

# STAT2361

## ملاحظات المحاضرات

### من إعداد موقع

## BZU-HUB



# Chapter 3

ملاحظة: المحاضرات من شرح الدكتور محمد مضية

## Chapter 3: Descriptive Statistics Numerical Measures

• parameters  $\leftrightarrow$  Population

• Statistics  $\leftrightarrow$  Sample

### 3.1: Measures of Location (Center)

• Three measures:

① Mean (Average)

② Median

③ Mode



#### ① Mean (Average)

$\bar{X} \equiv (x \text{ Bar}) \equiv \text{Sample Average}$

$\mu \equiv (\mu_n) \equiv \text{population Average}$

Raw Data  $\Rightarrow$  Average =  $\frac{\text{Sum}}{n}$

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

$n = \text{Sample size.}$

$$\mu = \frac{\sum x_i}{N}$$

$N = \text{population size.}$

• الجواب لأقرب منزلتين

## ② Median (الوسيط)

The median is the Value in the middle of arranged data Values. (In an increasing order)  
القيمة في المنتصف

How to Find The median for a raw data?

- ① arrange the values from smallest to largest
- ② Find the median position

Position: ① # of Values odd  $n$   
بترتيب للزوايد بـ بنزير 0.5  $\leftarrow$  Position =  $\frac{n+1}{2}$   
 $n = 17$        $p = 9$



② # of Values even  $m$

$\rightarrow$  2 positions:  $\frac{m}{2}$ ,  $\frac{m}{2} + 1$

median = the average of the two values in positions  $\frac{m}{2}$ ,  $\frac{m}{2} + 1$

$m = 20$        $p = 10, 11$

Median = 10.5

\*Example:

① 2, 17, 12, 15, 13, 11, 20, 21, 2

$\uparrow$ : 2, 2, 11, 12, 13, 15, 17, 20, 21

$n = 9 \Rightarrow$  Position =  $\frac{9+1}{2} = 5$

Median = 13

② 1, 7, 13, 2, 15, 18, 20, 24, 10, 25

↑: 1, 2, 7, 10, 13, 15, 18, 20, 24, 25

$$n=10 \rightarrow \text{Position} = \frac{10}{2} = \boxed{5}, \quad \frac{10}{2} + 1 = \boxed{6}$$

$$\text{median} = \frac{13 + 15}{2} = \boxed{14}$$

③ **Mode** (المنوال)

The value that occurs most frequently

القيمة الأكثر تكراراً

① 2, 2, 11, 17, 13, 2, 11, 11, 2

$$\text{mode} = 2$$

② 2, 2, 11, 17, 11, 2, 11, 11, 2

$$\text{mode} = 2, 11$$

③ 1, 3, 4, 6, 11, 13, 15

No mode



✳️ **Example:**

The following data are daily expenditure for 5 BZU student

30, 40, 25, 50, 200 → Extreme Value (outlier)

Find the:

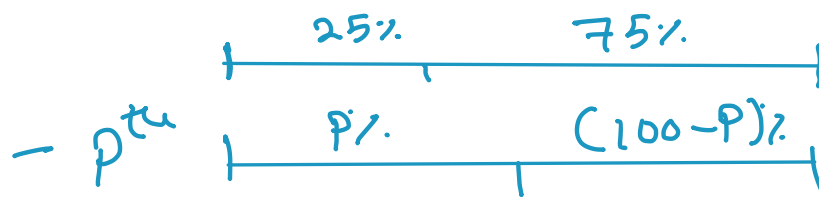
① Mean: 69

② Median: 40



## • The Percentiles (order) (Rank)

- The twenty fifth percentile  $\equiv P_{25}$



\* Median  $\equiv P_{50}$

A horizontal line with tick marks at each end and one interior point. Above the line, the segment from the left end to the interior point is labeled '50%', and the segment from the interior point to the right end is labeled '50%'.

