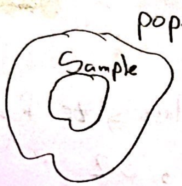
* Ch 3: Descriptive Statistics: Numerical Measures.



سواء كانت البيانات مأخوذة من عينة أو مجتمع هذه القراءات الرقمية التي نجدها من هذه البيانات تكون ضمن علم الإحصاء الوصفي «Descriptive statistics».

- Measures are computed for data from a sample: sample statistic.
- Measures are computed for data from a population: parameters.
- A sample statistic is the point estimator of the corresponding population parameter. } inferential statistic هنا تكون ضمن

أي قراءتها المتعلقة بـ population تسمى parameter.
// // // // sample تسمى statistic.

عند صعوبة دراسة المجتمع الكامل «pop» نلجأ إلى عينة «sample» وتعتبر قراءات العينة نقاط تقريبية لقراءات المجتمع كامل.

- Sec 3.1: Measures of location:-

↓

Measures of central tendency مقاييس الموضع أو مقاييس النزعة المركزية.

مقاييس تستخدم لقياس موضع تركز أو تجمع البيانات.

① Mean: «Arithmetic average» الوسط الحسابي

- If the data are for a sample, the mean is denoted by $\overline{X}$.
- If the data are for a population, the mean is denoted by μ.

→ the sample mean $\overline{X} = \frac{\sum x_i}{n}$; Σ: the summation sign إشارة المجموع , x_i: the data values , n: the sample size.

→ the population mean $\mu = \frac{\sum X_i}{N}$; N: the population size.

∴ الوسط الحسابي = $\frac{\text{مجموع القيم}}{\text{العدد الكلي}}$

-Ex: The following data represents the class size for a sample of 5 classes. 46, 54, 42, 46, 32

Find the sample mean.

$$\longrightarrow \bar{X} = \frac{46+54+42+46+32}{5} = 44$$

-Note: $\bar{X}$ is a point estimator for M.

(2) Median: الوسيط

The value in the middle.

• To find the median:

a- order the data (from smallest to largest). أو العكس

b- • If the size ((n)) is odd, the median is the middle value.

• If the size ((n)) is even, the median is the average of of two middle values.

-Ex: Given the sample: 46, 54, 42, 46, 32. Find the median.

n = 5 odd

ترتيب البيانات

$\longrightarrow$ 32, 42, $\boxed{46}$, 46, 54

the median is 46

بإمكاننا بعد ترتيب البيانات إيجاد رتبة الوسيط بالطريقة التالية
إذا n فردي $\frac{n+1}{2}$
ونبحث عن القيمة

- **Ex:** Given the sample: 13, 10, 15, 10, 10, 9, 12, 14,
Find the median.

→ $n = 8$ «even»

$\underline{9}$, $\underline{10}$, $\underline{10}$, $\boxed{10 \; , 12}$, $\underline{13}$, $\underline{14}$, $\underline{15}$ نرتب البيانات

the median $= \dfrac{10 + 12}{2} = \boxed{11}$

the rank of the median $= 4, 5$

بإمكاننا إيجاد رتبة الوسيط إذا كان العدد زوجي :-
$\frac{n}{2}$, $\frac{n}{2} + 1$

* Note that: 50% of data value are less than or equal the median, and 50% are greater than or equal the median.

* Extreme values: outliers القيم المتطرفة أو الشاذة تكون بعيدة عن البيانات.

- **Ex:** ① Given the sample: 4, 2, 1, 5, 4

→ 1, 2, $\boxed{4}$, 4, 5

the median is $\underline{4}$

② Given the sample: 4, 2, 1, 95, 4

→ 1, 2, $\boxed{4}$, 4, 95

the median is $\underline{4}$

نلاحظ أن البيانات في المثال ① لا يوجد فيها قيم متطرفة وكان Median = 4 ، في المثال ② كانت 95 عبارة عن قيمة متطرفة وبقي Median = 4 .

← الوسيط لا يتأثر بالقيم المتطرفة.

- Note that the median isn't affected by extreme values, but the mean على عكس الوسيط الحسابي is affected.

3. Mode: أكثر البيانات شيوعاً (تكراراً). المنوال.
the highest frequency data (that is, the data occurs with greatest frequency).

- Ex: (A) Given the sample: 4, 2, 2, 1, 3, 10.
→ The mode is 2. تكررت مرتين

(B) Given the sample: 4, 2, 2, 1, 4, 10
→ the mode is 2, 4 2,4: تكررت مرتين

(C) Give the sample: 4, 2, 3, 1, 10.
→ there ~~no~~ is no mode جميع البيانات لها نفس التكرار ∴ لا يوجد منوال.

- Note that:
We could have no modes, 1 mod, 2 modes, 3 modes, ...

- If the sample has 1 mode: unimodal data.
- If the sample has 2 modes: bimodal data.
- If the sample has more than 2 modes: multimodal.

- Ex: The following data represents the blood type for 10 students. A, O, O, O, AB, O, B, B, A, O
→ The mode is O. لكن في هذا المثال لا نستطيع ايجاد mean and median.

