Chapter 3 Descriptive Statistics

Numerical Measures

↳ Sample statistics is mesures are Computed for data from a sample

↳ احصاءيات العينة ← يتم حسابه لقياس للبيانات ماخوذة من عينة

↳ population parameters is mesures are Computed for data from a population

↳ يتم حساب لقياس للبيانات ماخوذة من المجتمع

↳ A sample statistics is point estimater of the Corresponcd population parameter

تقدير نقطي

3.1 Measures of location

[1] Mean - الوسط الحسابي - المعدل

[2] Median الوسيط

[3] Mode

[4] Percentiles

[5] Quartiles

[1] Mean (average) المعدل - الوسط الحسابي

↳ a measure of Central location for the data
مقياس للموقع المركزي للبيانات

Mean = $\frac{\text{sum of values}}{\text{number of values}}$ الوسط الحسابي

↳ for sample → $\bar{X} = \frac{\Sigma x_i}{n}$

$\bar{X}$
اكس بار

n = sample site

↳ for population → $\mu = \frac{\Sigma x_i}{N}$

μ
ميو

n = population size

$\bar{X}$ is a point estimater for μ

Example → Give the sample data 45, 95, 60, 77, 80 and Find the mean

sample mean = $\frac{\Sigma x_i}{n}$

$\bar{X} = \frac{45 + 95 + 80 + 60 + 77}{5} = 71.4$

How to use Calculater (تظبيط على الآلة الحاسبة)

خطوة

① Mode → 2 (SD appears on screan)

② 45 → M+ ⇒ n = 1
95 → M+ ⇒ n = 2
80 → M+ ⇒ n = 3
60 → M+ ⇒ n = 4
77 → M+ = n = 5

↓

نضغط على الشاشة عدد كل داتا

وهكذا حتى ندخل كل القيم

نصيحة: قبل تظبيط الآلة الحاسبة قبل البداية
Shift + mode + 3 + = + =

③ Shift [2] → نضغط على الشاشة

$\bar{x}$	x σ n	x σ n-1
1	2	3

جواب المثال اللي قبل [1] = answer ⇒ $\bar{x}$ = 71.4

* الآلة الحاسبة مهمة وبتهم احنا به الوقت في الامتحان بناءً على سرعة استخدام الآلة الحاسبة

* اهمية اختيار وضع الاصفار في البداية "خطوة 1"

[2] Median الوسيط

↳ is the value in the middle approximately 50% of data values are less than or equal median and approximately 50% of the value are greater than or equal the median

لـ الوسيط ← هي القيمة اللي بتكون بالوسط
50% من البيانات فوقه "اعلى منه"
50% من البيانات تحته "اقل منه"

How to find median

[1] arrange the data in ascending order
لـ رتب البيانات تصاعدياً

[2] Two cases are possible

① if n is odd "فردي"
↳ then median is the middle value

② if n is even "زوجي"
↳ then median is of the avarge of the two middle values

15, 20, 30, 23, 70, 80 ← هذا لـ
↳ Median = $\frac{30+23}{2} = \underline{\underline{26}}$
← نقرب العدد في حالة الأعشار

Example → Find the median → 22, 10, 18, 22, 16, 17, 20

① arrange → 10, 16, 17, 18, 20, 22, 22

n = 7 median = 18

[3] Mode المنوال

↳ the value(s) that has highest Frequency
"occure ll most"
↳ يعني الأعلى تكراراً

ex → Find the mode → 100, 140, 120, 123, 137
→ No mode

ex → Find the mode → 99, 140, 123, 140, 100, 140, 100
→ Mode = 140
▫ unimodal data → one mode

ex → Find the mode → 8, 12, 15, 8, 12, 18, 17, 10
→ Mode = 8, 12
▫ Bimodal data → two mode

Ex → Find the mode → A, AB, O; O, B, O, AB, A, O
→ Mode = O

Note → when ever the data set Contains extreme values
, the median is after the prefferd mesaure

→ ▫ Trimodal data → 3 mode

▫ Outlier مظاهر

↳ extreme values قيم متطرفة بعيدة عالية او قيم هابطة

لا يعني هذا الرواتب بين الـ 5000 ← 2000

وهناك شخص باخذ راتب 20,000

اذا هذا يعتبر outlier

outlier لها ← 300 لها واحد يوخذ

↳ الـ outlier لا يؤثر في الـ Mode

---

[4] Percentiles النسب المئوية/نتيل "مرعبة جداً"

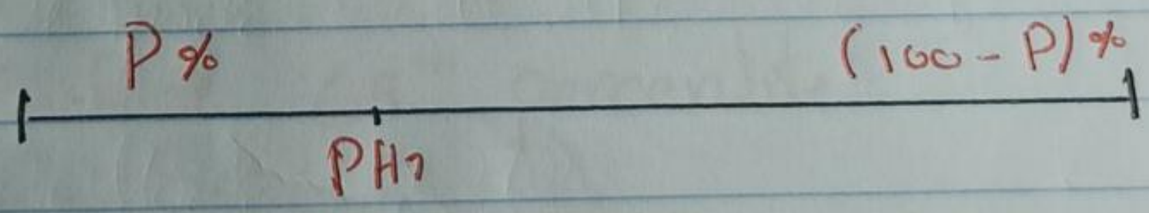
P^{th} percentile is the value that approximately P% of the values are less than or equal this value and (100-p)% of the values are greater or equal to this value

↳ يعني نقسم البيانات لـ 100 جزء

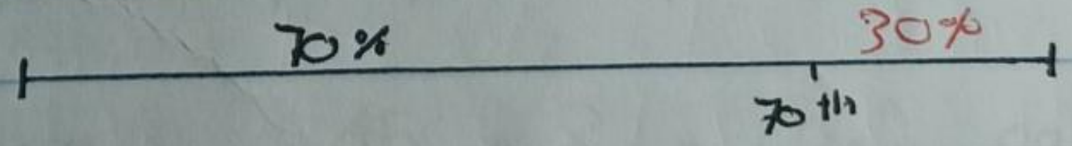
وتستخدم P% للسؤال عن جزء ما

ex→ P_{30} من 30 or more اعلى او يساوي

30 or less اقل او يساوي



ex→ P_{70}



How to find a Percentile

① arrange data in a ascending order
لازم ترتيب البيانات تصاعدي

② Compute $\underline{i}$ $\quad i = \frac{P}{100} \times n$

i = index (position) موقع القيمة المطلوبة $\quad n$ = حجم العينة sample size

[Case 1] if i is not integer, then the P^{th} percentile is the value in the next integer position
غير صحيح
لازم نقرب إلى الأعلى

[Case 2] if i is an integer, then the P^{th} percentile is the average of the two values in the i and $(i+1)$ position
يعني لذا كان عدد صحيح ناخذه ونأخذ إلي بعده

Example → Given the data 12, 18, 15, 25, 10, 8, 12, 9

Find the 65^{th} percentile

arrange → 8, 9, 10, 12, 12, 15, 18, 25

$i = \frac{P}{100} \times n = \frac{65}{100} \times 8 = \underline{5.2}$ not integer
نقرب للأعلى

P_{65} = The value of the 6th position

$P_{65} = 15$

② Find the P_{25} (25^{th} percentile)

arrange → 8, 9, 10, 12, 12, 15, 18, 25

$i = \frac{25}{100} \times 8 = \frac{2}{\text{integer}}$ ناخد بالقوي

P_{25} → The avarge of the vales in the secound and thired position

$P_{25} = \frac{9+10}{2} = 9.5$

→ Note

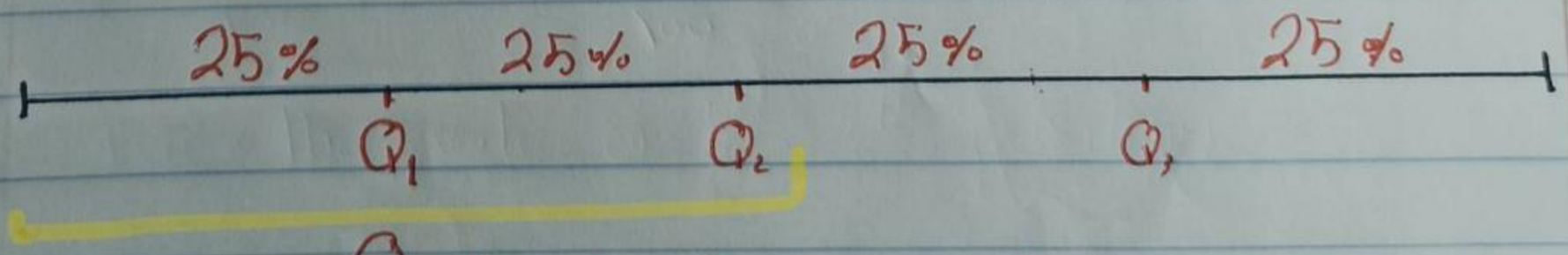
P_{50} (50^{th} percentile) = Median

[5] Quartiles الارباعي

① ↳ First quartile $Q_1 = P_{25}$

② ↳ Second quartile $Q_2 = P_{50}$ ~> Median

③ ↳ Third quartile $Q_3 = P_{75}$



Q_2 50% (25 + 25)

Example → Given the data → 12, 18, 16, 15, 20, 10

Find quartiles →

Step 1 arrange → 10, 12, 15, 16, 18, 20

1→ $Q_1 \rightarrow P_{25} = i = \frac{25}{100} \times 6 = 1.5$ not integer

2 ↵ نقربها Round up

$Q_1 = \underline{12}$

2→ Q_2 = Median "P_{50}" $= i = \frac{50}{100} \times 6 = 3$ integer

$Q_2 = \frac{15 + 16}{2} = 15.5$

3→ $Q_3 = P_{75} \rightarrow i = \frac{75}{100} \times 6 = 4.5$ not integer

نقربها ↳ 5

$Q_3 = 18$

يعني ثلاثة ارباع البيانات اقل او تساوي 18

or

75% من البيانات اقل او تساوي 18

3.2 Measures of variability "مقياس التباين"

1. Range
2. Interquartile range "IQR"
3. variance
4. standard deviation
5. Coefficient of variation "CV"

These numbers measure the spread "dispresion" هذه الارقام تقيس التشتت

يعني هنا عندما تكون علامات الطلاب قريبة من بعض

---

[1] Range "المدى"