## • How to find the $P^{\text{th}}$ Percentile?

1 Order (Sort) the data ( $\uparrow$ )

2 Compute the index  $i$

$$i = \left(\frac{P}{100}\right)(n) \quad n = \# \text{ of values}$$



3 If  $i$  is not integer, then round up to the next integer.

$$i = 13.2 \Rightarrow \text{position} \equiv 14$$

4 If  $i$  is an integer, then the  $P^{\text{th}}$  percentile is the average of the 2 values in positions:  $i, i+1$

\* Example: 14, 15, 17, 19, 20, 30 (Arranged)

$$P_{20} : \left(\frac{20}{100}\right) \times 6 = 1.2 \approx \boxed{2}$$

$$P_{50} : \left(\frac{50}{100}\right) \times 6 = 3 \text{ and } 3+1 \Rightarrow 4$$

$$P_{50} = \frac{17+19}{2} = \boxed{18}$$

## • special percentiles

\* Quartiles :-

- ① The first quartile  $\equiv Q_1 \equiv P_{25}$  (lower quartile)
- ② The second quartile  $\equiv Q_2 \equiv P_{50} \equiv \text{Median (Middle)}$
- ③ The third quartile  $\equiv Q_3 \equiv P_{75}$  (upper)



\* The Five number Summary : بيان الأرقام

Minimum,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , Maximum

\* Example (1):

Find the five number summary for the following data:

3, 5, 1, 8, 15, 9, 12, 14, 15, 35, 20

↑ 1, 3, 5, 8, 9, 12, 14, 15, 20, 35

$$n = 10$$

$$\bullet Q_1 = P_{25} = \left(\frac{25}{100}\right)(10) = 2.5 \text{ (round up)} = \boxed{3}$$

$$\bullet Q_2 = P_{50} = \left(\frac{50}{100}\right)(10) = 5 \text{ (integer)} \Rightarrow 5, 6$$

$$Q_2 = \frac{9+12}{2} = \boxed{10.5}$$

$$\bullet Q_3 = \left(\frac{75}{100}\right)(10) = 8 \Rightarrow \boxed{Q_3 = 15}$$

\* Five number summary :-

1, 3, 10.5, 15, 35

## \* 3.2 : Measures of Variation (Dispersion)

Variability

### 1) Range

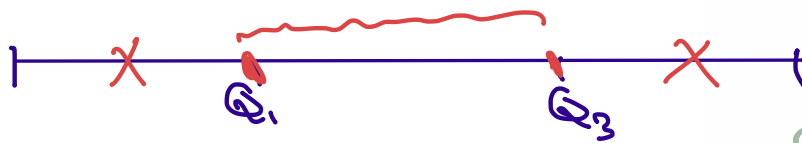
Range = Maximum Value - Minimum Value

\* Example: The range for the data of example 1 is:

$$35 - 1 = 34$$

### 2) The Interquartile range (IQR)

$$IQR = Q_3 - Q_1$$



\* Ex: Data of example (1):

$$IQR = 15 - 5 = 10$$



### 3) The Variance (التباين)

$\sqrt{\text{Variance}}$  = The Standard deviation S.D

$$(S.D)^2 = \text{Variance}$$

\* Notations :-

- $\sigma$  (Sigma)  $\equiv$  Population S.D (parameter)
- $s$  (s)  $\equiv$  Sample S.D (Statistic)

\* Formulas :-

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

- $\bar{x}$   $\equiv$  Sample mean
- $n$   $\equiv$  Sample size
- $x - \bar{x}$   $\equiv$  The deviation of the value  $x$  from  $\bar{x}$

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

- $\mu \equiv$  population mean
- $N \equiv$  population size

\*Example: Find  $S$  for the following data:

8, 7, 15, 18

$$\bar{x} = \frac{\sum x}{n} = \frac{48}{4} = \boxed{12}$$

|                  |                   |
|------------------|-------------------|
| $x - \bar{x}$    | $(x - \bar{x})^2$ |
| -4               | 16                |
| -5               | 25                |
| 3                | 9                 |
| 6                | 36                |
| <hr/>            | <hr/>             |
| Total<br>↙ ↘ ↖ ↗ | 86                |

$$S = \sqrt{\frac{86}{3}} = \boxed{5.35}$$

#### 4 The Coefficient of Variation

$$C.V = \frac{S.D}{Mean} * 100\%$$

\*Ex:  $\bar{x} = 12$  ,  $S = 5.35$

$$C.V = \frac{5.35}{12} * 100\%$$

$$= 44.5\%$$



\*Ex: Consider the following 2 Data Sets:-

|      | Set (1) | Set (2) |
|------|---------|---------|
| mean | 65      | 25      |
| S.D  | 5       | 4       |

Which Data set has more variation?

$$CV(1) = 7.84\%$$

$$CV(2) = 8.8\%$$

Data Set 2 has more Variation

Outline  
of  
chapter

## Center (Location)

Mean  
Median  
Mode

Percentiles

## Variation

Range  
IQR  
S.D

C.V

دائماً موجب

## \* Calculator

Raw Data: SD Mode

• Example data: 12, 5, 17, 8, 20

mean  
S.D  
C.V



1 Clean (Reset): Shift → Mode → 3 → = → =

2 Mode 2

3 \* Data Entry (إدخال):

12 M+ → 5 M+ → 17 M+ → 8 M+ → 20 M+

4 Statistics

• Shift → 2

$$\bar{x}: \text{Shift } 2 \quad 1 = \bar{x} = 12.4$$

$$s: \text{Shift } 2 \quad 3 = s = 6.19$$

$$\sigma: \text{Shift } 2 \quad 2 = \sigma = 5.84$$

• علاقة ال (S.D) بال (mean)

$$\boxed{1} \text{ C.V} = \frac{s}{\bar{x}} * 100 \%$$

$\boxed{2}$  The Z-Score (measure of Variation)  
 ↳ Standard القِيم

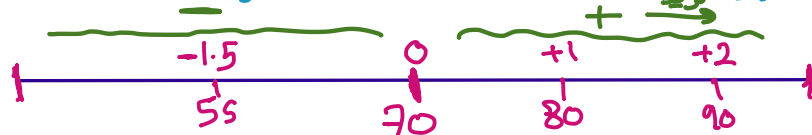
• Given a data Set with mean  $\mu$  and Standard deviation  $\sigma$ , then the Z-Score for a Value  $x$  is defined by:

$$Z = \frac{x - \mu}{\sigma} \rightarrow \text{Deviation}$$

Ex  $\mu = 70 \quad \sigma = 10$

Find the Z-Score for  $x = 90, 80, 70, 55$

$$Z_{90} = +2, \quad Z_{80} = +1, \quad Z_{70} = 0, \quad Z_{55} = -1.5$$



• Z-Score = # of S.D for a Value ( $x$ ) from the mean ( $\mu$ )

$$\begin{aligned} 80 &= 50 + 30 \\ &= 50 + (2)(15) \\ &= \mu + 2\sigma \end{aligned}$$

$$\begin{aligned} 20 &= 50 - 30 \\ &= 50 - 2(15) \\ &= \mu - 2\sigma \end{aligned}$$