- Note that: If the data is Qualitative, then we can just find the mode. في البيانات النوعية فقط نستطيع إيجاد mode

(4) Percentiles: المئينات توضح لنا سرعة تزايد البيانات من أقل قيمة لأعلى قيمة

the pth percentile is a value from the orderd such that at least p% are less than or equal to this value and at least (100 − p)% are greater than or equal to this value.

→ To find the pth percentile:-

1) order the data from smallest to largest.

2) Compute $i = \left(\frac{P}{100}\right) n$; p: the percentile.
n: the size.

3) If i isn't integer, round up: العدد الذي يليه "نقربه" هو موقع percentile

If i is integer, the pth percentile $= \frac{X_i + X_{i+1}}{2}$ القيمة التي موقعها i والقيمة التي موقعها i+1 نأخذ معدلهم.

Ex: Given the sample:

3450, 3550, 3650, 3480, 3355, 3310, 3490, 3730, 3540, 3925, 3520, 3480.

a) Find the 85th percentile « P_{85} »

أولاً نرتب البيانات

→ 3310, 3355, 3450, 3480, 3480, 3490, 3520, 3540, 3550, 3650, 3730, 3925

$i = \frac{P}{100} \times n = \frac{85}{100} \times 12 = 10.2 \xrightarrow{\sim} 11$ نجد i

∴ the 85th percentile is 3730. ← موقعها رقم 11
« P_{85} »

لتوضيح العنوان: 85% من البيانات أقل من 3730 و 20% أكثر منها

20% 80%
أقل 3730 أكثر

b) Find the 50th percentile « P_{50} »

$$i = \frac{50}{100} \times 12 = \underline{\underline{6}}$$ عدد صحيح ← معدل القيمة السادسة والسابعة

∴ the 50th percentile $= \dfrac{3490+3520}{2} = 3505$
↳ P_{50}

50% 50%
3505

– Note that: the 50th percentile is the median.

5 Quartiles: الربيعات

لو جزأنا البيانات بعد ترتيبها الى أربعة أجزاء كل جزء يأخذ 25% تقريباً.

– The division points are the quartiles:

25% 25% 25% 25%
Q_1 Q_2 Q_3

Q_1 = first quartile = P_{25} « 25th percentile »
Q_2 = Second quartile = P_{50} « 50th percentile » = median
Q_3 = third quartile = P_{75} « 75th percentile »

– Ex: Give the sample: 1, 4, 4, 3, 7, 3

a) Find the first quartile. → 1, 3, 3, 4, 4, 7 نرتب البيانات

$Q_1 = P_{25}$
$\rightarrow i = \frac{25}{100} \times 6 = 1.5 \longrightarrow \underline{\underline{2}}$

∴ $Q_1 = \underline{\underline{3}}$ القيمة الي موقعها 2

b) Find the second quartile.
$Q_2 = P_{50}$
$i = \frac{50}{100} \times 6 = 3$

∴ $Q_2 = \dfrac{3+4}{2} = 3.5$ معدل القيمة الثالثة مع القيمة الرابعة

c) Find the third quantile.

$Q_3 = P_{75}$

$i = \frac{75}{100} \times 6 = 4.5 \rightarrow \underline{5}$ القيمة التي موقعها 5

$\therefore Q_3 = 4$

$1, \boxed{3}, 3, \boxed{4}, 4, \boxed{4}, 7$

Q_1 Q_2 Q_3

- Sec 3.2 : Measures of variability: مقاييس التشتت

«Measures of dispersion»

تقيس هذه القراءات مقدار تشتت وتباعد القيم ، حيث أن مقاييس النزعة المركزية لوحدها لا تكفي ولا تعطي صورة كاملة عن البيانات لذا يجب قياس درجة تشتت القيم وتبعثرها.

1 Range المدى

→ Range = largest value − smallest value

المدى = أكبر قيمة − أصغر قيمة

- Ex: Give the sample: 52 , 33 , 75 , 80 , 91 , 45.
Find the range.

→ Range = 91 − 33 = 58

- Note:
The range is affected by outliers. بشكل كبير جداً

2 Interquartile range (IQR): نصف المدى الربيعي

the range for the middle 50% of the data.
« the difference between the third quartile Q_3 and the first quartile » المجال الذي تنتشر فيه 50% من البيانات

→ $IQR = Q_3 - Q_1$

- Ex: Given the sample 20, 40, 10, 25, 10.
Find the interquartile range.

$\rightarrow Q_3 = P_{75}$

$i = \frac{75}{100} \times 5 = 3.75 \rightarrow 4 \rightarrow$ القيمة التي موقعها 4 بعد ترتيب البيانات تصاعدي

$10, 10, 20, \boxed{25}, 40$

$\therefore Q_3 = 25$

$Q_1 = P_{25}$

$i = \frac{25}{100} \times 5 = 1.25 \rightarrow 2$ القيمة التي موقعها 2 بعد ترتيب البيانات تصاعدي

$10, \boxed{10}, 20, 25, 40$

$\therefore Q_1 = 10$

Now $IQR = Q_3 - Q_1 = 25 - 10 = \underline{15}$

- Note:
IQR isn't affected by outliers.

