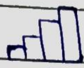
Lecture "8"

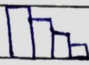
* Skewness

الحالة (unimodal) نستطيع عن تعرف الـ data

بالـ Histogram لو كان شكله



skewed to the left



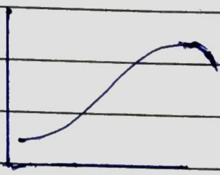
skewed to the right

symmetric distribution (for unimodal)

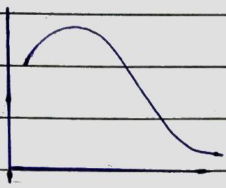
tail (حسب الذيل)

x	1	2	3	4	5
y					

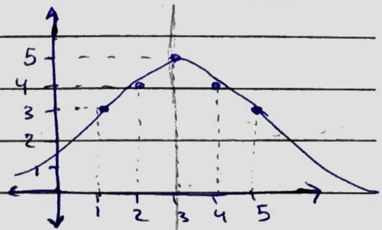
x	1	2	3	4	5
	3	4	5	4	3



skewed to the left

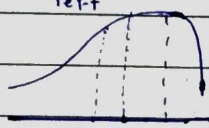


skewed to the right



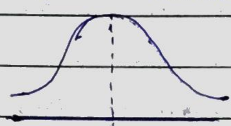
بعكس (mirror)

* negatively skewed to the left



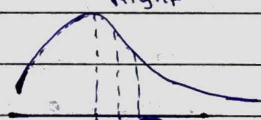
$\bar{x}$ Q_2 mode

symmetric distribution



$\bar{x} = Q_2 =$ mode

positively skewed to the Right

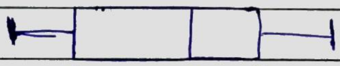


mode Q_2 $\bar{x}$

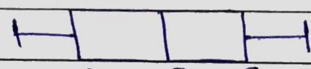
$\bar{x} < Q_2 <$ mode

$\bar{x} = Q_2 =$ mode

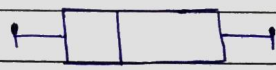
mode $< Q_2 < \bar{x}$



Q_1 Q_2 Q_3



Q_1 Q_2 Q_3



Q_1 Q_2 Q_3

$Q_2 - Q_1 > Q_3 - Q_2$

$Q_2 - Q_1 = Q_3 - Q_2$

$Q_2 - Q_1 < Q_3 - Q_2$

نقول عن حالة الـ symetric استخدمنا الـ $\bar{x}$ & S

نقول عن skewed to the Right استخدمنا الـ Q_2 / IQR

= = = skewed to the left = Q_2 / IQR

Most distribution have an index of skewness betwen −3 and 3

$$P = \frac{3(\bar{x} - \text{median})}{S}$$

When ⇒ * $P > 0$, the data are skewed right //

$P < 0$ → the data are skewed left // $P = 0$, the data are symmetric.

كل ما كلما اقتربت من الصفر كان اقرب للتوزيع symetrice

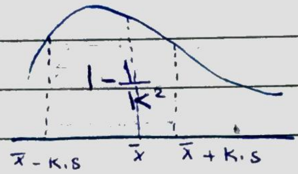
COLLEGE

Chebyshev's Inequality :-

* Chebychev's Theorem

skewed to the R / symmetric / skewed to the L ← بغض النظر اذا كانت data

For $k > 1$ & for any collection of data there are at least (على الأقل) $1 - \frac{1}{k^2}$ of the observation lie in $(\bar{x} - k.s , \bar{x} + k.s)$ ← Interval



Note: at most (على الأكثر) $\frac{1}{k^2}$ lie outside $(\bar{x} - ks , \bar{x} + ks)$

ولاحظ ان الاعتماد على الفترة $(\bar{x} - 2s , \bar{x} + 2s)$ يعتبر unusual ⇐ لأن $\boxed{K = Z\text{score}}$

eg) For a collection of 500 observations, we have:

$\bar{x} = 50$, $s = 5$ Find :-

a) The interval whose center is 50 & contains at least ~~90%~~ 450 of the observations.