\* Empirical Rule  $\Rightarrow$  % of data

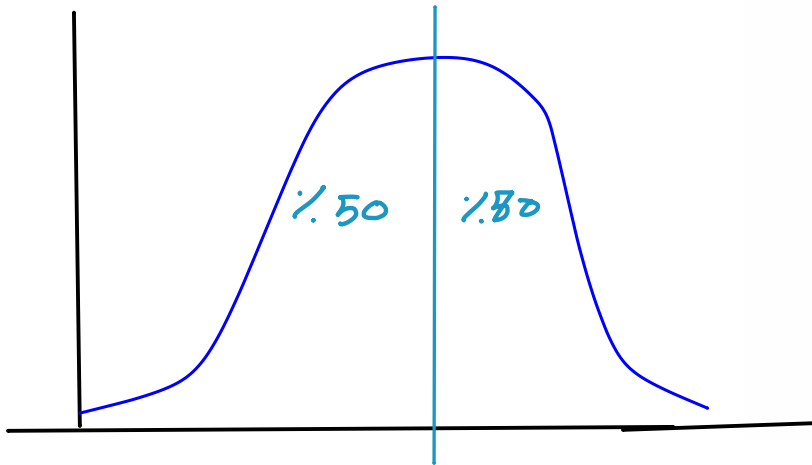


### 3.3: Skewness, z-score, Empirical Rule, Outliers

50 سے قبل ←

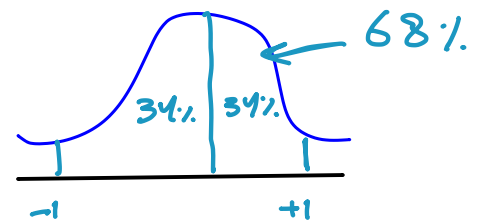
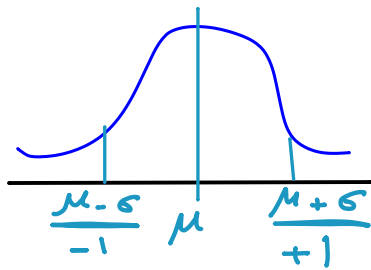
#### • The Empirical Rule

↳ Bell-Shape data ⇒ Symmetric

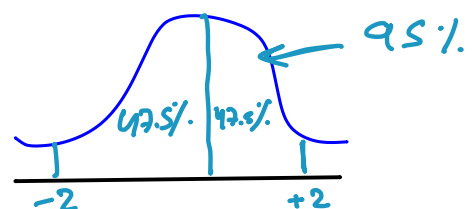
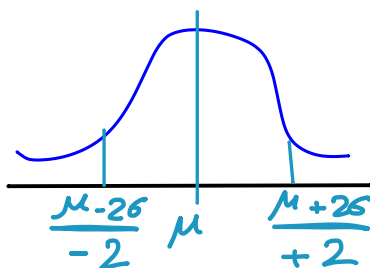


\* For a bell-shaped data

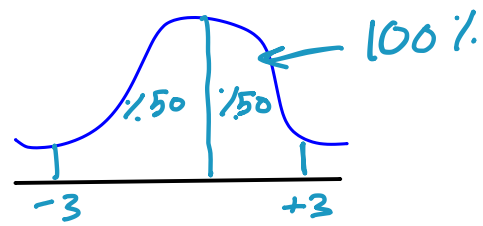
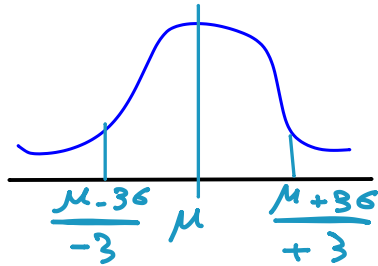
- 1 Approximately 68% of the data values are within 1 standard deviation from the mean



- 2 Approximately 95% of the data values are within 2 standard deviation from the mean



- 3 Almost all data values are within 3 S.D from the mean



● بناءً على Z-Score يمكن أن نعلم أن 95% من البيانات تقع بين  $z = \pm 1, \pm 2, \pm 3$  ، باقى الأرقام في سابت (6) له بالامثلة 2 يكون