3 Variance: التباين
متوسط مجموع مربعات انحرافات القيم عن وسطها الحسابي

the average of the squared deviations about the mean. «measure whether the data cluster around the mean».

CamScanner الممسوحة ضوئيا بـ

→ Population variance « δ^2 » = $\frac{\sum(x_i - M)^2}{N}$, x_i: data values.
M: pop mean.
N: pop size

→ Sample variance « S^2 » = $\frac{\sum(x_i - \overline{x})^2}{n-1}$; $\overline{x}$: Sample mean.
n: sample size.

صيغة أخرى
$S^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}$

→ S^2 is a point estimator of δ^2.

- Ex: Given the sample: 46, 54, 42, 46, 32.

→ $\overline{x} = \frac{46+54+42+46+32}{5} = 44$.

x	$x - \overline{x}$	$(x-\overline{x})^2$
46	2	4
54	10	100
42	-2	4
46	2	4
32	-12	144

Now $S^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{256}{4} = \boxed{64}$

- Note:
$\sum(x - \overline{x}) = 0$ « the sum of the diveations about the mean = 0 always ».

4) Standard diviation : الانحراف المعياري

الجذر التربيعي لمتوسط مجموع مربعات انحرافات القيم عن وسطها.

the positive square root of the variance.

→ Sample standard deviation « S » = $\sqrt{S^2}$

→ Population standard deviation « σ » = $\sqrt{\sigma^2}$

→ the sample standard deviation is a point estimator of the population standard deviation.

- Ex: Find the standard deviation for the last example.

→ $S = \sqrt{64} = 8$

- **Note:**
The variance and standard deviation are affected by outliers.

5) Coefficient of variation. معامل الاختلاف.

مقاييس التشتت تشير إلى مدى تقارب أو تباعد البيانات عن الوسط الحسابي. أما بالنسبة لمعامل الاختلاف فهو يستخدم للمقارنة بين مجموعتين مختلفتين.

→ the coefficient of variation is a relative measure of variability, it measures the standard deviation relative to the mean.

→ Coefficient of variation « CV » $= \dfrac{\text{Standard deviation}}{\text{mean}} \times 100\%$

- Ex: The mean and standard deviation for the grades of 2 sections in Stat 2311 were:

Section 1 : mean = 64.87 , standard deviation = 14.8.
Section 2 : mean = 70.45 , standard deviation = 12.71

Find the coefficient of variation , then determine which section has more variable grades.

→ Section 1: $CV = \dfrac{14.8}{64.87} \times 100\% = 22.81\%$

Section 2: $CV = \dfrac{12.71}{70.45} \times 100\% = 18.04\%$

- Section 1 has more variability in grades than section 2 .

- كلما كان CV أقل كانت البيانات متقاربة على بعضها مقارنة بالمجموعات الأخرى .

* Using calculator:-

1) to find the mean: sample بغض النظر كانت
population ولا

$\boxed{\text{Mode}}$ $\boxed{2}$

x_1 $\boxed{M+}$

x_2 $\boxed{M+}$

⋮

x_n $\boxed{M+}$

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{1}$ $\boxed{=}$

2) to find the sample standard deviation.

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{3}$ =

3) to find the population standard deviation.

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{2}$ =

4) to find the variance.

Just square $\underline{S}$ or $\underline{\sigma}$.

- Ex: Given the data 25, 30, 47, 80, 56, 62, 74, 80

a) Find the mean. $(\bar{x}, \mu)$.

$\boxed{\text{Mode}}$ $\boxed{2}$

25 $\boxed{M+}$ 30 $\boxed{M+}$ 47 $\boxed{M+}$ 80 $\boxed{M+}$ 56 $\boxed{M+}$
62 $\boxed{M+}$ 74 $\boxed{M+}$ 80 $\boxed{M+}$

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{1}$ $= 56.75$

b) Find the sample standard deviation. «S»

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{3}$ $= 21.47$

c) Find the pop. standard deviation. «σ»

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{2}$ $= 20.08$

d) Find the sample variance. «S^2»

$$S^2 = 21.47^2 = 460.96.$$

e) Find the pop. variance «σ^2»

$$\sigma^2 = 20.08^2 = 403.21$$

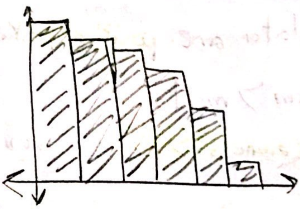
— Sec 3.3 : Measures of distribution shapes Relative location, and detecting outliers :-

هناك قراءات تحدد لنا شكل التوزيع للبيانات.

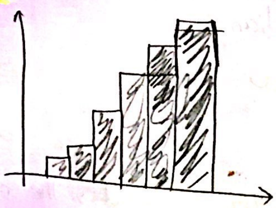
* Skewness :

A histogram is a graphical dispaly showing the shape of a distribution. We have 3 main distribution shape :-

1) positively skewed



2) negatively skewed.



