Standard Deviation

Data Mining

lec8 # Distribution Question (Mid & final)

Rules 8

Sample   Population

$$\text{mean} = \overline{X} = \frac{\Sigma xi}{n} \qquad\qquad = \mu = \frac{\Sigma xi}{n}$$

$$\text{variance} = S^2 = \frac{\Sigma (xi - \overline{X})^2}{n-1} \qquad\qquad = \sigma^2 = \frac{\Sigma (xi - \mu)^2}{n}$$

SD

$$\text{Standard deviation} = S = \sqrt{S^2} \qquad\qquad = \sqrt{\sigma^2} = \sigma$$

$$= \sqrt{\frac{\Sigma (xi - \overline{X})^2}{n-1}} \qquad\qquad = \sqrt{\frac{\Sigma (xi - \mu)^2}{n}}$$

Mode → أكثر أرقام مكرره في الأرقام الموجودة

Median/Medium → الرقم أو الرقمين الـ في النص بعد يترتب   Avg

Range = Maximum No – Minimum No.

Ex8 Sample ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩
{20, 19, 5, 0, 7, 9, 9, 15, 15, 12}

① Mode = 9, 15

② Medium   0
   5
   7
   9
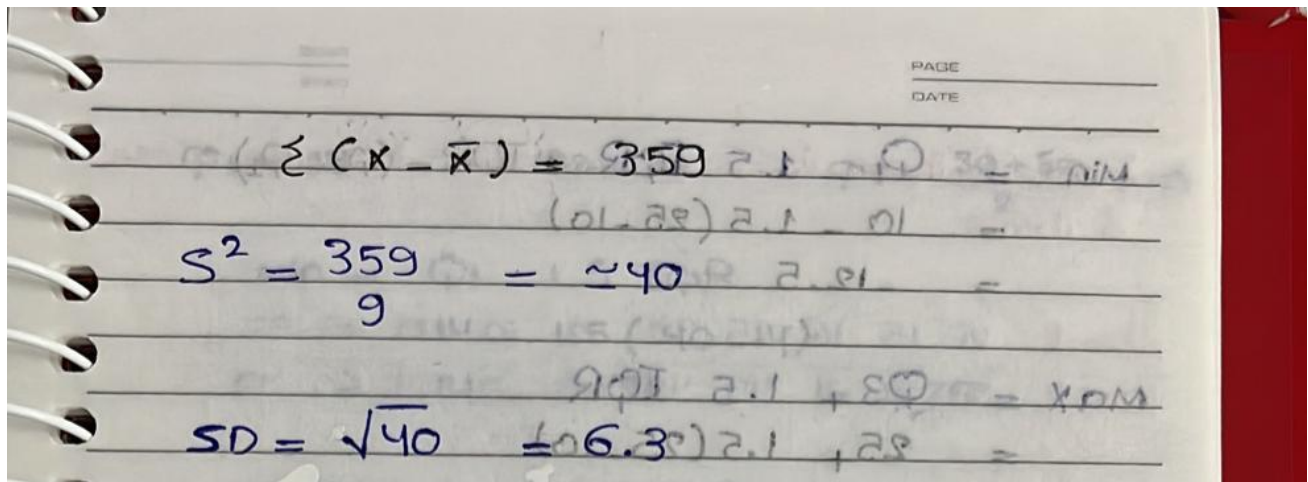   → 9
   → 12
   15
   15
   19
   20

$$\frac{9+12}{2} = \frac{21}{2} = 10.5$$

③ Mean = $\frac{111}{10}$ = 11.1 ≈ 11

④ Range = 20 − 0 = 20

⑤ variance = $\frac{\sum(x - \bar{x})^2}{n-1}$

$(x - \bar{x})$ $(x - \bar{x})^2$

0 − 11 = $(-11)^2$ = 121
5 − 11 = $(-6)^2$ = 36
7 − 11 = $(-4)^2$ = 16
9 − 11 = $(-2)^2$ = 4
9 − 11 = $(-2)^2$ = 4
12 − 11 = $(1)^2$ = 1
15 − 11 = $(4)^2$ = 16
15 − 11 = $(4)^2$ = 16
19 − 11 = $(8)^2$ = 64
20 − 11 = $(9)^2$ = 81