### \* Example:

The IQ Scores are bell-shaped with  $\mu$  mean = 100 and S.D = 15

- A What is the percentage of IQ scores between:-

- 1 85 to 115

$$z_{115} = \frac{115 - 100}{15} = +1, \quad z_{85} = -1$$

$$\% = 68\%$$

- 2 70 to 130

$$z_{130} = +2, \quad z_{70} = -2$$

$$\% = 95\%$$

- 3 55 to 145

$$z_{55} = -3, \quad z_{145} = +3$$

$$\% = 100\%$$

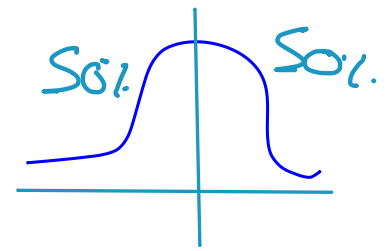




4 Greater than 100

$$\text{mean} = 100 \Rightarrow z_{100} = 0$$

$$\% = 50\%$$

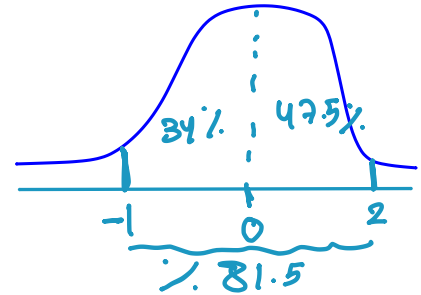


5 85 to 130

$$\downarrow$$
$$-1$$

$$\downarrow$$
$$+2$$

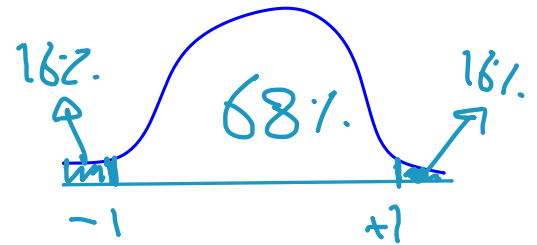
$$\% = 81.5\%$$



6 Greater than 115

$$\downarrow$$
$$z = 1$$

$$\% = 16\%$$



7 Less than 150

$$z = 3.33$$

$$\% = 100\%$$

B Is it possible to find a person with IQ score 150 or greater?

بالطبع ممكن

\* Extreme Value (Outliers)

Very Small

Very Large



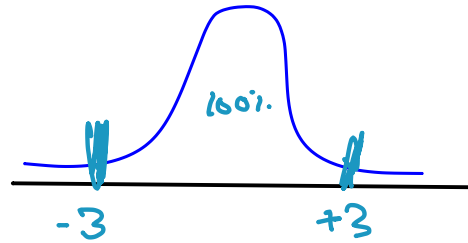
\* Why there is outliers?

1 incorrectly recorded data value

2 incorrectly included data value

3 unusual large (small) data value

\* How to determine the outliers?



$$\left. \begin{array}{l} z \geq 3 \\ z \leq -3 \end{array} \right\} \Rightarrow \text{outliers}$$

$$\frac{\mu - 3\sigma}{\text{أصغر}} , \frac{\mu + 3\sigma}{\text{أكبر}} \Rightarrow \text{outliers}$$