3) Normal (Bell-shaped)

«Symmetric»

→ Skewness: is a measure of distribution shape.

* The formula for the skewness of sample data =

$$\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s}\right)^3$$ } القانون: غير مطلوب

• For a symmetric distribution (normal or bell shaped):

mean = median = mode (at the center).

skewness = 0

• When the data are positively (right) skewed:

mean > median

skewness > 0 (+ve)

• when the data are negatively (left) skewed:-

mean < median

skewness < 0 (-ve).

– Note:

When data are highly skewed, the median is better than the mean.

* Z-Scores:

بالإضافة الى مقاييس التشتت ومقاييس الموضع وشكل التوزيع ، هناك قراءات توضح موقع البيانات بالنسبة لبعضها البعض.

Measures of relative location help us determine how far a particular value is from the mean.

Suppose a sample of n observations $\longrightarrow x_1, x_2, ---, x_n$ and assume the sample mean is $\bar{x}$ and standard deviation s

$\rightarrow$ Z-score:

$$Z_i = \frac{x_i - \bar{x}}{s}$$. «standardized value»

« the number of standard deviations x_i from the mean $\bar{x}$ ».

مقدار بعد القيمة عن الوسط الحسابي وهذا المقدار يكون كنسبة من الانحراف المعياري ،

- Ex: The following data represent the number of students in 5 classes. 46, 54, 42, 46, 32

a) Find the Z-score of the class size 46. $\left.\begin{array}{l}\bar{x} = 44 \\ S = 8\end{array}\right\}$ from the calculator

$$Z = \frac{46 - 44}{8} = 0.25$$

this indicates that $x_1 = 46$ is 0.25 standard deviations greater than the mean

أي أن $x=46$ تبعد عن الوسط الحسابي بمقدار 0.25 من الانحراف المعياري ،

-Notes:-

① if $X_i > \overline{X} \longrightarrow Z_i > 0$

② if $X_i < \overline{X} \longrightarrow Z_i < 0$

③ if $X_i = \overline{X} \longrightarrow Z_i = 0$

• If Z_i and $\overline{X}$, S are given then we can find X_i by:-

$$Z_i = \frac{X_i - \overline{X}}{S} \longrightarrow \boxed{X_i = Z_i S + \overline{X}}$$

-Ex: If the mean = 44.81 and standard deviation = 5.41, Find the data value whose Z-score is -2.01.

$$X_i = Z_i(S) + \overline{X}$$

$$\rightarrow X_i = (-2.01)(5.41) + 44.81 = 33.94.$$

* Chebyshev's theorem: skip.

* **Empirical Rule:** « based on the normal distribution »

the empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

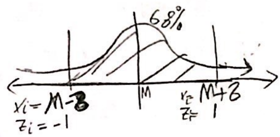
normal
. Bell-shaped خاص فقط في ←

→ **Empirical rule:**

For data having a bell-shaped distribution:-

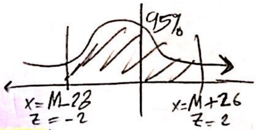
① Approximately 68% of the data values will be within one standard deviation of the mean.

أي أنه 68% من البيانات تقع بين $X_i = M+\sigma$, $X_i = M-\sigma$.

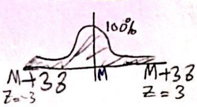


② Approximately 95% of the data values will be within 2 standard deviations of the mean.

95% من البيانات تقع بين $X_i = M+2\sigma$, $X_i = M-2\sigma$



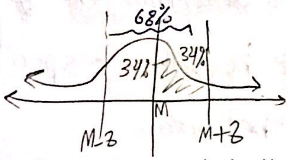
③ Almost all of the data values will be within 3 standard deviations of the mean.



أي أنه معظم البيانات تقع بين $M+3\sigma$, $M-3\sigma$
وأي قيمة تقع خارجهما تعتبر outliers

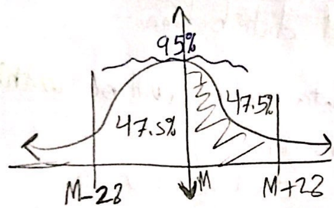
-Notes: since the bell shaped is symmetric, we have.

①

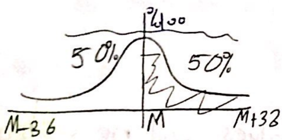


أي بين M و M+δ يقع ما نسبته 34% من البيانات، وما نسبته
34% من البيانات يقع بين M و M-δ.

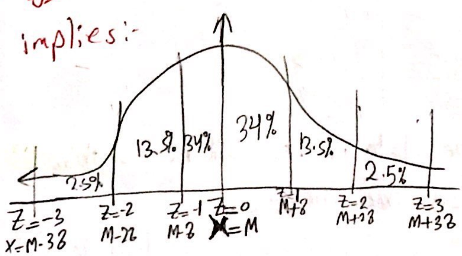
②



③



. this implies:-



- Ex: Assume the ages of employees in BZU have a bell-shaped distribution with mean = 44, and standard deviation = 8.

a) What is the percentag of employees with ages between 28 and 68 years?