↳ The simplest mesure of variability.
↳ ابسط مقياس للتغير

Range = largest value - smallest value
or
Maximum - Minimum

مهم جداً جداً مقياس عشان اعرف اذا البيانات ضربت او جمعت عن بعض

Example → 73, 78, 81, 81, 85, 90, 90, 93, 94

* Range is highly influenced by outliers

لما يكون في Outlier ← القيم الشاذة

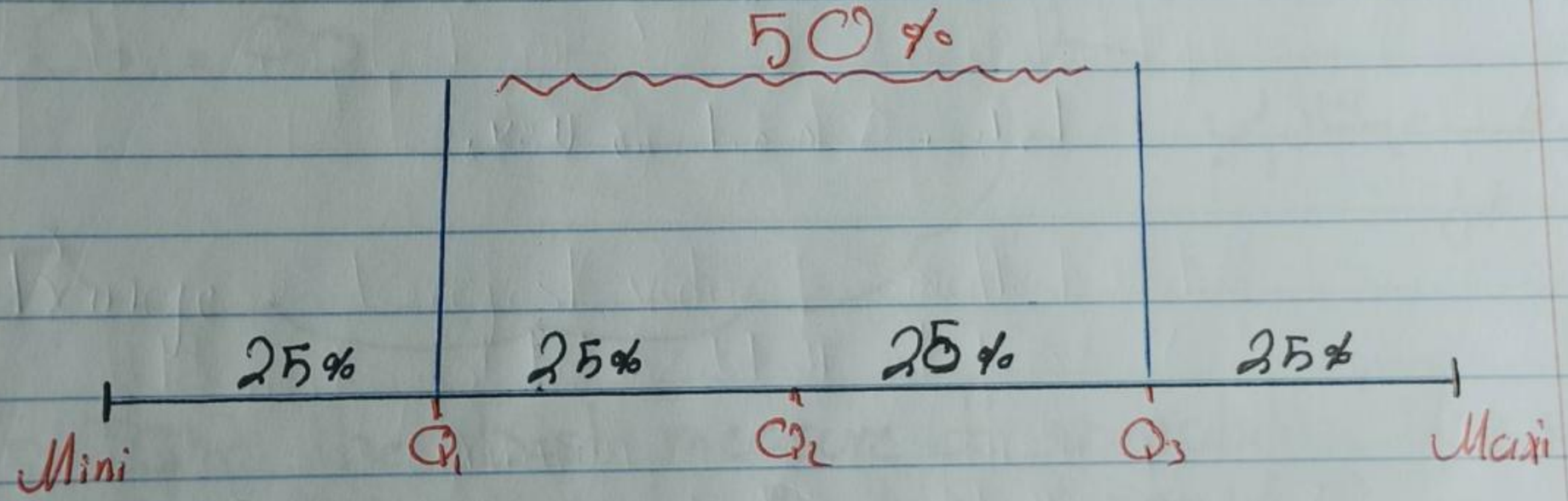
Range = Max - Mini
= 94 - 73
= 21

---

[2] Interquartile range "IQR" المدى الربعي

↳ The range of the middle [50%] of the data
المدى هو المدى لمنتصف البيانات

IQR = $Q_3 - Q_1$



- نوجد المدى لنطاق البيانات عادة هو موضح بالصورة
- الـ Range على IQR أكثر فيه
- الـ Range , IQR لا يمكن ان يكون سالب
ويمكن ان يساوي صفر في حالة عدم وجود أي قيمة

Example ⇝ $Q_1 = 81$
$Q_3 = 90$

Find IQR → $IQR = Q_3 - Q_1$

$= 90 - 81$

$= 9$

توضيح

$Q_1 = P_{25} = \frac{25}{100} \times 9$

$= 2.25 \rightsquigarrow 3$

$Q_1 = 81$

$Q_3 = P_{75} = \frac{75}{100} \times 9$

$= 6.75 \rightsquigarrow 7$

$Q_3 = 90$

* مثال الـ Range

حسب مثال لي قبل

- IQR The best measure of variability to use when extreme value "outlier" exist

يعني لـ IQR هو افضل مقياس عندما تكون هناك قيم شاذة "outlier"

لان الـ IQR لا تتأثر بالقيم الشاذة

لذلك فإن الـ IQR يعطي فكرة أفضل عن المجتمع

[3] Variance التباين

↳ a measure that utilize all the data.
لـ وهو مقياس يستخدم جميع البيانات

↳ Variance is the avarge of square deviation
لـ التباين هو متوسط إنحراف التربيعي

1. Population variance → التباين للمجتمع

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

sigma

N = pop size

μ = pop mean متوسط حسابي للمجتمع

2. sample variance → التباين للعينة

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

n = samp size

$\bar{x}$ samp mean متوسط حسابي للعينة

*we divided by $n-1$ to have the best estimation that is unbaised "غير متحيز" for the population

لـ هذه القوانين يمكن ايجادها وتطبيقها على الآلة الحاسبة اسرع واسهل

- S^2 is a point estimator to σ^2

note → $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

(deviation about the mean $x_i - \bar{x}$)

Example → Given the sample "مثال على هذه العينة"

x_1 x_2 x_3 x_4 x_5 x_6

15, 13, 17, 20, 22, 42

Find the variance "S^2" "$n = 6$"

$\bar{X}$ = مجموع القيم / عددهن = 21.5

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
13	13 - 21.5 = -8.5	72.25
15	" " = -6.5	42.25
17	= -4.5	20.25
20	= -1.5	2.25
22	= 0.5	0.25
42	= 20.5	420.25
Total	0	557.5

$\sum X_i - \bar{X} = 0$ دائماً ↔

هذا تحقق يطبق على أي مسألة ولكل عينة

$$S^2 = \frac{\sum_{i=1}^{6} (x_i - \bar{X})^2}{6 - 1}$$

$$S^2 = \frac{557.5}{5}$$

$$= 111.5$$

[4] Standard deviation "الانحراف المعياري"

↳ is the positive square root of the variance
له يجب هو الجذر التربيعي للتباين

Variance = (Standard devi.)^2

Standard deviation = $\sqrt{\text{Variance}}$

▫ حسب المثال اللي قبل

variance = 111.5

S. devi = $\sqrt{111.5} \approx 10.56$

▫ population standard deviation σ

▫ sample standard deviation S

S is a point estimator for σ

▫ Standard deviation and variance mesure whether the data values cluster around the mean
لـ قيم البيانات متقطة حول المتوسط

▫ The variance is useful in Comparing the variability of two or more variable
لـ التباين مفيد في مقارنة تباين متغيرين او اكثر

صفحة جديدة

- we can use the SD mode to find standard deviation and variance

Steps

[1] Reset all نعمل تنظيف

كيف نستخدم الآلة الحاسبة حسب كتاب مناهج

[2] Mode → 2 "SD"

[3] Data ندخل البيانات

15 → m+
13 → m+
17 → m+
20 → m+
22 → m+
42 → m+ → n = 6

[4] press shift + 2 → "4"

$\bar{x}$	xσn	xσn-1
1	2	3

[5] press "3" ⇝ then press "=" 10.5593 ≈ 10.56

في حالة طلب $\bar{x}$ (اكس بار) مباشرة نقوم بـ

[1] press shift + 2 ⇝ then press 1

[2] press "=" ⇝ 21.5

Notes

* IQR ← اقل اشي تأثر بال outlier

* Variance , S. deviation ← تأثر قليل بال outlier

* Range ← الاكثر تأثر بال outlier

---

* Standard. Deviation ← البيانات تبتعد عن المتوسط
والعكس صحيح

↑ البعد → ↑ S^2

↓ البعد → ↓ S^2

[5] Coefficient of variation معامل الاختلاف

↳ it ~~mean~~ measure the standard deviation relative to the mean

↳ يقيس الانحراف المعياري بالنسبة للمتوسط

$$CV = \left(\frac{S}{\bar{X}} \times 100\right) \%$$

Note → CV is used to Compare variability of two groups of data

↳ معامل الاختلاف يستخدم لمقارنة التشتت بين عينتين

Example ⇝

sample 1 → 10, 13, 12, 17, 18

sample 2 → 11, 16, 12, 9, 14

	sample 1	sample 2
[1] mean	$\frac{\Sigma Xi}{n} = \frac{70}{5}$	$= \frac{62}{5}$
	$= 14$	$= 12.4$
[2] S "الانحراف المعياري"	$= 3.39$	$= 2.7$
[3] CV	$= \left(\frac{3.39}{14} \times 100\right)\%$	$= \left(\frac{2.7}{12.4} \times 100\right)\%$
	$= 24.2\%$	$= 21.8\%$

* إذا كان الـ mean متشابه للعينتين فإن العينة ذات الانحراف المعياري الأعلى

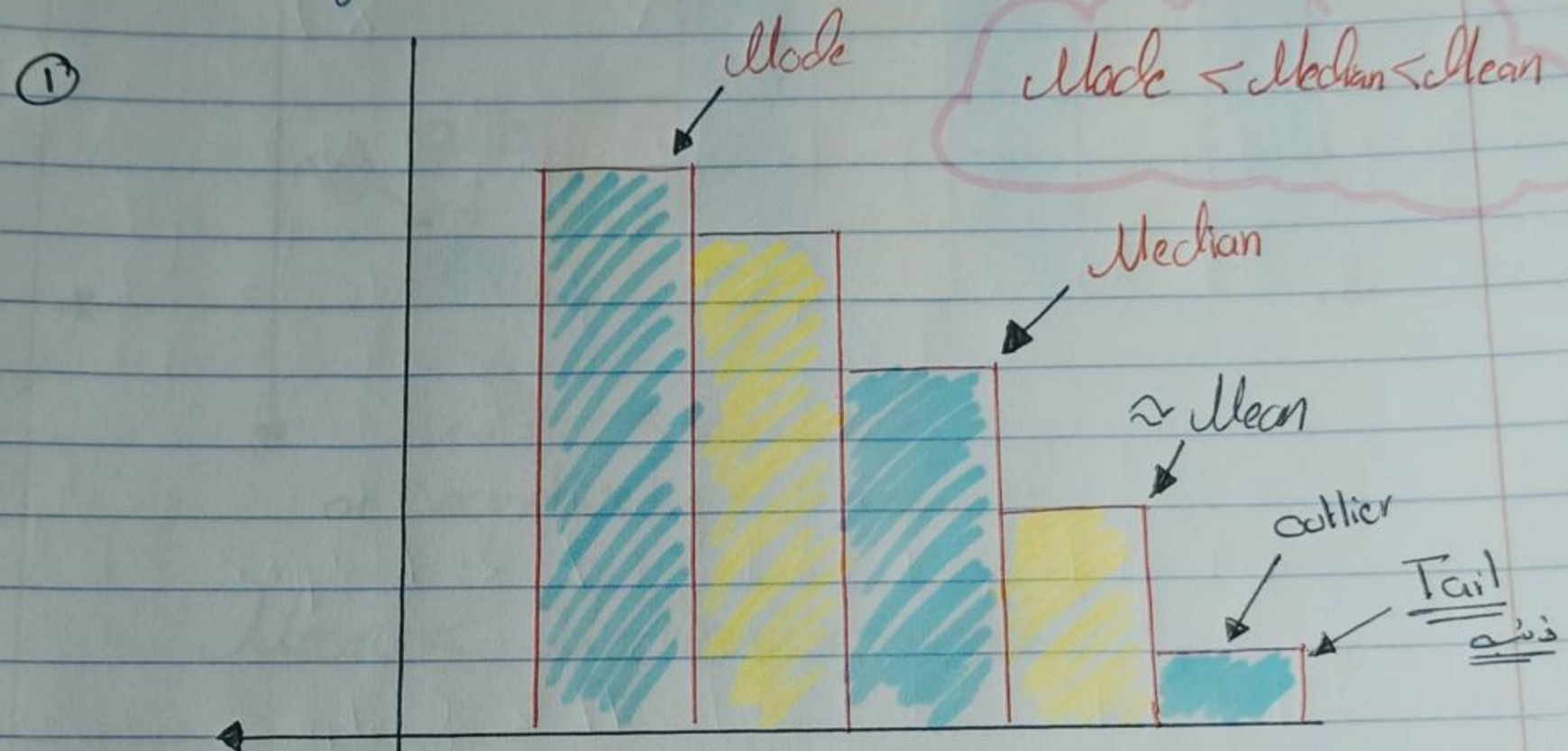
⇝ sample 1 has higher variability

* تختار القيمة الأعلى من بين العينتين " يعني ذات التشتت الأعلى "