$$|z| \geq 3$$



### 3.4: The Box Plot

Graph  $\rightarrow$  Summarize data  
 $\rightarrow$  outliers

\* All data

\* 5 number Summary

\* To construct a box plot

1 Find:  $Q_1, Q_2, Q_3$

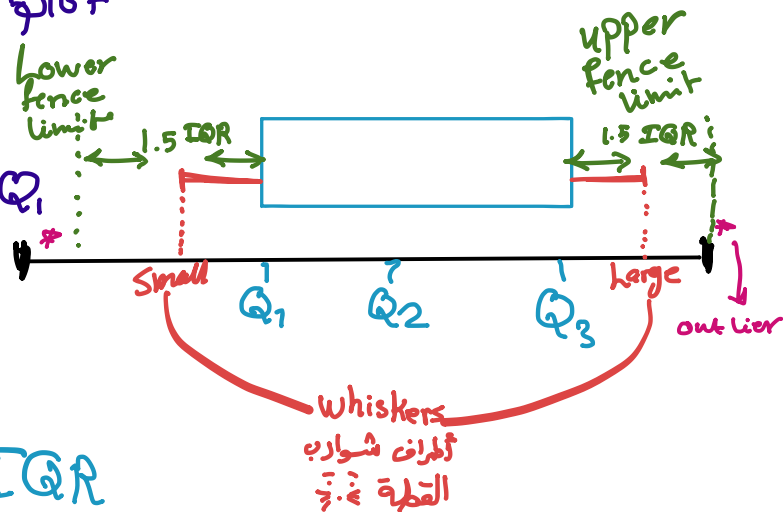
2 Find  $IQR \Rightarrow Q_3 - Q_1$

3 Find  $(1.5)(IQR)$

4 Find fence limits

$$\text{lower} = Q_1 - 1.5 IQR$$

$$\text{upper} = Q_3 + 1.5 IQR$$



● Outliers When:

$$x \geq \text{upper limit} = Q_3 + 1.5 IQR$$

$$x \leq \text{lower limit} = Q_1 - 1.5 IQR$$

### Ex

Construct a Box Plot for the following data:

1, 7, 20, 21, 22, 25, 30, 32, 42, 55

↑  
arranged data

•  $Q_1 = P_{25} \Rightarrow i = \left(\frac{25}{100}\right) 10 = 2.5 \Rightarrow 3$   $Q_1 = 20$

•  $Q_2 = P_{50} \Rightarrow 5 = i \Rightarrow$  positions 5, 6

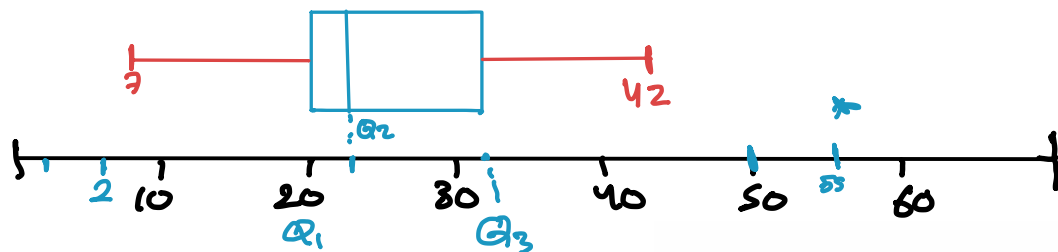
$\underline{Q_2} = \frac{22 + 25}{2} = \underline{23.5}$

•  $Q_3 = i = 7.5 \Rightarrow 8$  ,  $Q_3 = 32$

•  $IQR = Q_3 - Q_1 = 12$

•  $1.5 IQR = 18$

Whiskers: أعلى وأقل قيمة  
فصل الحدود



• Lower =  $20 - 18 = 2$

• Upper =  $32 + 18 = 50$



### 3.6: The weighted mean and grouped data

Frequency Dist.

#### • The Weighted mean:-

Give the values  $x_1, x_2, \dots, x_n$  with frequencies  $f_1, f_2, \dots, f_n$  respectively, then the (Weighted) mean is:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

\*Ex: Find the mean for the following distribution:

| <u>x</u> | <u>f</u>  | <u>x.f</u> |
|----------|-----------|------------|
| 10       | 15        | 150        |
| 12       | 5         | 60         |
| 18       | 12        | 216        |
| 20       | 8         | 160        |
|          | <u>40</u> | <u>586</u> |

$$\bar{x} = \frac{586}{40} = 14.65$$

#### • The mean for the grouped data

\*Ex: Find the mean for the following distribution

| <u>Class</u> | <u>Frequency</u> | <u><math>m_i</math></u> | <u><math>m_i \cdot f</math></u> |
|--------------|------------------|-------------------------|---------------------------------|
| 5-9          | 6                | 7                       | 42                              |
| 10-14        | 10               | 12                      | 120                             |
| 15-19        | 7                | 17                      | 119                             |
| 20-24        | 8                | 22                      | 176                             |
| 25-29        | 4                | 27                      | 108                             |
|              | <u>35</u>        |                         | <u>565</u>                      |



$$\bar{x} = \frac{\sum m_i f_i}{\sum f_i}, \quad m_i = \text{The midpoint for the interval } i$$