① نرسم Bell-shaped مستعينا في Empirical rule.

$\mu = 44, \ \sigma = 8$

$\rightarrow Z = -1 \ , \ X = \mu - \sigma = 36$

$Z = 1 \ , \ X = \mu + \sigma = 52$

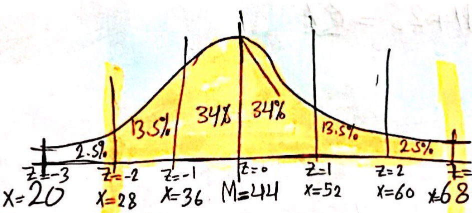
$\rightarrow Z = -2 \ , \ X = 28 \} \ \mu - 2\sigma$

$Z = 2 \ , \ X = 60 \} \ \mu + 2\sigma$

$\rightarrow Z = -3 \ , \ X = 20 \} \ \mu - 3\sigma$

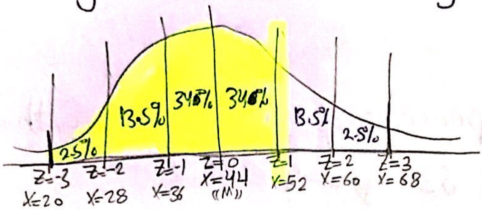
$Z = 3 \ , \ X = 68 \} \ \mu + 3\sigma$

$\rightarrow Z = 0 \ , \ X = \mu = 44$ (at the center).



$\therefore$ percentage $= (13.5 + 34 + 34 + 13.5 + 2.5)\%$
$= 97.5\%$

b) What is the percentage of ~~students in B~~ employees with ages less than 52 years?



∴ the percentage is 84%

-Ex: If the weights of students in BZU have a bell shaped distribution with $M = 70$ Kg and $\delta = 8.5$.
What is the percentage of students with weights between 61.5 Kg and 87 Kg? $M = 70$, $\delta = 8.5$

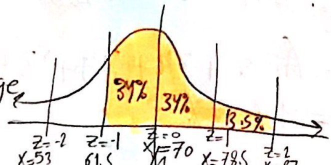
$\rightarrow Z = -1$, $X = M - \delta = 70 - 8.5 = 61.5$
$Z = 1$, $X = M + \delta = 70 + 8.5 = 78.5$

$\rightarrow Z = -2$, $X = M - 2\delta = 53$
$Z = 2$, $X = M + 2\delta = 87$

$\rightarrow Z = -3$
$Z = 3$

$\rightarrow Z = 0$

∴ the percentage = 81.5%



* Detecting outliers:
القيم المتطرفة

Sometimes a data set will have one or more observations with unusually large or small values. These are extreme values or outliers.

→ If the data has a bell-shaped distribution, we can use the empirical rule to detect outliers:-

- any value is greater than $\mu+3\sigma$ or less than $\mu-3\sigma$ is an outlier.

→ that is $(X > \mu+3\sigma \; , \; X < \mu-3\sigma)$ outliers.

$\quad Z > 3 \; , \; Z < -3$

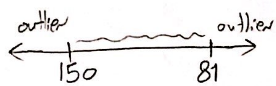
- Ex: Given a sample:

$310, 500, 470, 1920, 450, 380$, and

assume this sample is taken from a bell-shaped distribution with mean $\mu = 480$ and $\sigma = 110$. Use the empirical rule to find the outliers, if any.

$x_* = \mu - 3\sigma = 480 - 3(110) = 150$

$x^* = \mu + 3\sigma = 480 + 3(110) = 810$



∴ $X = 1920$ is an outlier.

- Sec 3.4 : Explatory data analysis :

• **Explatory data analysis :-**

we use simple arithmetic to draw pictures to summarize data.

→ by considering 5 number summaries and box plots.

• the following five numbers summary are used to summarize the data :-

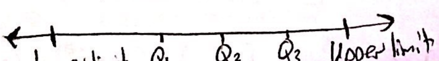
1) smallest value
2) First quartile « Q_1 ».
3) Median « Q_2 »
4) Third quartile « Q_3 »
5) largest value.

then we place the data in ascending order. تصاعدي (from smallest to largest)

• Box plot:-

is a graphical summary of data that is based on a five-number summary

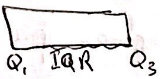
- The steps to construct the box plot:-

1) 

wehere lower limit $LL = Q_1 - 1.5IQR$

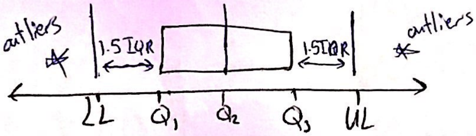
Upper limit $UL = Q_3 + 1.5\,IQR$; $IQR = Q_3 - Q_1$

2) a box is drawn with the ends located at Q_1, and Q_3.