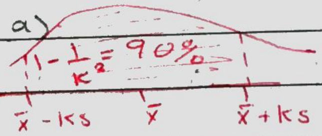
b) The interval whose center is 50 & at most ~~75%~~ 125 of the observation lie outside it.

c) the no. of observations that at ~~most~~ least in (35, 65). → $1 - \frac{1}{k^2}$

d) the no. of observation that at most outside (30, 70) → $\frac{1}{k^2}$

sol) $\bar{x} = 50$, $S = 5$, $n = 500$

a)



$\frac{450}{500} = 0.90$

$\therefore 1 - \frac{1}{k^2} = 0.90 \rightarrow \frac{1}{k^2} = 0.10 \rightarrow k^2 = 10 \rightarrow k = \sqrt{10} \approx 3.16$

$(\bar{x} - ks , \bar{x} + ks) : ((50 - 3.16(5)) , (50 + 3.16(5)))$

$: (34.2 , 65.8)$

at least 450 out of 500 observation lie in the interval (34.2 , 65.8)

b) $\frac{125}{500} = 0.25 = \frac{1}{4}$

$\frac{1}{k^2} = \frac{1}{4} \rightarrow k^2 = 4 \rightarrow \underline{\underline{k = 2}}$ اخذت +2 وتجاهل -2

لان بالشرط $\underline{\underline{k > 1}}$

$(\bar{x} - ks \,,\, \bar{x} + ks) : (50 - 2*5 \,,\, 50 + 2*5) : (40, 60)$

at most 125 out of 500 obs. lie outside $(40, 60)$

c) $\bar{x} + ks = 65$ or $\bar{x} - ks = 35$

$50 + 5k = 65 \rightarrow 5k = 15 \rightarrow \boxed{k = 3}$

at least $1 - \frac{1}{k^2} = 1 - \frac{1}{9} = \frac{8}{9}$ in $(35, 65)$

لي نسبة المشاهدات 0.889 من أصل 500

$\underline{444.4} = 500 * 0.889$ فكم عدد المشاهدات

at least 445 lie in this interval [in $(35, 65)$]

d) $\bar{x} + ks = 70$ or $\bar{x} - ks = 30$

$50 + 5k = 70 \rightarrow 5k = 20 \rightarrow \boxed{k = 4}$

at most $\frac{1}{k^2} = \frac{1}{4^2} = \underset{\text{العينة}}{\frac{1}{16}} = 0.0625$ lie outside $(30, 70)$

$31.25 = \frac{1}{16} * 500$ فعدد المشاهدات

at most 31 observations lie outside $(30, 70)$.

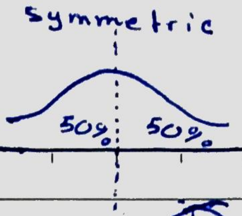
* The Emperical Rule:

For a bell-shaped (unimodal symmetric data) frequancy graph, we have

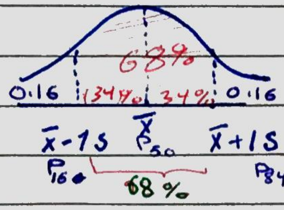
a) The percentage of data that lie within 1 standard deviation about the mean is about 68% $(\bar{x} - s, \bar{x} + s)$

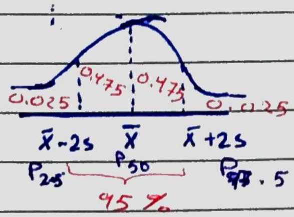
b) The percentage of data that lie within 2 standard deviation of the mean is about 95% $(\bar{x} - 2s, \bar{x} + 2s)$