3.3 Measure of Distribution shap

Relative location

- Distribution shap
  - Histogram



Mode < Median < Mean

- Right - skewed or positively skewed
  مائل لليمين

موقع ال "Tail" بـ Histogram

Outliers ← هناك شرطين ① تكرار قليل

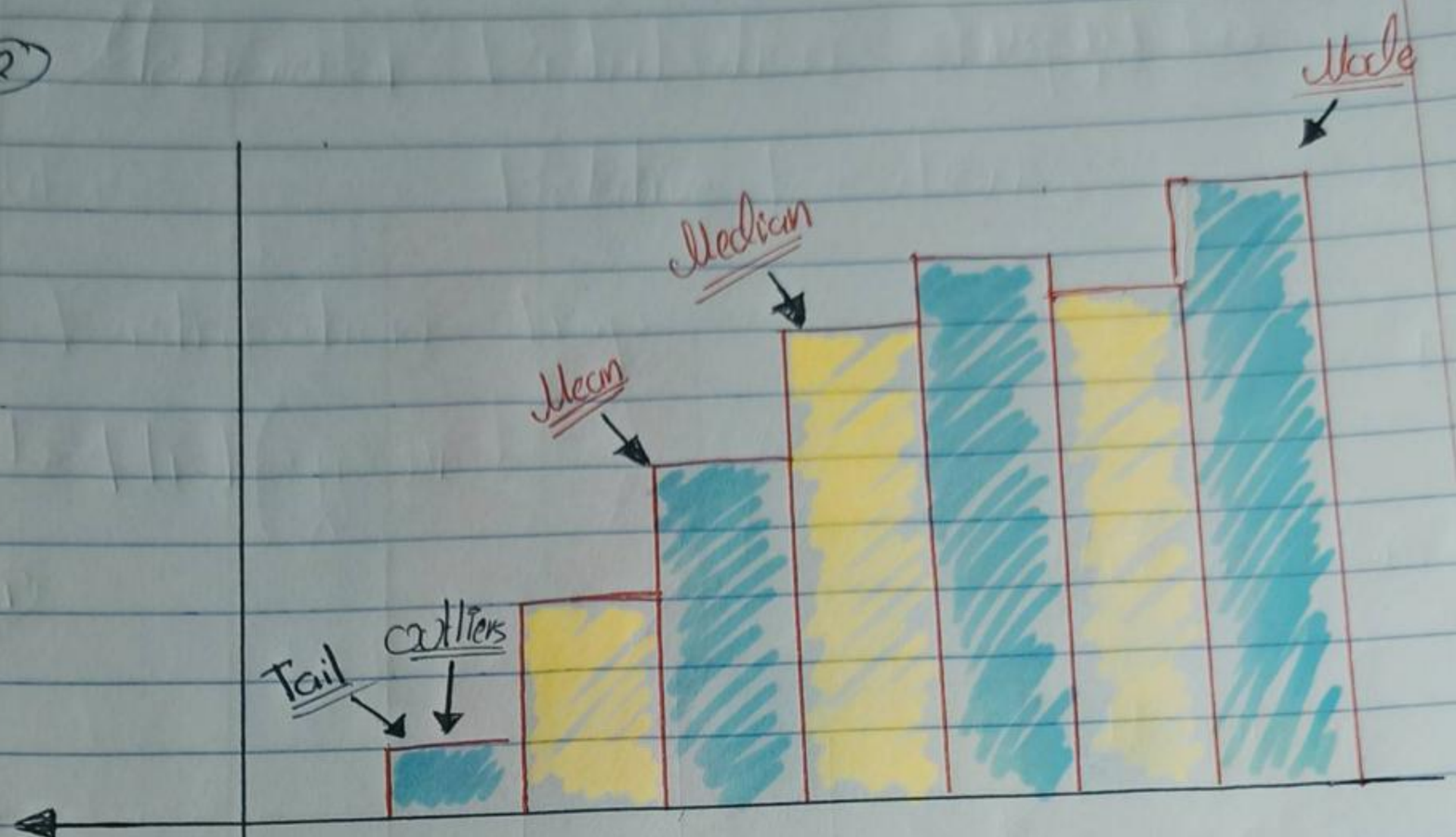
② قيمة عالية جداً
او منخفضة جداً

Mode ← هي القيمة الأعلى تكراراً

Mean ← يتأثر بال Outliers

Outliers تجعل الـ Mean بائجا هيها

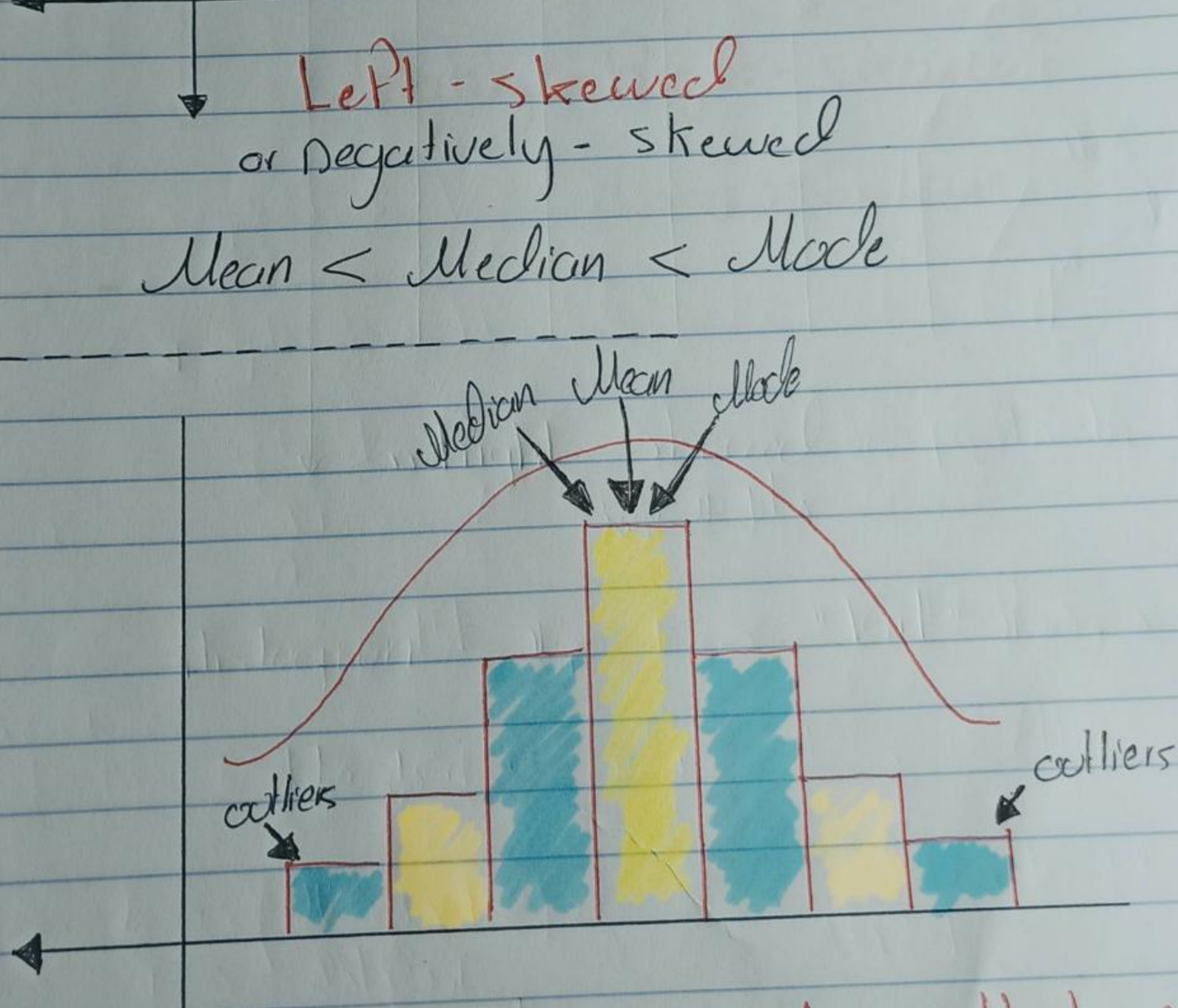
(2)



Left - skewed
or Negatively - skewed

Mean < Median < Mode

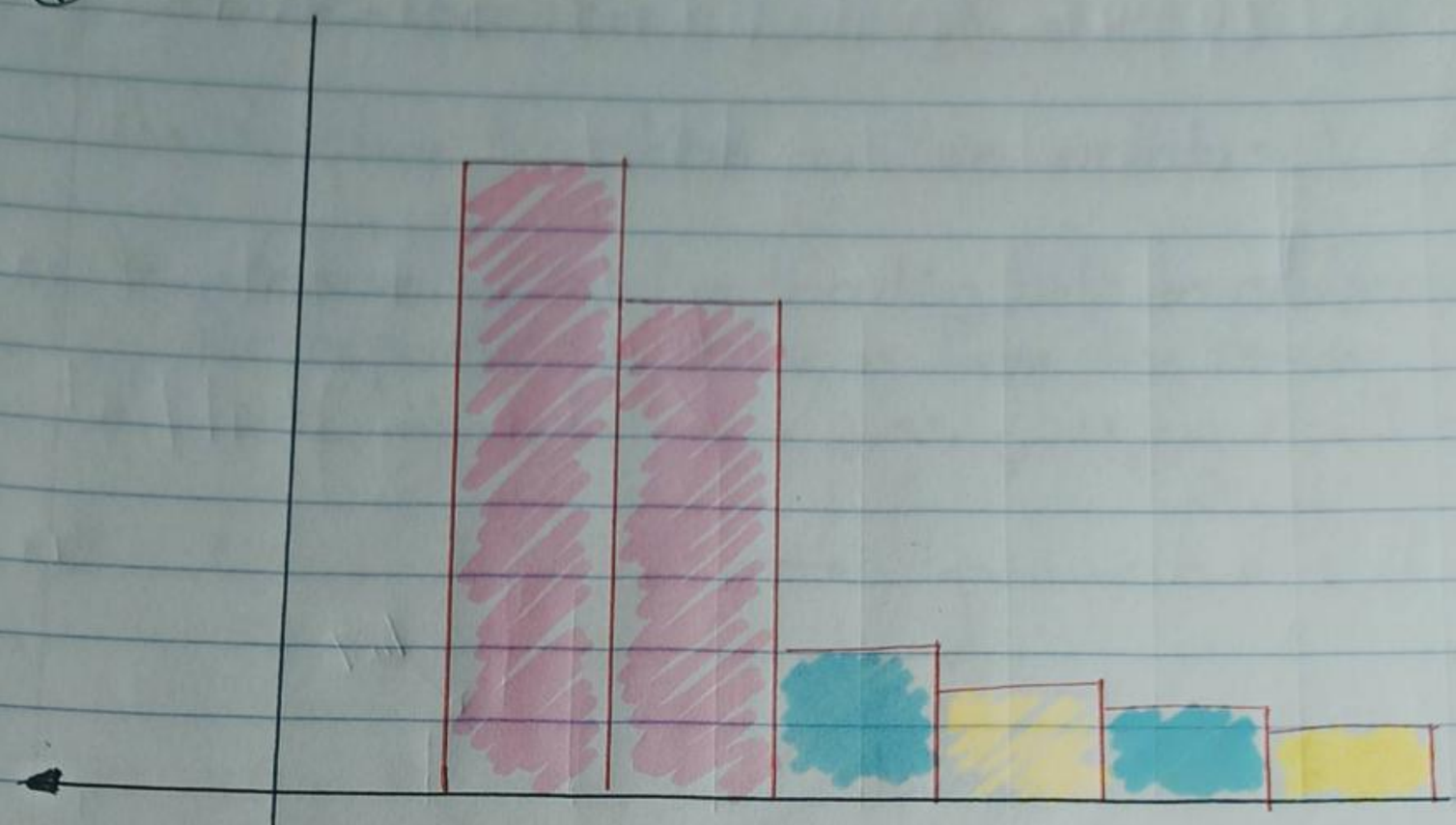
(3)