3) A vertical line is drawn at the location of the median « Q_2 »

4) dat outside the limits are considerd outliers. « * »



-Ex: Construct a box plot for:

12, 10, 18, 13, 25, 18

→ $Q_1 = P_{25}$ 10, 12, 13, 18, 18, 25

$i = \frac{25}{100} \times 6 = 1.5 \rightarrow \boxed{Q_1 = 12}$

→ $\boxed{\therefore Q_2 = \frac{13+18}{2} = 15.5}$ median أو P_{50}

→ $Q_3 = P_{75}$

$i = \frac{75}{100} \times 6 = 4.5 \rightarrow \boxed{\therefore Q_3 = 18}$

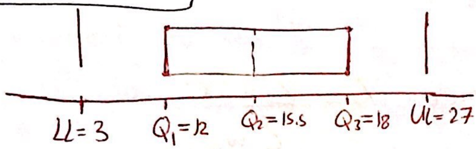
→ $LL = Q_1 - 1.5\,IQR$; $IQR = Q_3 - Q = 6$

$\therefore \boxed{LL = 12 - 1.5(6) = 3}$

$\rightarrow UL = Q_3 + 1.5\, IQR$

$= 18 + 1.5(6)$

$\therefore \boxed{UL = 27}$



there is no outlier.

- Ex: Consider the sample: 110, 150, 210, 180, 210, 70, 400. Use the box plot to find outliers.

$\rightarrow$ 70, 110, 150, 180, 210, 210, 400

$Q_1 = P_{25} \rightarrow i = \frac{25}{100} \times 7 = 1.75 \rightsquigarrow \underline{\underline{2}} \quad \therefore \boxed{Q_1 = 110}$

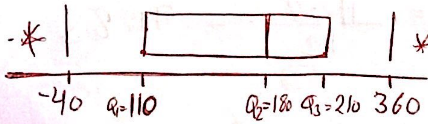
$Q_2 = P_{50} \rightarrow i = \frac{50}{100} \times 7 = 3.5 \rightsquigarrow 4 \quad \therefore \boxed{Q_2 = 180}$

$Q_3 = P_{75} \rightarrow i = \frac{75}{100} \times 7 = 5.25 \rightsquigarrow 6 \quad \therefore \boxed{Q_3 = 210}$

$IQR = Q_3 - Q_1 = 210 - 110 = 100$

$LL = Q_1 - 1.5\, IQR = 110 - 1.5(100) = -40.$

$UL = Q_3 + 1.5\, IQR = 210 + 1.5(100) = 360$



so 400 is an outlier because $400 > UL$.

- Sec 3.6: The weighted mean and working with grouped data:-

① **Weighted mean:**

→ The weighted mean $\bar{X} = \dfrac{\sum w_i x_i}{\sum w_i}$

where x_i = value of observation i
w_i = weight for observation i.

if we have population, the weighted mean is μ.

- Ex: The grades of 4 courses were:-

	«W»	«X_i»
English	2 hours	95
Arabic	3 hours	80
Calculus 2	3 hours	92
Stat 2311	4 hours	99

Find the mean of grades.

→ $\bar{X} = \dfrac{\sum x_i w_i}{\sum w_i} = \dfrac{95 \times 2 + 80 \times 3 + 92 \times 3 + 99 \times 4}{2+3+3+4}$

$= \dfrac{1102}{12} = 91.83$

② Grouped data:-

data are available only in a grouped or frequency distribution. We show how the weighted mean formula can be used to obtain approximation of the mean, Variance and standard deviation for grouped data.

→ Sample mean for grouped data:-

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

; M_i = the midpoint for class i.
f_i = the frequency for class i
n = the sample size.

→ Sample variance for grouped data:-

$$S^2 = \frac{\sum f_i (M_i - \bar{X})^2}{n-1}$$

→ population variance for grouped data:-

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

- Ex: Consider the following data:-

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

1) Find the sample mean.

$\frac{\text{lower limit} + \text{upper limit}}{2}$ = نجد midpoints لكل فئة

class	f	midpoint M	Mf
10-14	4	$\frac{10+14}{2} = 12$	$4 \times 12 = 48$
15-19	8	17	$8 \times 17 = 136$
20-24	5	22	$5 \times 22 = 110$
25-29	2	27	$2 \times 27 = 54$
30-34	1	32	$1 \times 32 = 32$

$$\therefore \bar{X} = \frac{\sum Mf}{\sum f} = \frac{48 + 136 + 110 + 54 + 32}{4 + 8 + 5 + 2 + 1} = \frac{380}{20} = 19$$

2) Find the sample variance. $\overline{X} = 19$ « from part 1) »

class	f (midpoint)	M	$M - \overline{X}$ (نطرح من كل mid الوسط $\overline{X}$)	$(M - \overline{X})^2$ (نربع العمود الرابع)	$f(M - \overline{X})^2$ (نضرب العمود الخامس في العمود الثاني)
10-14	4	12	$12 - 19 = -7$	49	196
15-19	8	17	$17 - 19 = -2$	4	32
20-24	5	22	$22 - 19 = 3$	9	45
25-29	2	27	$27 - 19 = 8$	64	128
30-34	1	32	$32 - 19 = 13$	169	169

نجمع العمود السادس

$$\therefore S^2 = \frac{\sum f(M - \overline{X})^2}{\sum f - 1} = \frac{196 + 32 + 45 + 128 + 169}{(4 + 8 + 5 + 2 + 1) - 1}$$

مجموع العمود الثاني

$$\longrightarrow S^2 = \frac{570}{19} = \underline{30}$$

3) Find the sample standard deviation. ناخذ فقط الجذر

$$\longrightarrow S = \sqrt{30} = \underline{5.48}$$

4) If these data are from population, find the population variance. نجدها كما اوجدنا sample، ولكن نقسم $\sum f$.