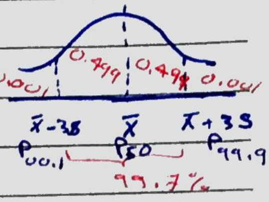
c) The " -- -- " -- -- 3 standard deviation about the mean is about 99.7% $(\bar{x} - 3s, \bar{x} + 3s)$



Bell-shaped distribution
→ unimodal // symmetric







eg) For a bell-shaped distribution with 500 observation

$\bar{x} = 50$ & $S = 5$ Find:

a) the percentage of data that lie in (40, 60)

b) " " " " " in (45, 65)

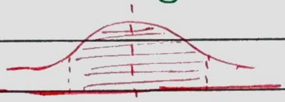
c) " " " " " " in (55, 65) shifted mean

d) " " " " " " in (35, 45)

e) How many observation in (40, 60) ? $\frac{95}{100} * 500 = 475$

sol)

a)



40 50 60

$(50 - ts)$ $(50 + ks)$

$50 - t(5) = 40$ $50 + 5k = 60$

$\boxed{t = 2}$ $\boxed{k = 2}$

$\bar{x} - 2s$ $\bar{x} + 2s$

About 95% within 2 standard deviation.

∴ 95% lie in (40, 60)

b)

45 50 65

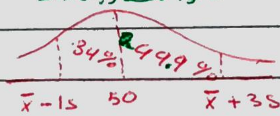
$(50 - 5t)$ $(50 + 5k)$

$45 = 50 - 5t$ $65 = 50 + 5k$

$\boxed{t = 1}$ $\boxed{k = 3}$

$\bar{x} - 1s$ chybyshev's $\bar{x} + 3s$

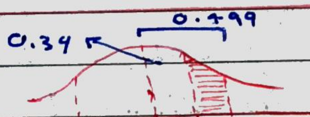
بيناف كانوا متساويان دائما



$0.34 + 0.499 = 0.839$

83.9% lie in (45, 65)

c)



$\bar{x} - ts$ 50 $\bar{x} + ks = 65$

$55 = \bar{x} + ts$

$50 + 5t = 55$ $50 + 5k = 65$

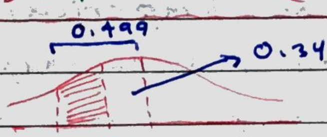
$\boxed{t = 1}$ $\boxed{k = 3}$

$\bar{x} + 1s$ $\bar{x} + 3s$

$0.499 - 0.34 = 0.159$

15.9 % in (55, 65)

d)



35 45 50

$\bar{x} - ts$ $\bar{x} - ks$

$\bar{x} - ts = 35$ $\bar{x} - ks = 45$

$50 - 5t = 35$ $50 - 5k = 45$

$\boxed{t = 3}$ $\boxed{k = 1}$

$0.499 - 0.34 = 0.159$

15.9 % in (35, 45)

COLLEGE

lecture "[illegible]"

اشياء لم تشرح على اليوتيوب لسنة 2020 واخا فقط عالمني لعام 2022

* levels of mesurements
* Q_1 & Q_3 for raw data طريقة حسابها بعض الأمثلة

$$\underbrace{1,\ 2,\ 3}_{Q_1},\ \underbrace{4}_{Q_2},\ \underbrace{5,\ 6,\ 7}_{Q_3}$$

ونعوض بالـ median للجدول — الخطوة الأولى — منوخذ الـ median الهم

$IQR = Q_3 - Q_1$ ومنها *

* standard deviation & variance :-

Population | Sample

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} = \frac{\sum x^2}{N} - (\mu)^2$$