Symmetric - Histogram
Zero - skewed

Median = Mean = Mode

(4)



Highly - Right - skewed

الفرضاته كبيرة جداً

- Relative - location موقع نسبي
  - ↳ Relative location within adataset
  - ↳ Measure of relative location help to determian how far a particular value is from the mean
    لـ بـ الموقع النسبي يتحديد مدى بعد قيمة معينة عن المتوسط
    ← باستخدام ← Z - score ← وحدة معيارية
    لـ تستخدم لقياس الموقع النسبي
- using the mean and standard deviation we can determine the relative location of any observation
  لـ باستخدام المتوسط والانحراف المعياري يمكننا تحديد الموقع النسبي لأي ملاحظة
- if we have values $X_1, X_2, X_3 \dots X_n$
- and assume that the sample mean is $\bar{X}$
- and the sample standard deviation is S

Remmember that $S = \sqrt{\frac{\sum (Xi - \bar{X})^2}{n-1}}$

$\bar{X} = \frac{\sum Xi}{n}$

↳ We Can Compute a new value Called Z-score
for each value we can Compute Z-score

- Z-score

$$Z_i = \frac{X_i - \bar{X}}{S}$$

Z_i = Z-score of x_i

$\bar{X}$ = The sample mean

S = sample standard deviation

↳ The Z-score is called the standerized value

Explanation of Z-score

↳ is the number of standard deviation, x_i is from the mean $\bar{X}$

Ex ⇝ $S = 10$, $\bar{X} = 80$ $\quad X_1 = 70 \quad X_2 = 85 \quad X_3 = 95$

$$Z_1 = \frac{X_1 - \bar{X}}{S} = \frac{70 - 80}{10} = -1$$ ⇝ $X_1 = 70$ is -1 standard deviation less than the mean $= \bar{X} = 70$

$$Z_2 = \frac{85 - 80}{10} = 0.5$$ ⇝ $X_2 = 85$ is 0.5 standard deviation more than the mean

$$Z_3 = \frac{95 - 80}{10} = 1.5$$ ⇝ $X_3 = 95$ is 1.5 standard deviation more than mean

* Z-score ← تستخدم لمعرفة كم تبعد كل قيمة عن الـ mean

Ex ~> when Z-score = 0

this means that the value of x is the same as value of mean

$\bar{X} = 80 \quad S = 10 \quad X_4 = 80$

$Z_4 = \frac{80 - 80}{10} = 0$

---

also, Z-score for any observation can be interpretted as a measure of the relative location of the observation in the data set

↓ يعني ايضاً تفسير ~~درجة~~ مقياس لوصف بياني لاي ملاحظة على انها مقياس للموقع النسبي للملاحظة في مجموعة بيانات

Example ~> $\bar{X} = 80$ $\quad \bar{y} = 20$

$S = 10 \quad S = 2$

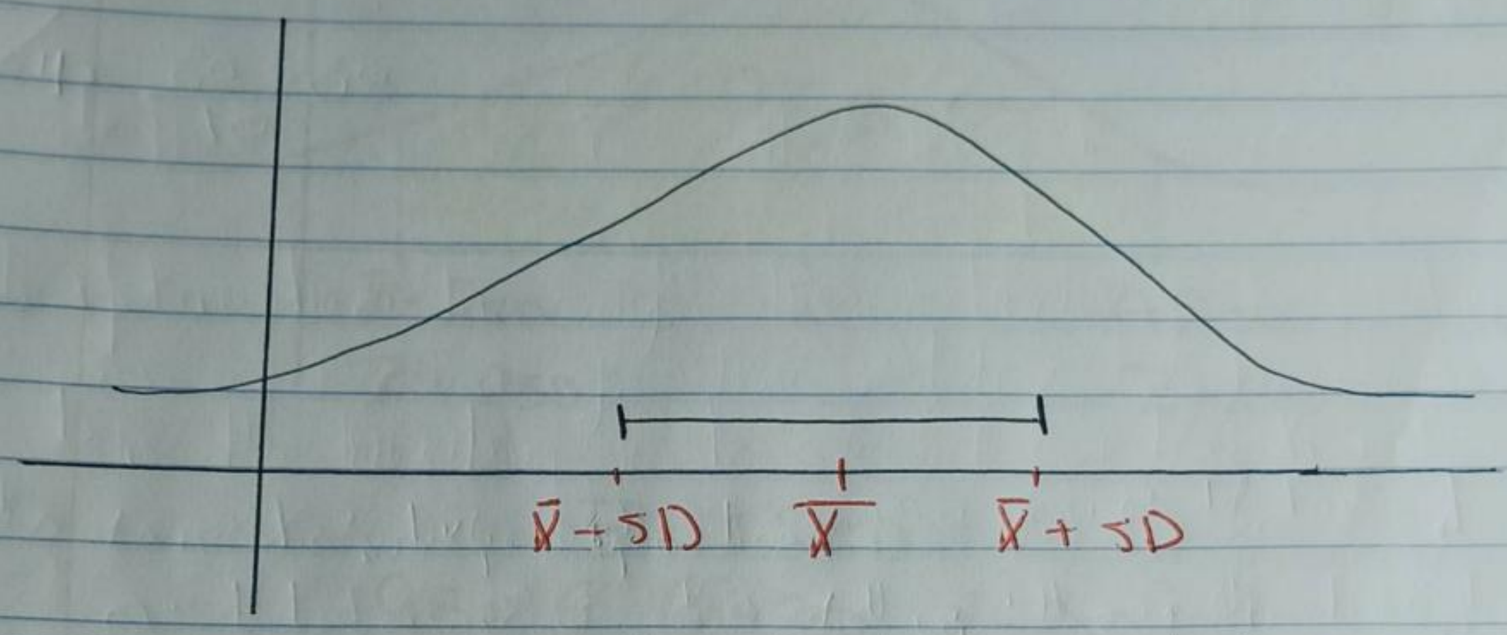
$X_1 = 100 \quad y_1 = 24$

$X_1 = 100$ and $y_1 = 24$ have some relative location

$Z_{x_1} = \frac{100 - 80}{10} = 2 \qquad Z_{y_1} = \frac{24 - 20}{2} = 2$

▫ Empirical Rule حكم التجريبية

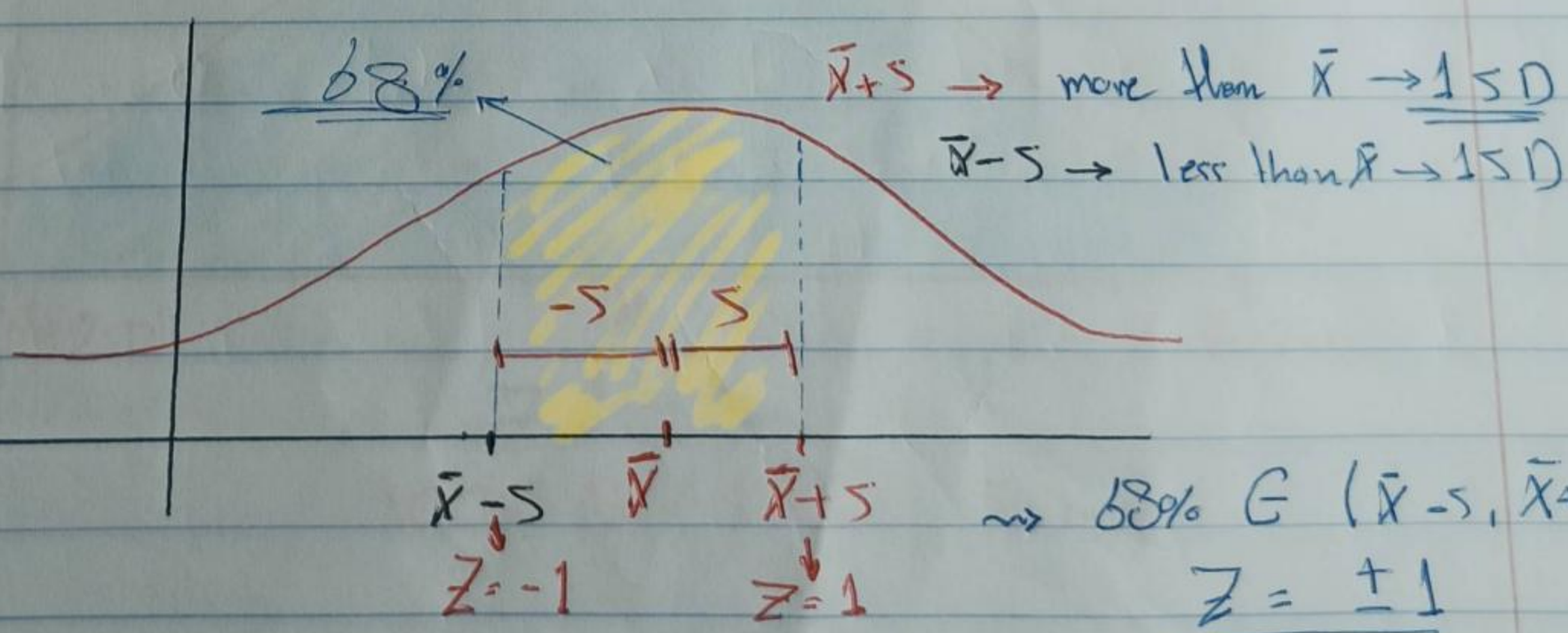
↳ if the data set exhibits, a symmetric shaped or bell-shaped distribution like
لو اذا عرضت مجموعة البيانات شكل متماثل او توزيع على شكل جرس فان



Then we can apply the empirical Rule.