$= \frac{U+L}{2}$

$$\bar{x} = \frac{565}{35} = 16.14$$

- S.D for grouped data:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 \cdot f_i}{(\sum f) - 1}}$$

### Calculator :

1 Reset (clean)  
Shift mode 3 = =

2 Mode 2

3 Data Entry ← *أهم خطوة*

بنكرها حسب ال Values  $\mu + 15$  و Shift  $10$  ← Value  
*من*

لازم يطالع آخر 1  $n=40$  *From*

4  $\bar{x}$  : Shift 2 1 = 14.65  
 $s_x$  : Shift 2 3 = 4.3

① في الآلة كسبة بنحو المثل  $\mu$  و  $s$   
لزم يطالع 35

Shift 2 1 = 16.14

$s =$  Shift 2 3 = 6.47



## 3.5 + 12.2 : Measures of Association between 2 Variables

$X$ : Independent (price) (Income)  
 $Y$ : Dependent (quantity) (Daily expenditure)

- Select a Sample Size  $n$ 
  - $\Rightarrow (x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$
  - $\Rightarrow$  Relationship between the 2 Variables (Linear)
- Graph = Scatter Diagram  $\rightarrow$  indicator (مؤشر)

### \* Numerical Measures <sup>مستقلة</sup> $E = \bar{N}_0$

- 1 The Covariance
- 2 The Correlation  $\rightarrow$  strength of the relation
- 3 Regression  $\rightarrow$  Trendline equation

#### 1 The Covariance

$s_{xy}$  = Sample Covariance

$\sigma_{xy}$  = population Covariance

\* Formulas:

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

$$\sigma_{xy} = \frac{\sum (x - \mu_x) \cdot (y - \mu_y)}{N}$$

$s_{xy} > 0 \Rightarrow +$  relationship

$< 0 \Rightarrow -$  relationship

$= 0 \Rightarrow$  No relationship



## 2 The Correlation Coefficient (معامل الارتباط)

↳ Strength of the relation

$r_{xy}$  = Sample C.C

$\rho_{xy}$  = population C.C

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

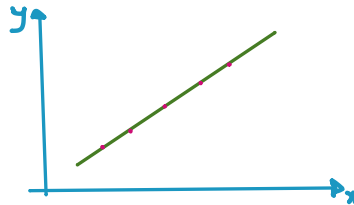
$S_{xy}$  = Sample Covariance  
 $S_x$  = Standard deviation  $x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$   
 $S_y$  = S.D for  $y$