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

N: population size

n: sample size

((المتوسط لا الانحراف لجذر القيم وسيطها))

eg) Find the standard deviation for the following sample set of data: 2, 7, 5, 11, 5

sol) $\sum x = 30 \rightarrow \bar{x} = \frac{\sum x}{n} = \frac{30}{5} = \boxed{6}$

x : 2, 5, 5, 7, 11

$(x-\bar{x})$: -4, -1, -1, 1, 5 — (+) أكبر من $\bar{x}$ ، (−) أقل من $\bar{x}$

deviation : انحراف القيمة عن الوسط

$\sum(x-\bar{x}) = 0$ ((sum of deviation = zero)) دائماً

$(x-\bar{x})^2$: 16, 1, 1, 1, 25

$\sum(x-\bar{x})^2 = 44$ • s^2 : variance $= \frac{44}{4} = 11$

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{44}{5-1}} = \sqrt{11} \approx 3.32$$

eg: x : 1, 4, 4, 7

$\bar{x} = \frac{16}{4} = 4$

$(x-\bar{x})$: −3, 0, 0, 3
$(x-\bar{x})^2$: 9, 0, 0, 9
$\Sigma(x-\bar{x})^2 = 18$

$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{18}{3}} = \sqrt{6}$

S.D $= \sqrt{6} \approx 2.45$ متوسط بعد القيم عن وسطها

Variance $= s^2 = 6$

إذا ضفنا قيمة $(x = 4)$ للعينة

x : 1, 4, 4, 4, 7
mean $\bar{x} = \frac{20}{5} = 4$

عند اضافة قيمة تساوي الـ mean يبقى الـ mean كما هو (the same) اما اذا كانت أكبر من الـ mean ← $\bar{x}$ is increses
اما اذا كانت اقل من الـ mean ← $\bar{x}$ is decreses

أما بالنسبة لل S.D اذا ضفنا قيمة (الفرق بينها وبين $(\bar{x})$) كان اقل من S.D سيقل S.D واذا كانت = SD سيبقى S.D the same واذا كان > S.D سيزداد S.D

$x-\bar{x}>0 \Rightarrow x>\bar{x}$ ولأن

eg) If a Value x is added to the data such that $x-\bar{x} > s$ اذا كانت موجبة, then the mean will
a) increase b) decrease c) remains the same
& the S.D will
a) increase b) decrease c) remains the same

إذا ضفنا قيمة وكانت كالتالي

eg) $x - \bar{x} = s$

mean will increse — S.D remains the same

because $x > \bar{x}$

الخلاصة مهمة جداً: إذا ضفنا قيمة على العينة:-

$x - \bar{x} > s \rightarrow$ S.D سيزداد

$x - \bar{x} < s \rightarrow$ S.D سيقل

$x - \bar{x} = s \rightarrow$ S.D سيبقى كما هو

$x > \bar{x} \rightarrow$ $\bar{x}$ سيزداد

$x < \bar{x} \rightarrow$ $\bar{x}$ سيقل

$x = \bar{x} \rightarrow$ $\bar{x}$ سيبقى ثابت كما هو

إذا كان $s = 0$ تعني ذلك ان كل data نفس بعض (متشابهين)

* The standard score

z score $\quad z = \dfrac{x - \bar{x}}{s} \quad \dfrac{\text{Deviation}}{\text{standard deviation}}$ *

مثلاً إذا كان لدينا data الآتية:-

$1, 2, 3, 10 \rightarrow \bar{x} = 4$

$x - \bar{x}$: $-3, -2, -1, 6$

$(x - \bar{x})^2$: $9, 4, 1, 36 \rightarrow \Sigma(x - \bar{x})^2 = 50$

$$s = \sqrt{\frac{50}{4-1}} = \sqrt{\frac{50}{3}} = 4.08$$

لنجيب z score لل $x = 1 \leftarrow z = \dfrac{1-4}{4.08} = -0.735$

أقل من ال mean بمقدار about 0.7 S.D. صار في وحدة قياس

$x=2 \rightarrow z = \dfrac{2-4}{4.08} = -0.49$ → below the mean // $x=3 \rightarrow z = \dfrac{-1}{4.08} = -0.245$

$x=10 \rightarrow z = 1.47 \rightarrow x=10$ is 1.47 S.D above the mean.