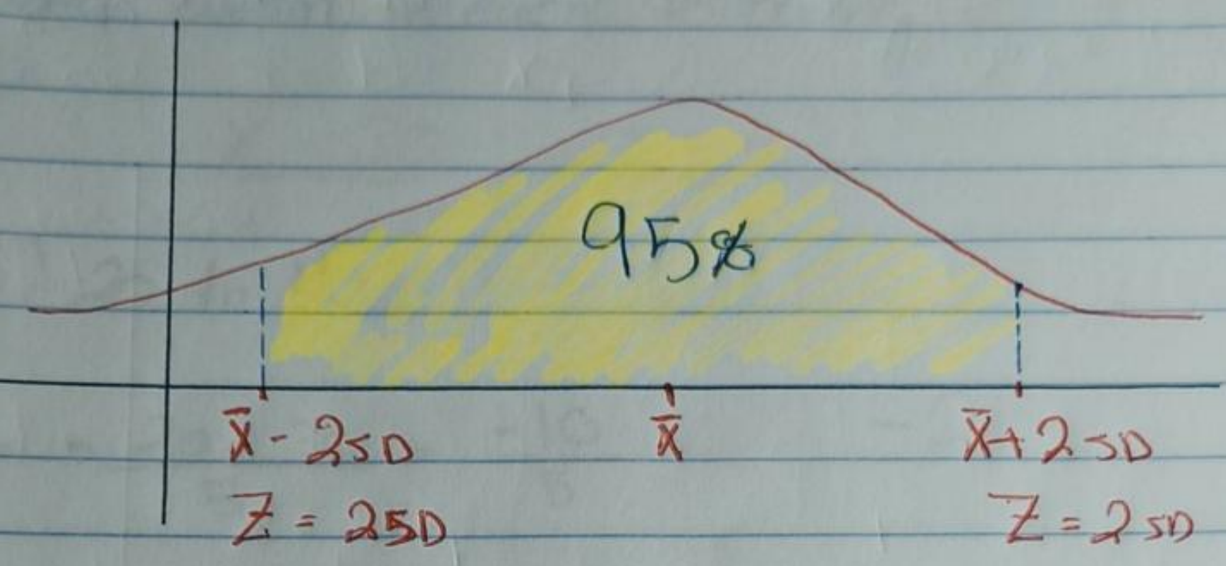
↳ Empirical rule for data having a Bell-shape distribution

▫ approximately 68% of the data values will be within 1 standard deviation of the mean
لـ 68% من البيانات موجودة في 1 انحراف معياري
لـ بين Z1 and Z-1 من 68% البيانات



- Approximately 95% of the data values will be within 2 standard deviation of the mean

لـ 95% من بياناته في 2 انحراف معياري

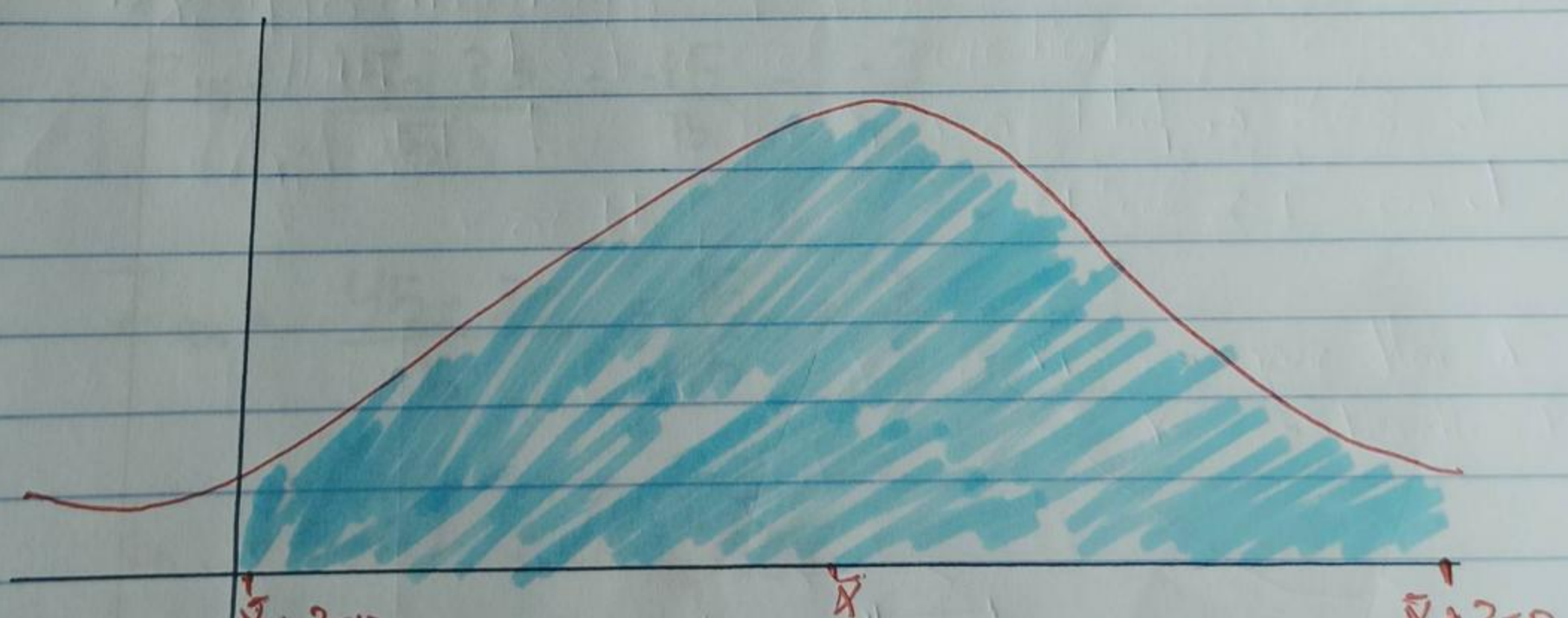


$$Z = \pm 2SD$$

$$95\% \in (\bar{X} + 2SD, \bar{X} - 2SD)$$

- Almost all of the ~~data~~ values will be within 3 standard deviations of the mean

لـ تقريبا جميع البيانات محصورة في 3 انحراف معياري



$$Z = \pm 3SD$$

$$100\% \in (\bar{X} + 3SD, \bar{X} + 3SD)$$

Ex suppose that the data have a bell-shaped distribution with a mean of 30 and standard deviation of 5

use the empirical Rule to determine the percentage of data within of each following

$\bar{x} = 30$ $s = 5$

A) 20 to 40

$$z_{20} = \frac{20 - 30}{5} = \frac{-10}{5} = -2$$

$$z_{40} = \frac{40 - 30}{5} = \frac{10}{5} = 2$$

↳ within 2 standard deviation, we have 95% of data

B) 15 to 45

$$z_{15} = \frac{15 - 30}{5} = \frac{-15}{5} = -3$$

$$z_{45} = \frac{45 - 30}{5} = \frac{15}{5} = 3$$

↳ Almost all data

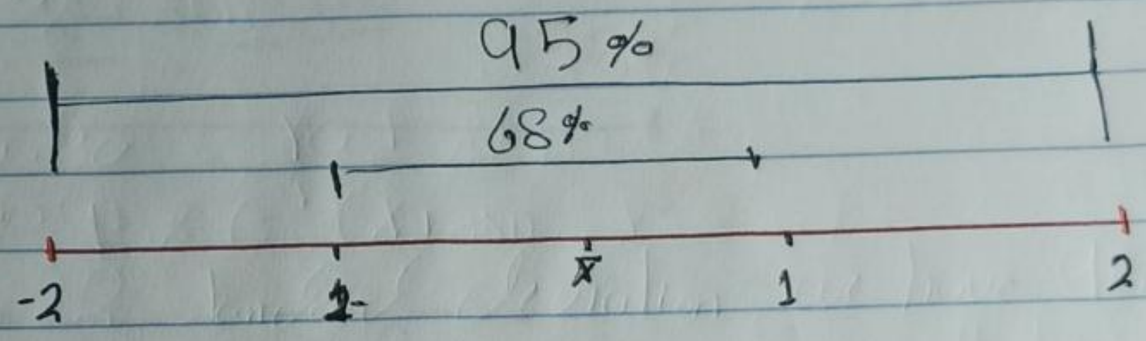
Ex → Assume that the grades of stat 236 are Bell-shaped with mean = 60 and standard deviation = 10

$\bar{X} = 60$ $\quad$ $S = 10$

A) Find percentage of data that are between 50 and 80

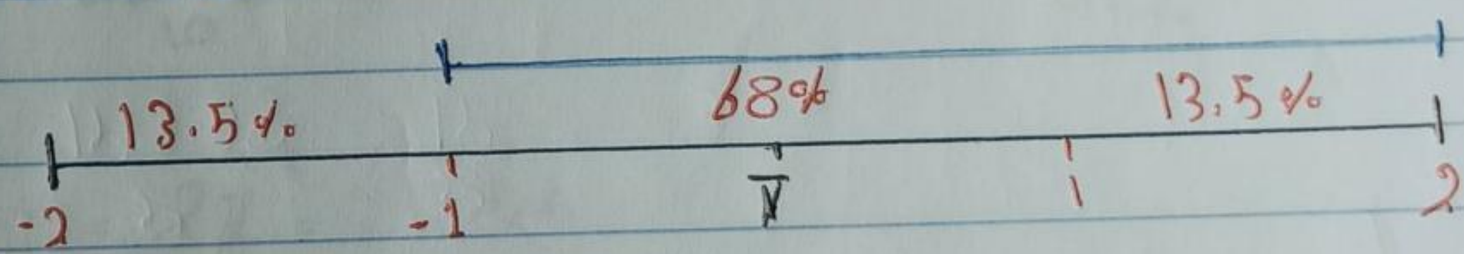
$$Z_{50} = \frac{50-60}{10} = \frac{-10}{10} = -1$$