$$-1 \leq r_{xy} \leq 1$$

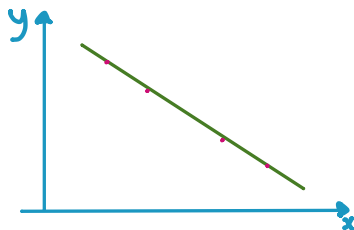


•  $r_{xy} = 0 \rightarrow$  no relationship

•  $r_{xy} = 1 \rightarrow$  perfect positive relationship



•  $r_{xy} = -1 \rightarrow$  perfect negative relationship



•  $|r_{xy}| \xrightarrow{\text{تقريب}} 0$   
(Weak) تست کبر

•  $|r_{xy}| \rightarrow 1$   
(Strong) تست اقل



### 3 Regression

↳ equation of the trend line

Regression equation

Trend line  $\Rightarrow$   $y = mx + b$

Method: Least Squared Method

$\hat{y}$   $\equiv$  estimated regression line

$\hat{y} = b_0 + b_1x = A + Bx \rightarrow$  Calculator

$$b_1 = B = \frac{\sum xy}{\sum x^2}, \quad b_0 = \bar{y} - b_1 \bar{x}$$



Example :-

Consider the following distribution

|          | $x$ | $y$ | $x - \bar{x}$ | $(x - \bar{x})^2$ | $y - \bar{y}$ | $(y - \bar{y})^2$ | $(x - \bar{x}) \cdot (y - \bar{y})$ |
|----------|-----|-----|---------------|-------------------|---------------|-------------------|-------------------------------------|
|          | 2   | 3   | -2            | 4                 | -2            | 4                 | 4                                   |
|          | 4   | 5   | 0             | 0                 | 0             | 0                 | 0                                   |
|          | 6   | 7   | 2             | 4                 | 2             | 4                 | 4                                   |
| $\Sigma$ | 12  | 15  | 0             | 8                 | 0             | 8                 | 8                                   |

①  $\bar{x} = 4$  , ②  $\bar{y} = 5$

③  $\sigma_x = \sqrt{\frac{8}{2}} = 2$  ④  $\sigma_y = \sqrt{\frac{8}{2}} = 2$

⑤  $\sigma_{xy} = \frac{8}{2} = 4 > 0$

⑥  $r_{xy} = \frac{4}{2 \cdot 2} = 1 \rightarrow$  Strong perfect positive relationship

⑦  $b_1 = \frac{4}{2^2} = 1$

⑧  $b_0 = 5 - (1)(4) = 1$

⑨  $\hat{y} = b_0 + b_1x = 1 + x$



## \*Example (for Calculator)

| $x$ | $y$ |
|-----|-----|
| 8   | 17  |
| 10  | 22  |
| 4   | 10  |
| 5   | 12  |
| 7   | 15  |
| 14  | 30  |



• Use Reg Mode to find:-

- ①  $\bar{x}$     ②  $s_x$     ③  $\bar{y}$     ④  $s_y$     ⑤  $s_{xy}$     ⑥  $r_{xy}$     ⑦  $b_0$     ⑧  $b_1$

⇒ Steps:-

① Reset

② Reg Mode ⇒ mode 3 1 (lin)

③ Data Entry:

$x, y$  M+

8, 17 M+  $n=1$

⋮

14, 30 M+  $n=6$

④ Statistics:-

$x$ :  $\bar{x} \rightarrow \text{Shift } 2 \ 1 = 8$  ①

$s_x \rightarrow \text{Shift } 2 \ 3 = 3.63$  ②

$y$ :  $\bar{y} \rightarrow \text{Shift } 2 \ 1 = 17.67$  ③

$s_y \rightarrow \text{Shift } 2 \ 3 = 7.34$  ④

$x \ z \ y$ : Shift 2 2 2

$A(b_0) \rightarrow \text{Shift } 2 \rightarrow 1 = 1.54$  ⑦

$B(b_1) \rightarrow \text{Shift } 2 \rightarrow 2 = 2$  ⑧

$r_{xy} \rightarrow \text{Shift } 2 \rightarrow 3 = 0.99$  ⑥

\*  $r_{xy} = \frac{S_{xy}}{s_x \cdot s_y} \rightarrow$  بعضها يدعى

$S_{xy} = s_x \cdot s_y \cdot r_{xy} = 26.58$  ⑤

9)  $\sum xy = 26.58 > 0 \Rightarrow$  Positive Relationship

10)  $r_{xy} = 0.99 \Rightarrow$  Strong Positive Relationship

11) Equation of the Line

$$\hat{y} = b_0 + b_1 x$$

$$= A + Bx$$

$$\hat{y} = 1.54 + 2x$$

12) Find the estimated value of  $y$  if  $x=30$

$$\hat{y} = 1.54 + 2(30) \rightarrow \text{ليه بنسوي هاي للمادة؟}$$
$$= 61.54 \quad \text{- عشان نتوقع لبعدين}$$

