $\quad < -1$

$> 4 + 1.5 * 2 = 7$ $\quad > 7$

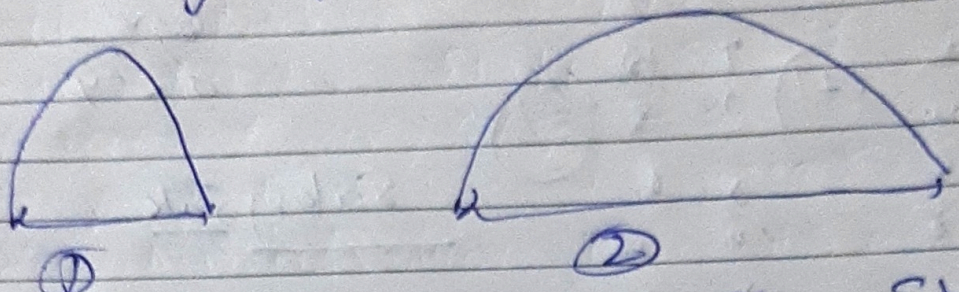
* Central Tendency ⇒ 1. Mean, Median, Mode } Center



Normally distributed

Mean = Median = Mode

measure of spread



② — more Std deviation
more Variance

Covariance & correlation

x	y	$(x-\bar{x})(y-\bar{y})$
1	2	(1-3)(2-6)
2	4	(2-3)(4-6)
3	6	(3-3)(6-6)
4	8	(4-3)(8-6)
5	10	(5-3)(10-6)

$$\text{Covariance} = \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$$

$$\text{correlation} = \frac{\text{Covariance}}{\sigma_x \times \sigma_y}$$

$\Rightarrow$ Std deviation

mean = 15/5

= 4

x	y
1	10
2	8
3	1
4	4
5	2
	4

mean

Cover

Temp ↓ Fever Cases ↑

Temp ↑ Ice cream Sale ↑

किफायती
Date
Pages
Rituraj

Correlation

X	Y
1	2
2	4
3	6
4	8
8	16

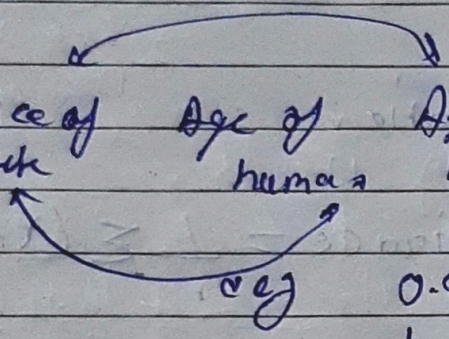
(+1)

X	Y
1	10
2	8
3	6
4	4
8	2

(-1)

cov
σ_{x-y}

Salary Sale 3L



eg 0.02

-0.8