$$Z_{80} = \frac{80-60}{10} = \frac{20}{10} = 2$$



95 - 68 = 27

$\frac{27}{2} = 13.5$



so between -1 and 2 = 68% + 13.5% = 81.5%

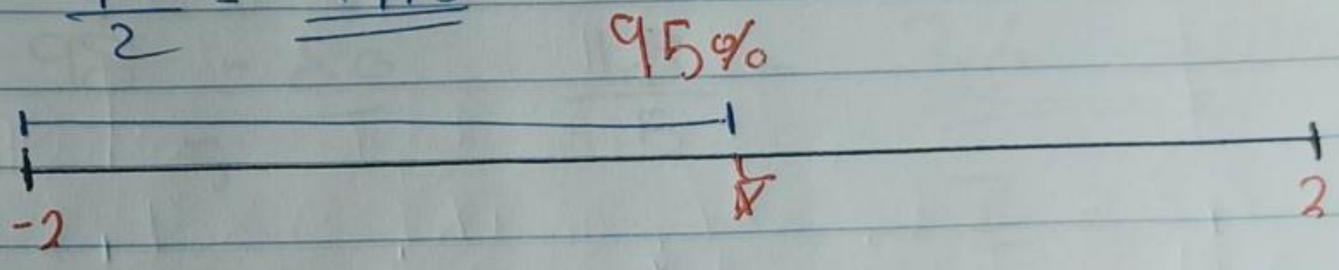
(B) between 40 and 60

$$Z_{40} = \frac{40-60}{10} = \frac{-20}{10} = -2$$

$$Z_{60} = \frac{60-60}{10} = \frac{0}{10} = 0$$

So we have
↳ 47.5 % of grades are between 40 and 60

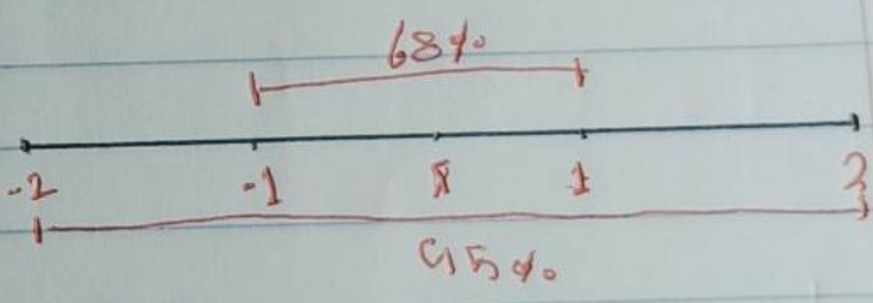
$$\frac{95}{2} = 47.5$$



(C) between 70 and 40

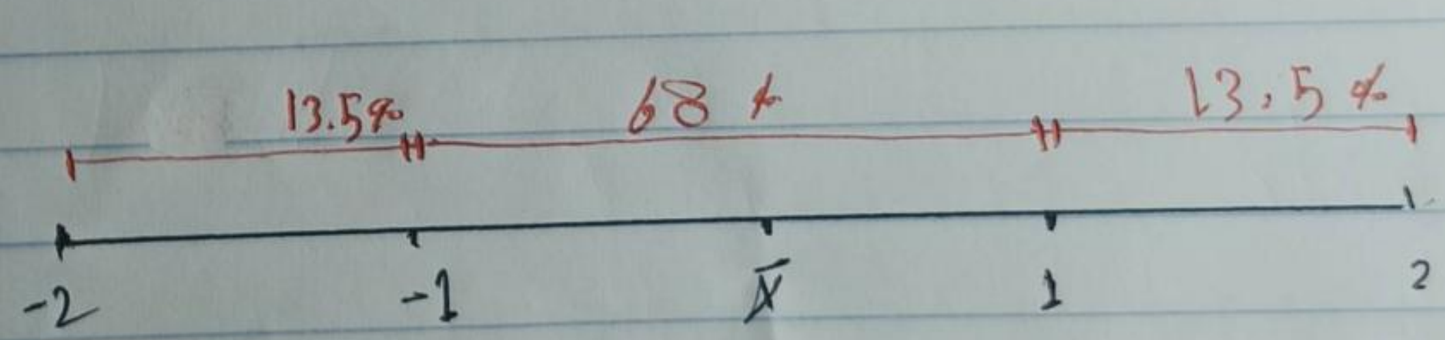
$$Z_{70} = \frac{70-60}{10} = 1$$

$$Z_{40} = \frac{40-60}{10} = -2$$



$$95 - 68 = \frac{27}{2} = 13.5\%$$

so between 1 and -2
$= 68 + 13.5 = 81.5\%$



30

Detecting outliers ⇒ Bell-shaped

↳ Extreme values are called also outliers

if $Z > 3$ or $Z < -3$

then this Z-score belong to outliers

Example → $\bar{X} = 80$

$S = 5$

Z_{98} ?

$$Z_{98} = \frac{98 - 80}{5} = \frac{18}{5} = 3.6$$ outliers

$Z > 3$

3.4 Exploratory Data Analysis ↔ تحليل البيانات

- 5-number summary

↳ The five following numbers are used to summarize data

1. smallest value
2. first Quartile Q_1
3. Median
4. Third Quartile Q_3
5. Largest value

- Example ⇝ Construct the five number summary of the following numbers

3310 , 3355 , 3450 , 3480 , 3490
3520 , 3540 , 3550 , 3650 , 3925

- Smallest value is ⇝ 3310

- First Quartile $Q_1 = P_{25}$

↳ So we have to compute the positions index i

$$i = \frac{P}{100} \times n \rightsquigarrow P_{25} = \frac{25}{100} \times 12 = 3$$

↳ i is integer ⇝ so the first Quartile is the avarage of the third and fourth values

$$Q_1 = \frac{3450 + 3480}{2} = 3465$$

- Median
  - ↳ we have two ways to compile th median

طريقة 1

1. we have 12 values so the median will be th avarge of the 6th and 7th values

$$\text{Median} = \frac{3490 + 3520}{2} = 3505$$

طريقة 2

2. Median = P_{50}

$$i = \frac{50}{100} * 12 = 6$$

i is integer so the median is th avarge of 6th and 7th values

$$\text{Median} = \frac{3490 + 3520}{2} = 3505$$

- Q_3
  - ↳ $Q_3 = P_{75}$

$$i = \frac{75}{100} * 12 = 9$$

i is the avarge of 9th and 10th values

$$Q_3 = \frac{3550 + 3650}{2} = 3600$$

- The largest value
  - ↳ the largest value = 3925

So the 5-number summary is

3310, 3465, 3505, 3600, 3925

Approximately 25% of the data are between any two adjacent "بيانات" #

لـ 25% من البيانات موجودة بين أي رقمين ←

"Box Plot"

Box Plot is a graphical summary of data that is based on Five Summary numbers

1) A box is drawn with the end of the box located at first and third quartiles
→ The box contains 50% of data

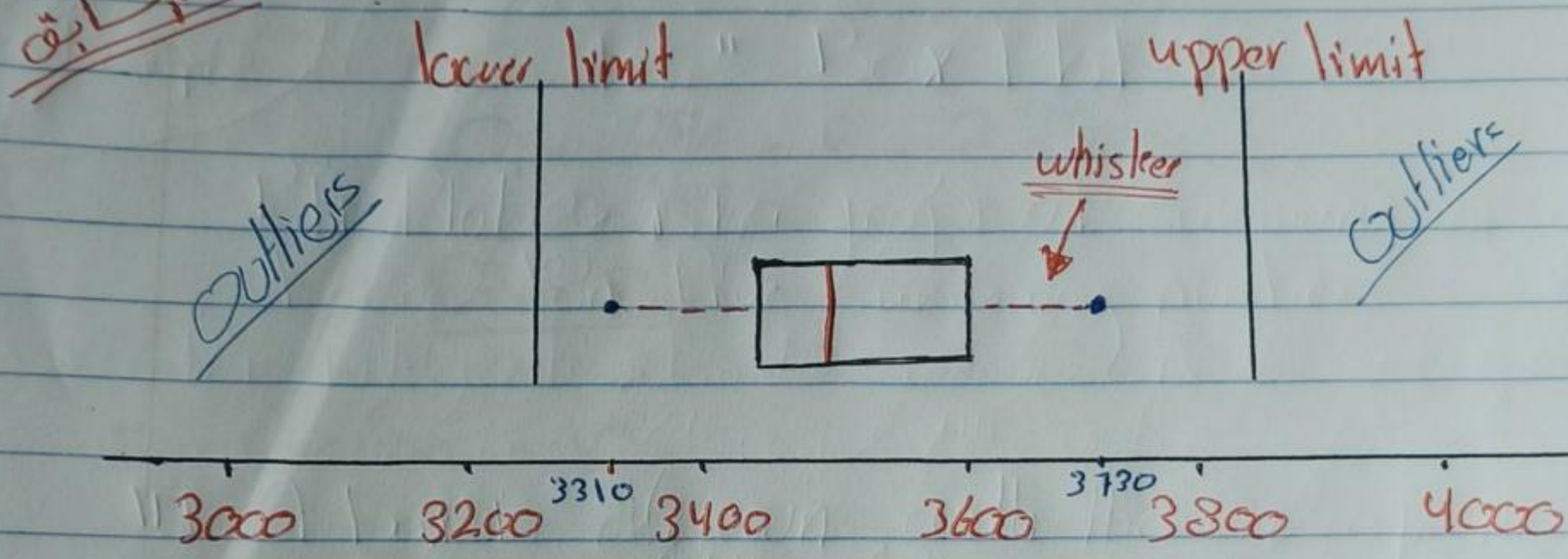
2) A vertical line is drawn in the box at location of median

3) By using the interquartile range limits are located, the limits for the box plot are (1.5)(IQR) below Q_1 and (1.5)(IQR) more than Q_3

④ dash lines are called Whisker " شارب "
- The whiskers are drawn from the end of the box to the smallest and largest values inside the limits compate in step 3

⑤ Finally the location of each outliers is shown with the symbol.

حسب السابق



$Q_1 = 3465$
$Q_3 = 3600$

median = 3505

$IQR = Q_3 - Q_1 = 3600 - 3465$
$= \underline{135}$

$1.5\ IQR = (1.5)(135) = 202.5$

$Q_1 - 1.5\ IQR = 3465 - 202.5 = 3262.5$ lower limit

$Q_3 + 1.5\ IQR = 3600 + 202.5 = 3802.5$ upper limit

Smallest # inside the limits is 3310 } → داخل الحدود
Largest # inside the limits is 3730 }