$$\Sigma (x - \bar{x}) = 359$$

$$S^2 = \frac{359}{9} = \simeq 40$$

$$SD = \sqrt{40} = 6.3$$

Boxplot Percentile

Boxplot_Presentile

Five number Summary.
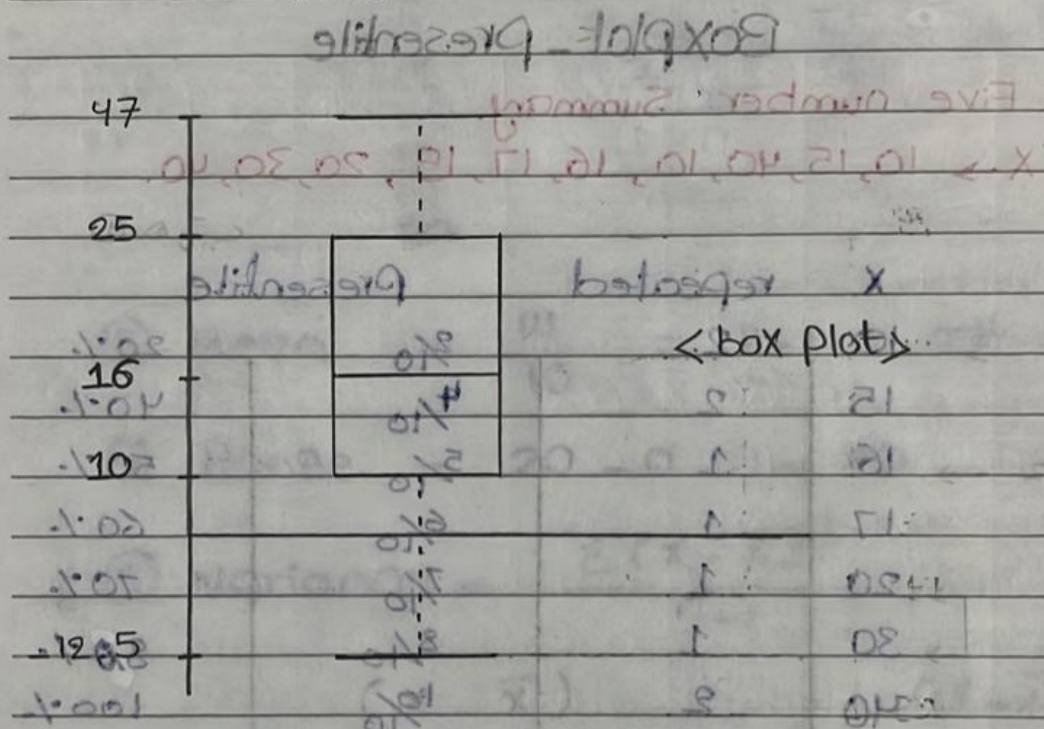
$X \rightarrow$ 10,15,40,10, 16, 17, 15, 20,30, 40.

① هرتب

	X	repeated		Presentile		
$Q_1 \rightarrow$	10	2		$2/10$		20%
	15	2		$4/10$		40%
$Q_2 \rightarrow$	16	1		$5/10$		50%
	17	1		$6/10$		60%
	20	1		$7/10$		70%
Q_3	30	1		$8/10$		80%
	40	2		$10/10$		100%

Questions

$Q_1 = 10$

$Q_2 = 16$

$Q_3 = \dfrac{20+30}{2} = 25$

$Q_1 \rightarrow 25\%$

$Q_2 \rightarrow 50\%$

$Q_3 \rightarrow 75\%$

$$Min = Q_1 - 1.5 \; IQR \qquad IQR = (Q_3 - Q_1)$$
$$= 10 - 1.5 (25 - 10)$$
$$= -12.5$$