رمز الوسط pop

$$\longrightarrow \sigma^2 = \frac{\sum f(M - \mu)^2}{\sum f} = \frac{570}{20} = \underline{28.5}$$

5) find the population standard deviation.

$$\longrightarrow \sigma = \sqrt{28.5} = \underline{5.34}$$

* Using SD mode. «calculator f_x - 82 MS».

1) To find the weighted mean.

x_i	w_i
x_1	w_1
x_2	w_2
⋮	⋮
x_n	w_n

① $\boxed{\text{Mode}}$ $\boxed{2}$

② x_1 $\boxed{M+}$ x_2 $\boxed{M+}$ ...... x_n $\boxed{M+}$. ندخل القيم وبين كل قيمة وأخرى نضغط M+

③ $\boxed{\triangledown}$ $\boxed{\triangledown}$ w_1 $\boxed{=}$ $\boxed{\triangledown}$ $\boxed{\triangledown}$ w_2 $\boxed{=}$ $\boxed{\triangledown}$ $\boxed{\triangledown}$ w_3 $\boxed{=}$ ----

$\boxed{\triangledown}$ $\boxed{\triangledown}$ w_n $\boxed{=}$

④ $\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{1}$ $=$

2) To find the mean of grouped data.

classes	f

$$M_i = \text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

① We find the midpoints $(M_1, M_2, \dots, M_n)$

② $\boxed{\text{mode}}$ $\boxed{2}$

③ M_1 $\boxed{M+}$ M_2 $\boxed{M+}$ .... $\boxed{M_n}$ $M+$ ندخل midpoint وبين كل قيمة وأخرى نضغط M+

④ $\boxed{\triangledown}$ · $\boxed{\triangledown}$ f_1 $\boxed{=}$ $\boxed{\triangledown}$ $\boxed{\triangledown}$ f_2 $\boxed{=}$ ---- $\boxed{\triangledown}$ $\boxed{\triangledown}$ f_n $\boxed{=}$

⑤ $\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{1}$ =

3) To find the sample standard deviation of grouped data.

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{3}$ =

4) To find the ~~sample~~ population standard deviation of grouped data.

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{2}$ =

5) To find the sample variance of grouped data.
we just square s .

6) To find the population variance of grouped data.
we just square σ .

- Ex: Solve the previous example using calculator.

class	f	Midpoints
10-14	4	12
15-19	8	17
20-24	5	22
25-29	2	27
30-34	1	32

(1) Find the sample mean.

(1) mode 2

(2) 12 M+ 17 M+ 22 M+ 27 M+ 32 M+

(3) ▽ ▽ 4 = ▽ ▽ 8 = ▽ ▽ 5 = ▽ ▽ 2 = ▽ ▽ 1 =

(4) shift 2 1 = 19

(2) Find the sample standard deviation.

shift 2 3 = 5.48

(3) Find the population standard deviation.

shift 2 2 = 5.34

(4) Find the sample variance, population variance.

$S^2 = 30$, $\sigma^2 = 28.5$

"square part (2)" "square part (3)"

CamScanner المسوحة ضوئيا بـ

- Sec **3.5** : Measures of Association Between two variables

العلاقة بين متغيرين

أخذنا قراءات لتلخيص بيانات بمتغير واحد ، وهنا سنوضح بعض القراءات التي تربط بين متغيرين.

① Covariance :

التغاير بين متغيرين

أو التباين بين متغيرين

For a sample of size n with the observations:-

(x_1, y_1) , (x_2, y_2) , ------ , (x_n, y_n)

→ the sample covariance "S_{xy}" $= \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$;

$\bar{x}$: sample mean for variable X.

$\bar{y}$: sample mean for variable Y.

→ the population covariance $\sigma_{xy} = \dfrac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$;

μ_x : population mean for X.

μ_y : population mean for Y.

N : population size

- Ex: The following table summarizes the number of absences (x) and the grade (y) of a sample of students.

x_i	6	3	1	4	5
y_i	65	81	94	76	69

• Find the sample covariance.

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

→ ① $\bar{x} = \frac{\sum x_i}{n} = \frac{6+3+1+4+5}{5} = 3.8$

② $\bar{y} = \frac{\sum y_i}{n} = \frac{65+81+94+76+69}{5} = 77$

③

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
6	65	$6 - 3.8 = 2.2$	$65 - 77 = -12$	$(2.2)(-12) = -26.4$
3	81	-0.8	4	-3.2
1	94	-2.8	17	-47.6
4	76	0.2	-1	-0.2
5	69	1.2	-8	-9.6

④ ∴ $S_{xy} = \frac{-26.4 + -3.2 + -47.6 + 0.2 + -9.6}{5-1} = -21.75$

ملاحظة: لو كانت من population نقسم على N.