COLLEGE

eg

	statistics	physics
$\bar{x}$	15	18
s	2	3
x	18	20

وين اداءك أفضل.

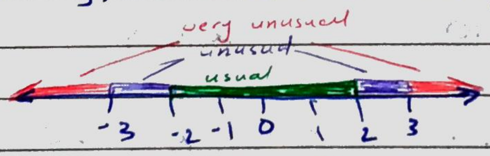
$$Z \text{ score} = \frac{x - \bar{x}}{s}$$

statistics $x = 18 \rightarrow Z \text{ score} = \frac{18 - 15}{2} = \frac{3}{2} = +1.5$

physics $x = 20 \rightarrow Z \text{ score} = \frac{20 - 18}{3} = \frac{2}{3} = +0.6$

The student did better in statistics than in physics.

* $-2 < Z < 2$ usual
* $2 < Z < 3$ او $-3 < Z < -2$ unusual
* $Z < -3$ او $Z > 3$ very unusual



Example 7

Finding Z scores The mean speed of vehicles ... is 56 miles per hour with a standard deviation of 4 miles per hour. You measure the speeds of three cars travelling along ... as 62 miles per hour, 47 miles per hour and 56 miles per hour. Find the z score to each speed. Assume the distribution of the speeds is approximately bell-shaped.

COLLEGE

sol) $Z \text{ score} = \frac{x - \bar{x}}{s}$

$x = 62$ mph	$x = 47$ mph	$x = 56$ mph
$Z = \frac{62 - 56}{4} = 1.5$	$Z = \frac{47 - 56}{4} = -2.25$	$Z = \frac{56 - 56}{4} = 0$
Usual	Unusually slow	
[1.5 SD above the mean]	[2.25 SD bellow the mean]	[0 SD equal the mean]

* Cofficient of variation (C.V)

ال S.D = 2 ليس كافي لتحديد التشتت كبير ام صغير.

نفس S.D بس نحسبوا كنسبة من ال (mean)

$$C.V = \frac{s}{\bar{x}} * 100\%$$

eg)

	section I	section II
mean	60	70
standard deviation	4.5	5

للوهلة الأولى بس ان تشتت السعة الثانية أكثر من الأولى

لكن ال Average للجب دور مهم بالحساب لذلك الأفضل (C.V)

section I $\rightarrow C.V = \frac{s}{\bar{x}} * 100\% = \frac{4.5}{60} \times 100\% = \boxed{7.5\%}$

section II $\rightarrow C.V = \frac{5}{70} * 100\% = 7.14\%$

C.V of section I > C.V of section 2

تشتت section I > تشتت sec II

* The variability of section I is more than the " " " II.

Find the cofficient of variation for this population :

$$\mu = \frac{873}{12} = \frac{291}{4} = 72,75 \approx 72.8$$

$$\sigma = \frac{\Sigma (x-\mu)^2}{N}$$

$(x-\mu)$: (-0.8) / 1.2 / (-4.8) / 3.2 / 1.2 / -3.8 /
(-0.8) / 6.2 / (-2.8) / (-3.8) / (4.2) / (0.2)

$$\hookrightarrow \frac{\Sigma (x-\mu)^2}{N} = \frac{130.28}{12} \approx \underline{\underline{10.9}}$$

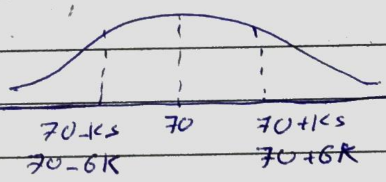
$$\sigma^2 = 10.9 \Rightarrow \boxed{\sigma = 3.3}$$

$$C.V_{Heights} = \frac{\sigma}{\mu} * 100\% = \frac{3.3}{72.8} * 100\%$$

$$= 4.54\% \approx 4.5\%$$

* For a collection of 200 observations with mean 70 and standard deviation 6, the interval $(70-a, 70+a)$ contains at least 150 observations. The value of a =