3925 is an outlier

Example

5, 15, 18, 10, 8, 12, 16, 10, 6

→ Constrct the Box plot

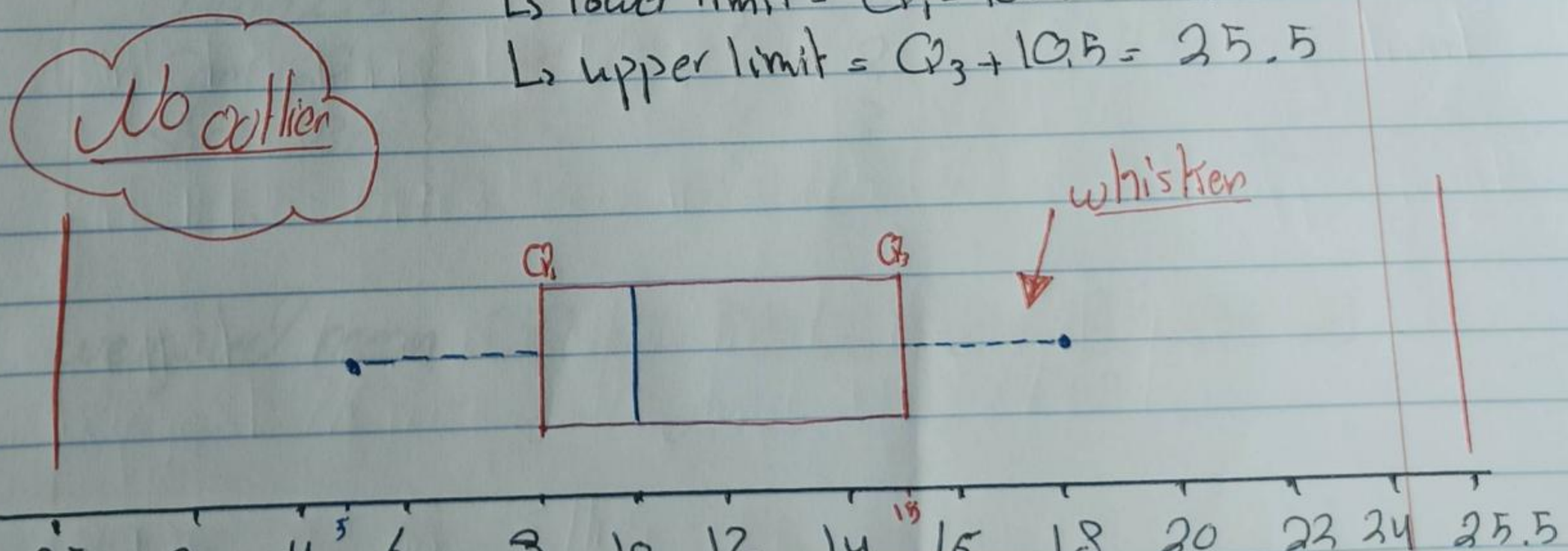
↳ 5, 6, 8, 10, 10, 12, 15, 16, 18

- Smallest value = 5
- $Q_1 = P_{25}$
  ↳ $Q_1 = \frac{25}{100} \times 9 = 2.25 \approx 3$
  So $\boxed{Q_1 = 8}$
- Median = 10
- $Q_3 = P_{75}$
  ↳ $Q_3 = \frac{75}{100} \times 9 = 6.75 \approx 7$
  $Q_3 = 15$

IQR = 15 - 8 = 7 ⇝ 1.5 IQR = 1.5 × 7 = 10.5

↳ lower limit = $Q_1 - 10.5 = -2.5$

↳ upper limit = $Q_3 + 10.5 = 25.5$



3.6 The weighted Mean and working with grouped data

→ The Mean of a sample

$$\bar{X} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

weighted mean ⇝ عندما نريد ان نوجد معدل لقيم لها مكافئة

$$\bar{X} = \frac{\sum x_i\, w_i}{\sum w_i}$$ ⇝ نفس طريقة حساب معدل الكامنترس

x_i = value of observation i ⟵ القيم

w_i = weight for observation i ⟵ وزن القيم

Example: You have 3 Cources ↔ لديك ثلاث مواد

Cources	Grade	# of hours
Math 235	90	3
Com 221	80	2
stat 236	95	3

ساعة المادة

weighted mean $(\bar{X}) = \frac{(90)\times(3) + (80)\times(2) + (95)\times(3)}{3+2+3}$

$\bar{X} = 89.375$

Example ~>

Clothes

Cost per meter (x_i)	# of meter (w_i)
3 $	20
5 $	15
2 $	70

Compute the weighted mean (cost per meter)

(x_i دي)

جسم w_i لم يكن ما بتذكر بالسؤال

$$\bar{X} = \frac{(3 \times 20) + (5 \times 15) + (2 \times 70)}{20 + 15 + 70}$$

$$= \frac{60 + 75 + 146}{85} = \frac{275}{85}$$

= 3.23 $ per meter

▫ Grouped data بيانات مبوبة

Classes	Frequency
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1

sample mean for grouped data

$$\bar{x} = \frac{\sum f_i M_i}{n}$$

$M_i \rightarrow$ The midpoint for class i

$f_i \rightarrow$ frequency of class i

$n \rightarrow$ sample size

Class	Midpoint	f_i	$M_i \times f_i$
10-14	$\frac{10+14}{2} = 12$	4	48
15-19	// = 17	8	136
20-24	// = 22	5	110
25-29	// = 27	2	54
30-34	// = 32	1	32
		20	380

$$\bar{x} = \frac{\sum f_i M_i}{n} = \frac{380}{20} = 19 \text{ days}$$

- Sample variance = S^2

$$S^2 = \frac{\sum P_i (x_{ip} - \bar{x})^2}{n-1}$$

مثال لـ لـ ف

Classes	P_i	x_{ip}	$(x_{ip} - \bar{x})$	$(x_{ip} - \bar{x})^2$	$P_i (x_{ip} - \bar{x})^2$
10 – 14	4	12	$12 - \underline{19} = -7$	$(-7)^2 = 49$	196
15 – 19	8	17	$17 - 19 = -2$	$(-2)^2 = 4$	32
20 – 24	5	22	// = 3	// = 9	45
25 – 29	3	27	// = 8	64	128
30 – 34	1	32	// = 13	169	169
	20				Σ 570

$\underline{19}$ $\bar{x}$ مثال لـ لـ ف

$$S^2 = \frac{570}{19} = \frac{30}{20-1}$$

- population Mean For grouped data

$$\mu = \frac{\sum f_i x_i}{n}$$

x_i لـ لـ مجتمع

- population varrance For grouped data

$$\sigma^2 = \frac{\sum P_i (x_{ip} - \mu)^2}{n}$$

3.5 Measures of Association between two variables

↳ In this section we present descreptive measuer of the relationship between two variables

Covariance التباين

↳ For n-observations $(X_1, Y_1)(X_2, Y_2)$

sample Covariance $S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

Example ~> we have two variables # of advertisement per month and sales for a Company (monthly)

# of ads = x (x)	sales (1000) (y)	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	50	-1	8	-8
1	20	-3	-22	66
4	30	0	-12	0
5	60	1	18	18
7	50	3	8	24
				100