- Interpretation of the covariance:-

The covariance is a measure of the linear association between 2 variables. يقيس نوع العلاقة الخطية بين متغيرين

→ ① if $S_{xy} > 0$ «positive», indicates a positive linear relation between X and Y. «if X increases y increases»

② if $S_{xy} < 0$ «-ve», indicates a negative linear relation between X and Y. «if X increases, y decreases»

③ if $S_{xy} = 0$, indicates no linear relation between X and Y.

② Correlation Coefficient: معامل الارتباط

measures the type and strength of the relation between X and Y.

→ the sample correlation coefficient «r_{xy}» $= \frac{S_{xy}}{S_x S_y}$;

S_{xy}: sample covariance

S_x: sample standard deviation of X.

S_y: sample standard deviation of Y.

→ the population correlation coefficient «ρ_{xy}» $= \frac{\sigma_{xy}}{\sigma_x \sigma_y}$;

σ_{xy}: population covariance.

σ_x: population standard deviation of X.

σ_y: population standard deviation of Y.

→ The sample correlation coefficient r_{xy} is the point estimator of the population correlation coefficient. « r_{xy} is a point estimator ~~of~~ ρ_{xy} »

• Interpretation of the correlation coefficient:-

① $-1 \leq r_{xy} \leq 1$

② if ~~xy~~ $r_{xy} = 1$, then there is a perfect ~~linear~~ positive linear relationship between X and Y.

③ if $r_{xy} \cong 1$, a strong positive linear relationship.

④ if $r_{xy} = -1$, a perfect negative linear relationship.

⑤ if $r_{xy} \cong -1$, a strong negative linear relationship.

⑥ if $r_{xy} = 0$, there is no linear relationship.

⑦ if $r_{xy} \cong 0$, there is a weak linear relationship.

~~Ex: Given the~~

-Ex: Consider the following sample data.

x_i	y_i
5	10
10	30
15	50

find the ~~ex~~ sample correlation coefficient.

① $\bar{X} = \frac{\sum x_i}{n} = \frac{5+10+15}{3} = 10$

$\bar{y} = \frac{\sum y_i}{n} = \frac{10+30+50}{3} = 30$

② S_x , S_y

x_i	y_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
5	10	-5	25	-20	400
10	30	0	0	0	0
15	50	5	25	20	400

$\therefore S_x^2 = \frac{\sum (x_i - \bar{X})^2}{n-1} = \frac{25+0+25}{3-1} = \frac{50}{2} = 25$

$\longrightarrow \boxed{S_x = \sqrt{25} = 5}$

$\therefore S_y^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{400+0+400}{3-1} = \frac{800}{2} = 400$

$\longrightarrow \boxed{S_y = \sqrt{400} = 20}$

③ S_{xy}.

$(x_i - \bar{x})(y_i - \bar{y})$
$(-5)(-20) = 100$
0
100

$\therefore S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{200}{2} = 100$

④ $r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{100}{(5)(20)} = \underline{1}$ $\therefore$ there is a perfect linear relationship

* Using calculator «SD mode» to find r_{xy}, S_{xy}.

(x_1, y_1) (x_2, y_2) . . . — — (x_n, y_n).

① $\boxed{\text{mode}}$ $\boxed{3}$ $\boxed{1}$

② x_1 $\boxed{,}$ y_1 $\boxed{M+}$

x_2 $\boxed{,}$ y_2 $\boxed{M+}$

⋮

x_n $\boxed{,}$ y_n $\boxed{M+}$

③ To find the sample mean of X. → $\bar{X}$

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{1}$ $\boxed{=}$

لو كان pop. نفس الخطوات
لـ M_x

④ To find the sample mean of Y. → $\bar{Y}$

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{\triangleright}$ $\boxed{1}$ $\boxed{=}$

«لو كان pop نفس الخطوات
M_y

⑤ To find the sample standard deviation of X → S_x

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{3}$ $\boxed{=}$

«لو كان pop. σ_x بدل $\boxed{3}$ نضغط $\boxed{2}$ أي
«$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{2}$»

⑥ To find the sample standard deviation of Y. → S_y

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{\triangleright}$ $\boxed{3}$ $\boxed{=}$

لو كان pop : σ_y

$\boxed{\text{shift}}$ $\boxed{2}$ $\boxed{\triangleright}$ $\boxed{2}$ $=$

⑦ To find the ~~ss~~ sample correlation coefficient → r_{xy}.

shift | 2 | ▷ | ▷ | 3 | =

((لو كان pop. (ρ_{xy}) نفس الخطوات.))

⑧ To find the sample covariance → S_{xy}.

((لو كان pop. (σ_{xy})
→ $\sigma_{xy} = \rho_{xy}\,\sigma_x\,\sigma_y$

$$S_{xy} = r_{xy}\,S_x\,S_y.$$