$$Max = Q_3 + 1.5 \; IQR$$
$$= 25 + 1.5 (25 - 10)$$
$$= 47.5$$

Question 8   10, 14, 15, 15, 15, 20, 30, 50, 90, 120

	value	count	fraction	percent
	10	1	$\frac{1}{10}$	10 %
$Q_1 \rightarrow$	14	1	$\frac{2}{10}$	20 %
$Q_2 \rightarrow$	15	3	$\frac{5}{10}$	50 %
	20	1	$\frac{6}{10}$	60 %
	30	1	$\frac{7}{10}$	70 %
$Q_3 \rightarrow$	50	1	$\frac{8}{10}$	80 %
	90	1	$\frac{9}{10}$	90 %
	120	1	$\frac{10}{10}$	100 %

$$\phi_1 = 14 \qquad \phi_2 = 15 \qquad \phi_3 = \frac{30+50}{2} = 40$$

$$\text{Min} = \phi_1 - 1.5 \; IQR$$

$$= 14 - 1.5(40-14)$$

$$= -25$$

$$\text{Max} = \phi_3 + 1.5 \; IQR$$

$$= 40 + 39$$

$$= 79$$

Pearson Correlation Coefficient

$$\frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

16/100 → 0.16

15/100 → 0.15

Correlation

Q Find if There is any Correlation between These 2 items.

Item1	16	15	16	25	20	22	20	40	45	35
Item2	30	25	31	45	40	45	30	54	60	49

answer :- n = 10

Item1 (x)	item2 (y)	xy	x^2	y^2
16	30	480	256	900
15	25	375	225	625
16	31	496	256	961
25	45	1125	625	2025
20	40	800	400	1600
22	45	990	484	2025
20	30	600	400	900
40	54	2160	1600	2916
45	60	2700	2025	3600
35	49	1715	1225	2401
Sum = 254	= 409	= 11441	= 7496	= 17953

Pearson Correlation Coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

$$r = \frac{(10 * 11441) - (254)(409)}{\sqrt{[10 * 7496 - (254)^2][10 * 17953 - (409)^2]}}$$

$$r = 0.9304586776$$

Normalization

Minmax:

Normalization:

1. Min Max Scaller. Scale $(-3 \to 3)$

$$v` = \frac{v - min\,A}{max\,A - min\,A} \cdot (new\,max\,A - new\,min\,A) + new\,min\,A$$

$$v`_1 = \frac{16-15}{45-15}\left(3-(-3)\right)+(-3)) = -2.8$$

$$v`_2 = \frac{15-15}{45-15}\left(3-(-3)\right)+(-3)) = -3$$

$v`_3 = -2.8$ (16)	(22) -1.6	(45) 3
$v`_4 = -1$ (25)	(20) 2	(35) 1
$v`_5 = -2$ (20)	(40) 2	

Decimal Scaller:

2. Decimal Scaller.

① biggest Number. digits Number. Divide

$45 \to 00$ ✓

$16/100 \to 0.16$

$15/100 \to 0.15$

Z-Score

3. Z - Score

$$\mu = \frac{\Sigma Xi}{n} = \frac{260}{8} = 32.5$$

$$SD = \sqrt{\frac{\Sigma(V-\mu)^2}{n}}$$

X
20
15
30
40
50
45
35
25

$$= \sqrt{\frac{1050}{8}}$$

X	V−μ	(V−μ)²
20	−12.5	156.25
15	−17.5	306.25
30	−2.5	6.25
40	7.5	56.25
50	17.5	306.25
45	12.5	156.25
35	2.5	6.25
25	−7.5	56.25
	Sum =	1050

① $\frac{20 - 32.5}{12.247449} = -1.020620702$ & so on

QQPlot

Example:

	Percentile %
$Q_1 = 10$	20 %
$Q_2 = 16$	50 %
$Q_3 = 25$	75 %

Values

50
40
30
20
10

Percentile.

10% 20% 30% 40% 50% 60% 70% 80% 90%

Prices are in range 20 – 80.

Values

50
40
30
20
10

Percentile.

10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

QQ plot:

→ Prices of Amazon are higher. because max is higher.

→ Amazons' values are less Than Google.

→ Google Prices are equal / same.