Compute the Covariances of the above sample

$$S_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}$$

Note that
$\sum (x-\bar{x}) = 0$
$\sum (y-\bar{y}) = 0$

Example

$$\bar{x} = \frac{3+1+4+5+7}{5} = \frac{20}{5} = 4$$

$$\bar{y} = \frac{50+20+30+60+56}{5} = \frac{216}{5} = 42$$

$S_{xy} = \frac{100}{5-1} = \frac{100}{4} = 25$ two relationship

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

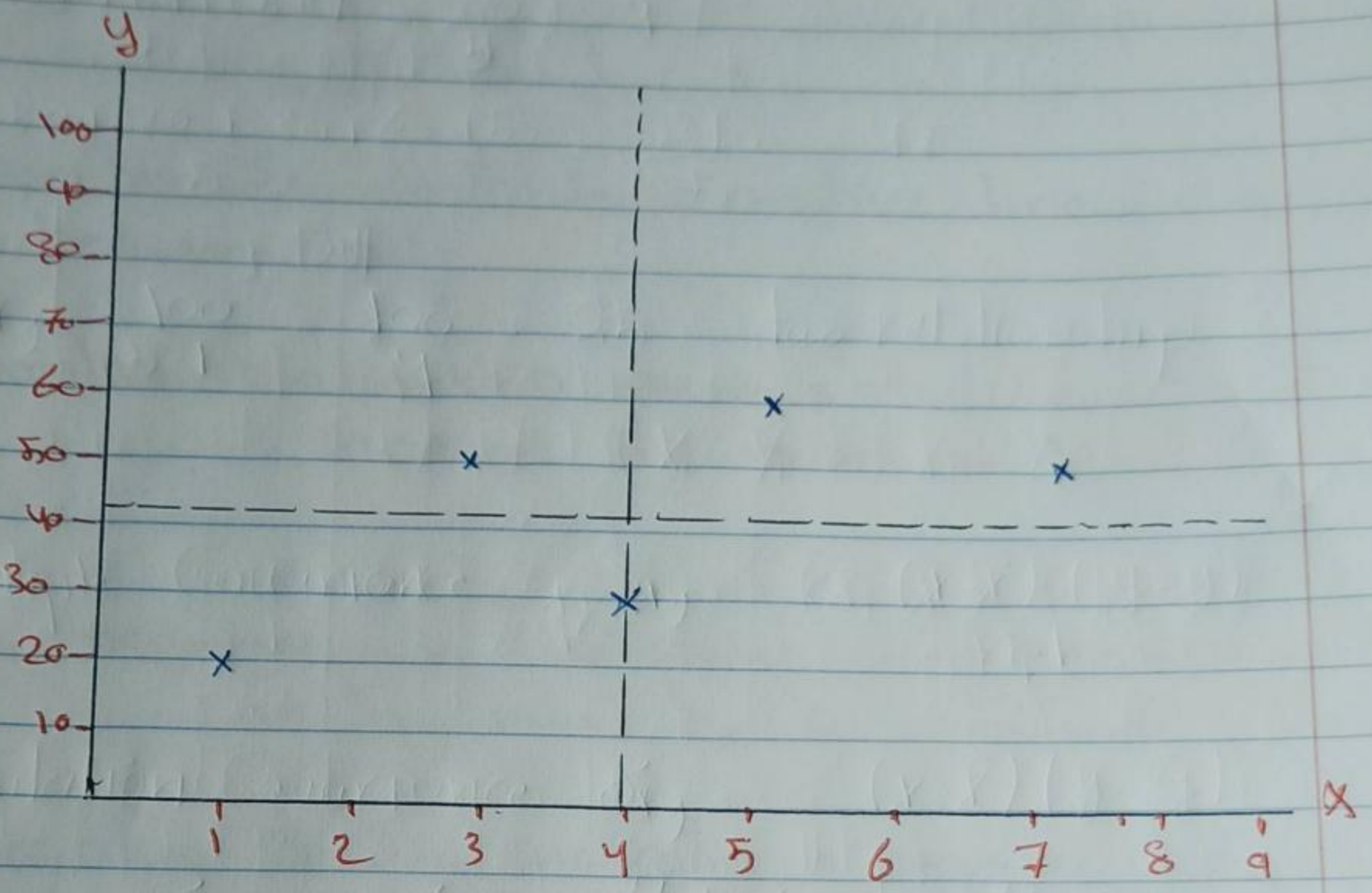
- population Covariance

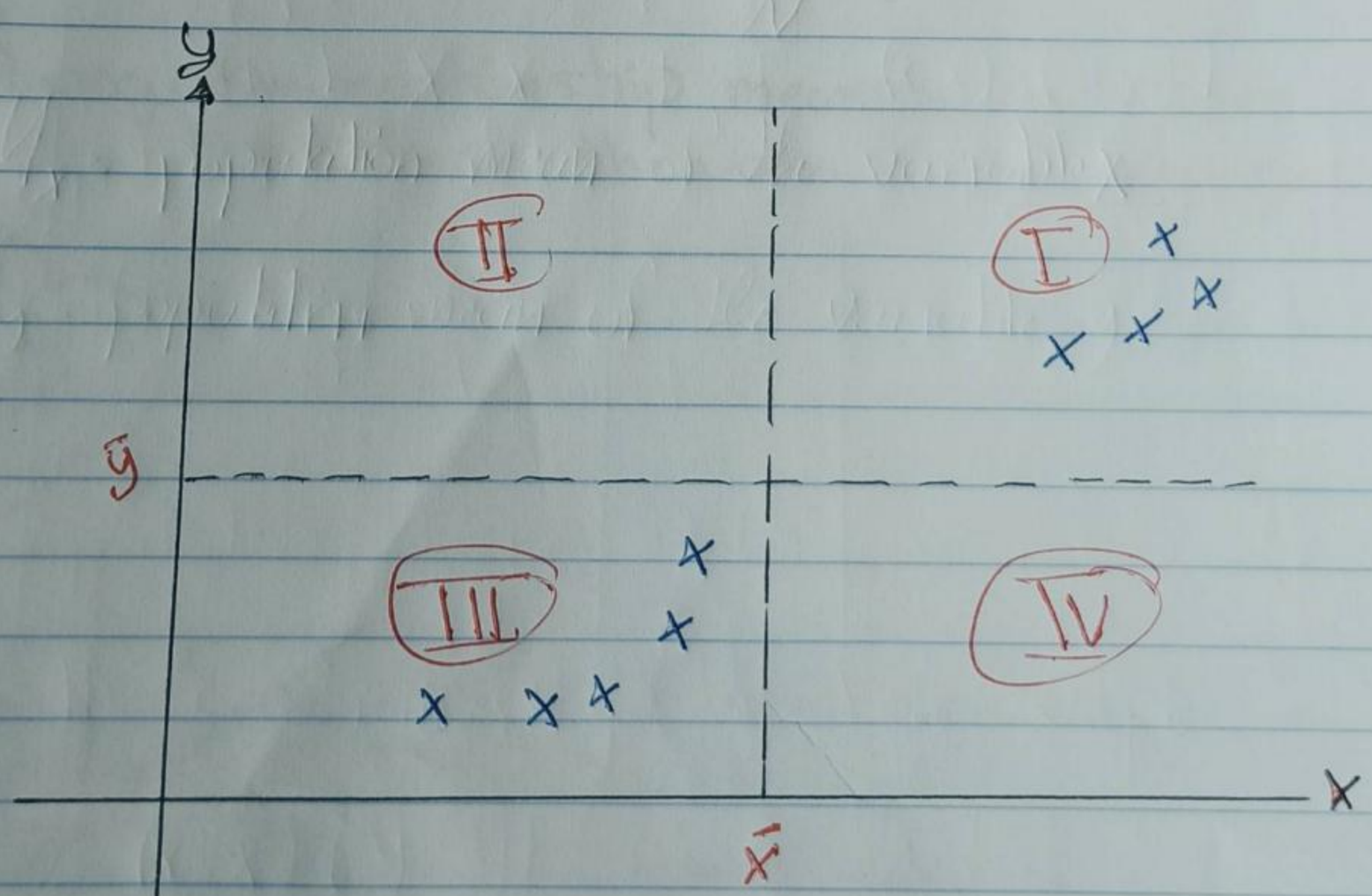
$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

μ_x = population mean of the variable x

μ_y = population mean of the variable y

- The Covariance is a measuer of linear Association between two variablee





الدرس الثاني

if S_{xy} positive ⇝ The point with greatest influence on S_{xy} must be in quadrants (I) and (III)

حسب الصفة في الكلف

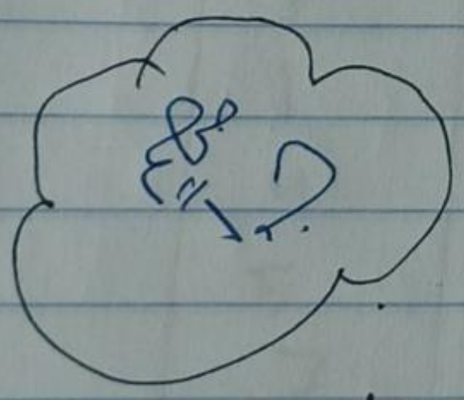
A positive S_{xy} indicates a positive linear association between x and y

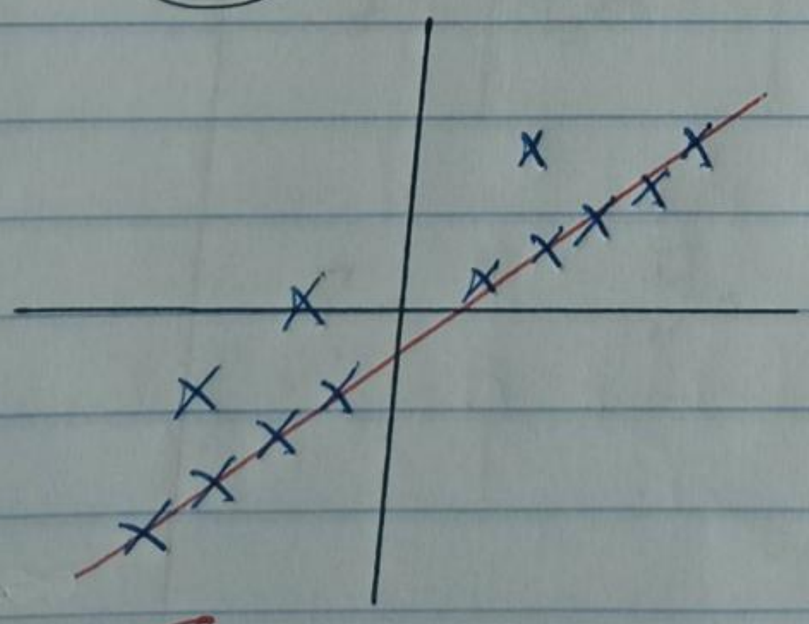
A positive relationship means ⇝ As values of X increase the values of Y also increase.

if S_{xy} negative ⇝ The point with greatest influence on S_{xy} must be in quadrants (II) an (IV)

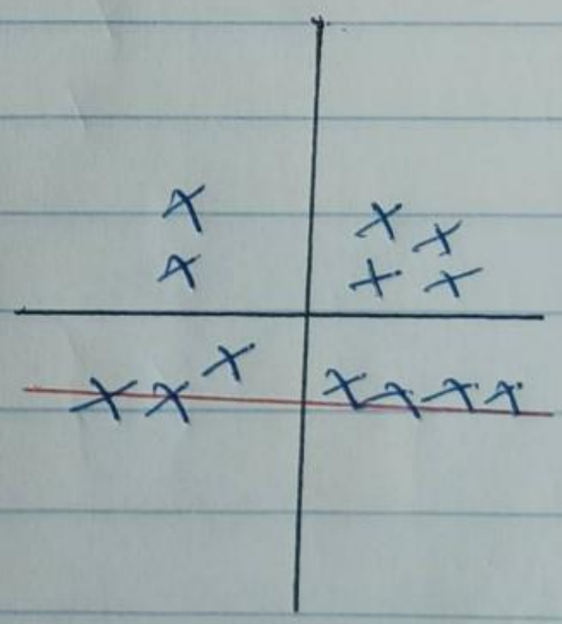
A negative S_{xy} indicates a negative linear association between X and Y

A negative relationship means ⇝ As values of X increase the value of y decreas

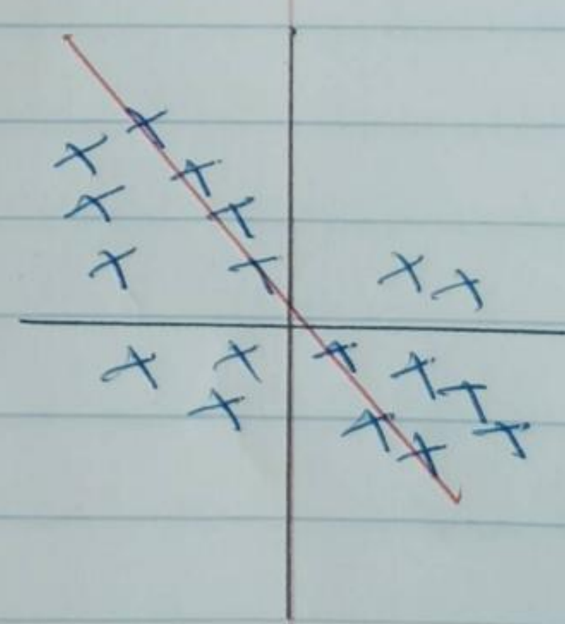




$S_{xy} = +$



$S_{xy} \approx 0$



$S_{xy} = -$

- One probleme ureing Covariance as a measuer of the strength of the linear relatron ship. is that the value of the Covariance depends on the units of measuerment of x and y

لـ وحدة الوحيدة لهذا القياس انه يتأثر بـ وحدات قوة العلاقة بين متغيرين ايضا يعتمد على الوحدة

A measure of the relationship between two variables that is not affected by the unit of measurement for x and y is the

← Correlation Coefficient مقياس لا يتأثر بالوحدة ← على العلاقة الخطية

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
3	50	-1	8	1	64
1	20	-3	-22	9	484
4	30	0	-12	0	144
5	60	1	18	1	324
7	50	3	8	9	64
				20	1080

$\bar{x} = 4 \quad \bar{y} = 42$

$$S_x \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{20}{4}} = \sqrt{5} \approx 2.23$$

$$S_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{1030}{4}} = \sqrt{270}$$

$$= 16.43$$

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = 25$$

$$\therefore r_{xy} = \frac{25}{(2.23)(16.43)} = 0.682$$

- Correlation Coefficient of population data

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Correlation Coefficient of sample data

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

↳ The Correlation Coefficient ranges from -1 to 1

- values Close to -1 ⇝ strong negative relationship
- values Close to 1 ⇝ strong positive relationship
- values Close to 0 ⇝ weake relation ship

جزء مهم جداً

- كيفية التطبيق على الآلة الحاسبة

shift → Mode , 3 , = , =

1) بتعمل تنظيف ←

2) بتضغط Mode

3) بتضغط "REG" 3

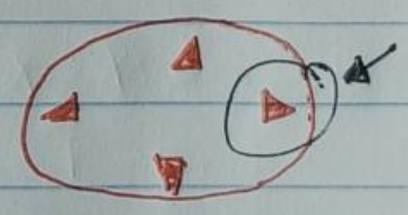
4) بتضغط "Lin" 1 ← Linear

5) ثم تدخل البيانات حسب الجدول في هذه الطريقة

x	y
3	50
1	20
4	30
5	60
7	50

بتضغط:

3 , 50 M+ n = 1
1 , 20 M+ n = 2
4 , 30 M+ n = 3
5 , 60 M+ n = 4
7 , 50 M+ n = 5

6) ثم نضغط مرتين ← 

لو ظهرت معك

A	B	r
1	2	3

Correlation Coefficient "r" ← 3 بتضغط

r_{xy}

0.680 هو الناتج معنا