$n = 200$ $\quad \bar{x} = 70$



$(70 - 6k, 70 + 6k)$

$(70 - 12, 70 + 12)$

$$1 - \frac{1}{k^2} = 0.75$$

$$\frac{1}{k^2} = \frac{1}{4} \rightarrow \boxed{k = 2}$$

$a = ks \rightarrow a = 2 * 6 = 12$

Graphical representat

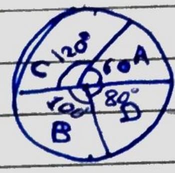
New syllabus Pie chart → 

Qualitative ⤴ ينزبط الـ
Quantitative ⤶

(Θ) Angle لازم نعرف كيف نحسب → $\Theta = \frac{f}{\Sigma f} \times 360^\circ$ (r.f)

e.g

	A	B	C	D	sum
f	3	5	6	4	18

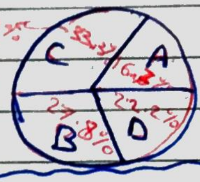


A → $\frac{3}{18} \times 360^\circ = 60^\circ$ (r.f) B → $\frac{5}{18} \times 360^\circ = 100^\circ$

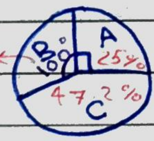
C → $\frac{6}{18} \times 360^\circ = 120^\circ$ D → $\frac{4}{18} \times 360^\circ = 80^\circ$

$$\boxed{f = \frac{\Theta}{360^\circ} \times \Sigma f}$$

r.f × 100%
$\frac{6}{18} \times 100\%$
= 33.3%



eg) A total of 36 items
find the frequancy
for each section A, B, and C



لاحظ أن
25% +
27.8% +
47.2%
= 100%

A → $f = \frac{\Theta}{360} \times \Sigma f = \frac{90}{360} \times 36 = \boxed{9}$

B → $f = \frac{100}{360} \times 36 = \boxed{10}$

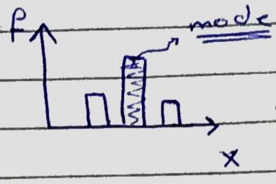
C → $f = \frac{360 - (100 + 90)}{360} \times 36 = \frac{170}{10} = \boxed{17}$

or : $36 - (9 + 10) = 36 - 19 = \underline{\underline{17}}$

	A	B	C	sum
f	9	10	17	36

COLLEGE

New syllabus [2] Bar graph (Pareto graph)

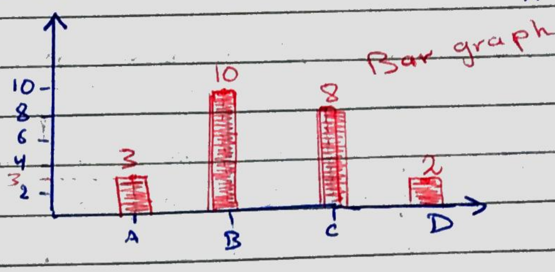


عادةً" للـ Discret أو للـ data غير المتصلة

eg)

	A	B	C	D
f	3	10	8	2

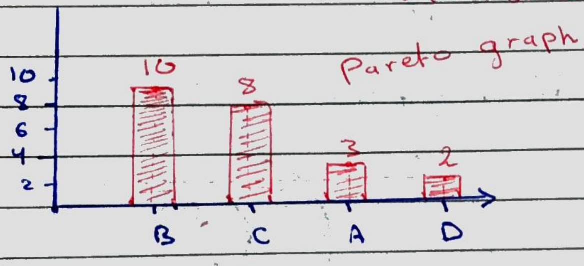
* draw Bar graph



B is modal mark لاحظ ان

بينما اذا كان مطلوب ← Pareto لازم يكونوا مرتبين B → C → A → D

الـ (Pareto) حالة خاصة من (Bar) .