→ Both have Same accepted minimum Values.
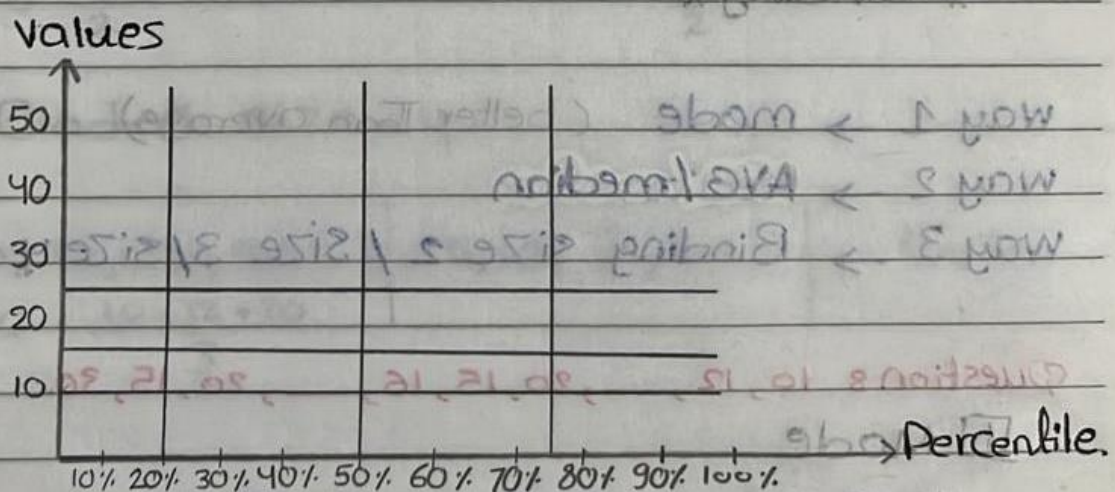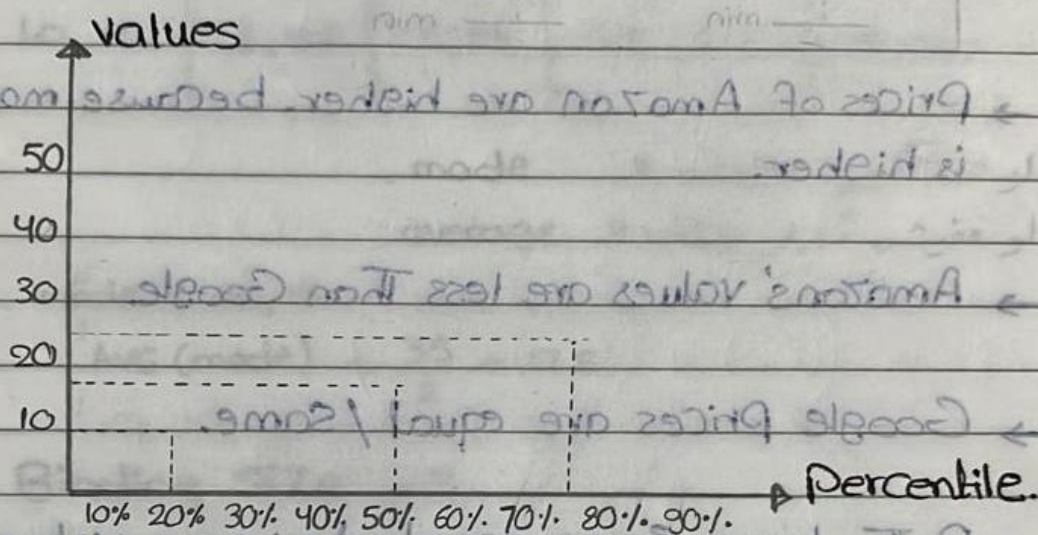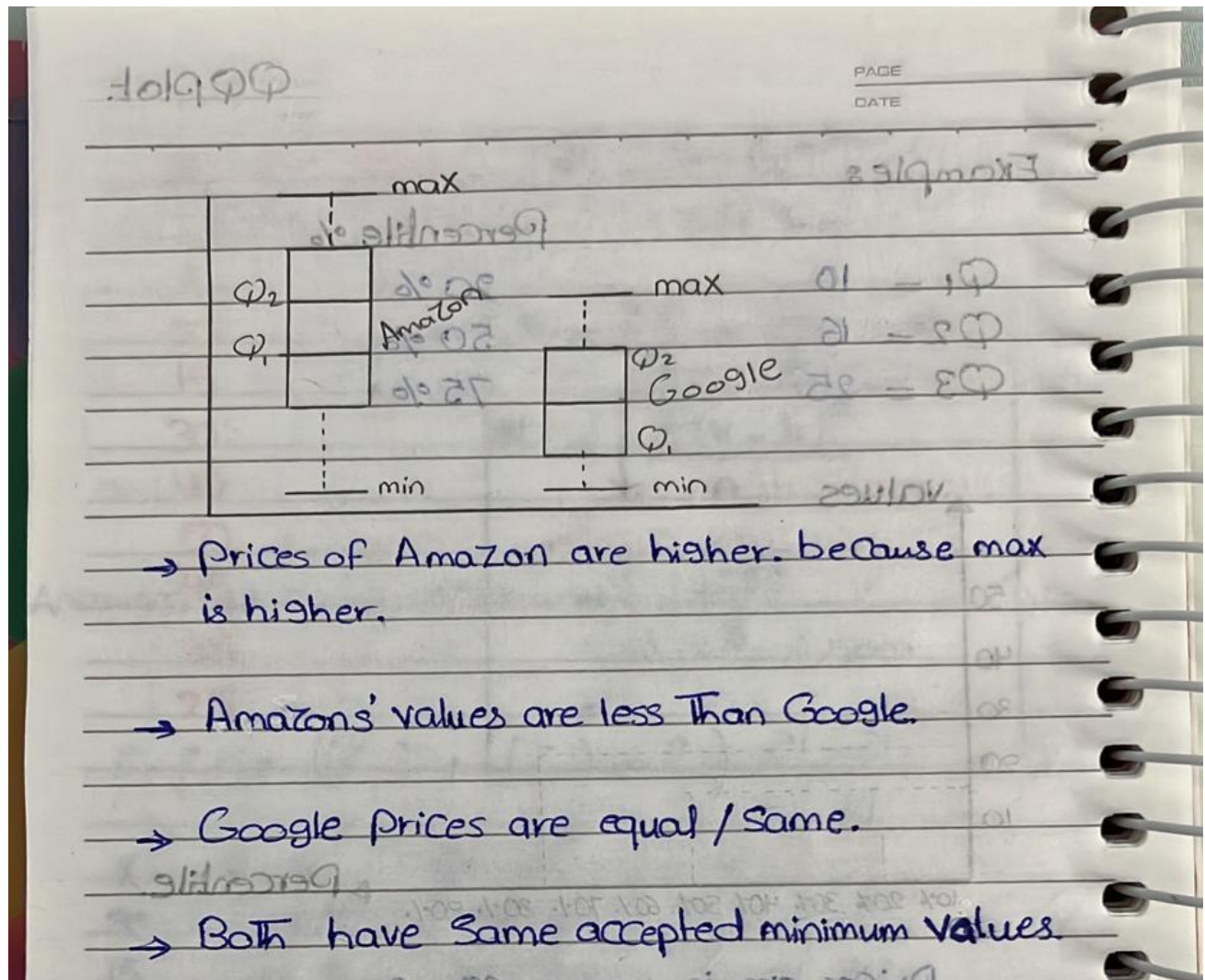
Binning

way 1 → mode (better Than average)

way 2 → Median

way 3 → Binding size 2 / size 3/ size 4

Question 8 10, 12, ___, 20, 15, 16, ___, 20, 15, 30.

[1] Mode

10, 12, 20, 20, 15, 16, **20**, 20, 15, 30.

[2] Median.   10, 12, ___, 15, 15, 16, ___, 20, 20, 30

$$\frac{15+16}{2} = 15,5.$$

[3] Binding Size = 2.

10, 12	17.5, 20	15, 16	17.5, 20	15, 30

→ لو فى تكرار بناخد ال mode.

→ لو مفيش تكرار بناخد ال average.

mode = 20, 15

$$AVG\ (mode) = \frac{35}{2} = 17.5$$

Binding Size = 3

10, 12, 11	20, 15, 16	17.5, 20, 15	30
AVG $\frac{22}{2}$		$\frac{35}{2}$	

Binding size = 4

10, 12, 14, 20	15, 16, 17, 20	15, 30
AVG $\frac{10+12+20}{3}$		