كل Pareto graph تعتبر Bar graph

لكن لا يكون شرط العكس

[3] Scatter plot اذا كان عندي مجموعتين عدديات متعلقتان ببعض

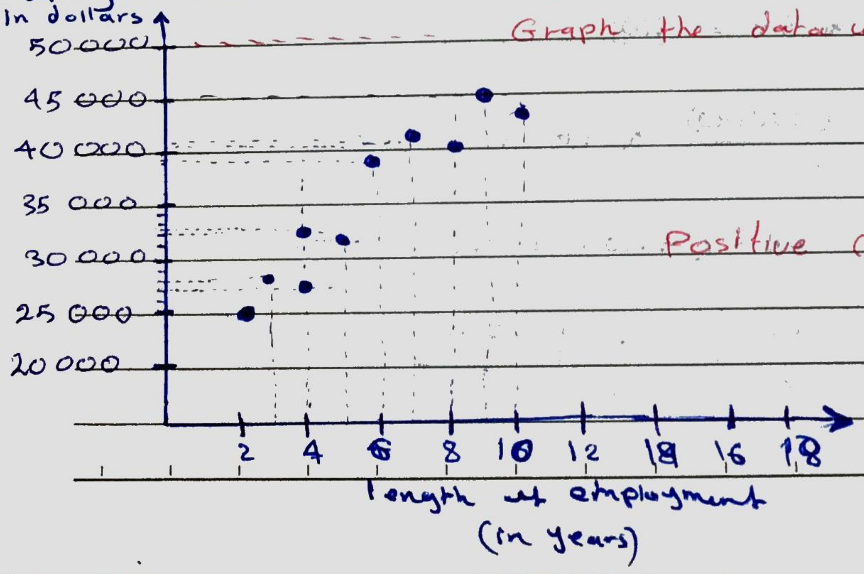
اذا الاثنين بيزيدوا مع بعض بيكون Positive corellation

مثال عليها (Petal length vs Width) للعالم Fisher

بينما لو احدهما بيزداد والآخر بيقل تكون العلاقة Negative correlation

TRY IT YOURSEIF 6 → Page 83

Graph the data using a scatter plot.



Positive correlation

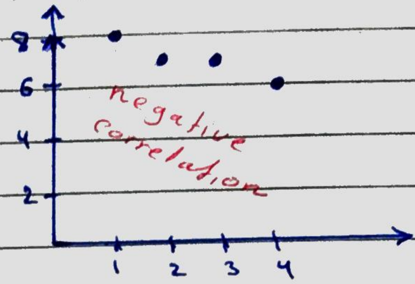
* Scatter plot

eg)

independent variable ← x

dependent variable ← y

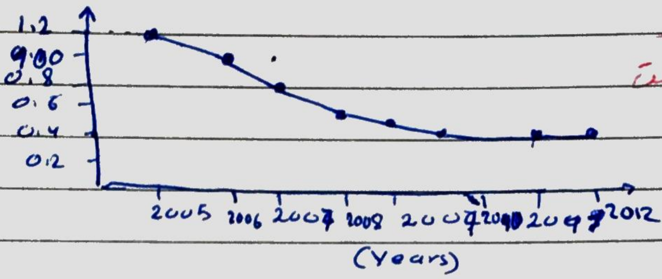
x	1	2	3	4
y	8	7	7	6



* as x increases → y decreases

4 Time series chart

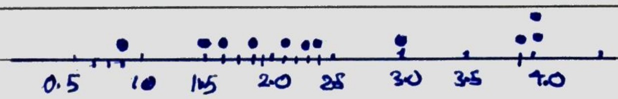
يُستخدم في السوق المالي كثيراً
عادةً متعلقة بأمور
لها علاقة بمرور الوقت
أيام، أشهر، سنوات ...
حتى يطلعوا عندهم توقع للمستقبل



5 dot plot

0	8
1	5 6 8
2	1 3 4 5
3	0 9
4	0 0

key : 0|8 = 0.8



eg)

x	5	8	10	15
f	1	2	3	5