

# ΣΤΑΤΙΣΤΙΚΗ

email: stamatiou.dimitrios@gmail.com

email ucy: dimopoulos.stamatiou@ucy.ac.cy

## Αξιολόγηση

Ενδιάμεση εξέταση 40% (10/11)

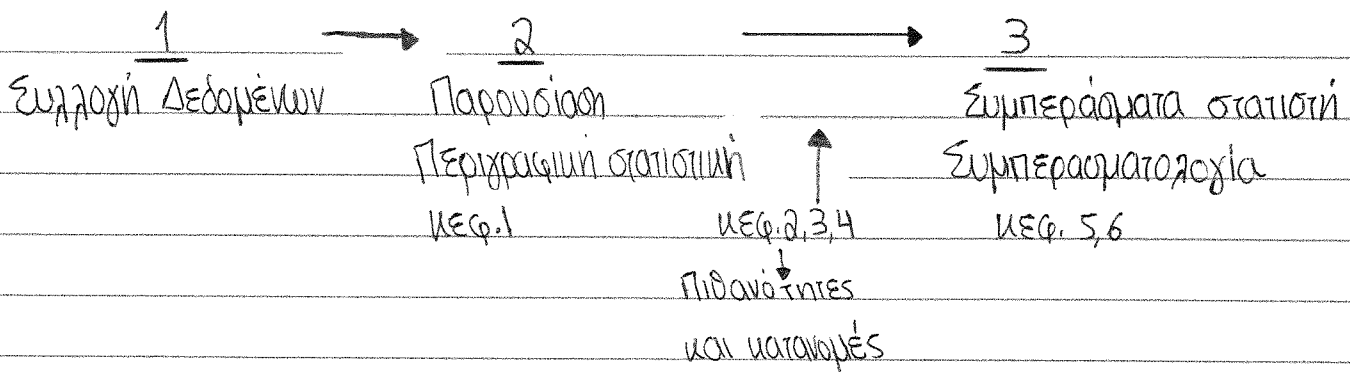
Τελική εξέταση 60%

Φροντιστήριο: Τετάρτη: 9-10

Στατιστική: Μελέτη μέρους ενός συνόλου → συμπεράσματα για όλο το σύνολο

π.χ πρόθεση γάλακτος

έλεγχος: οι πτήσεις τον χειμώνα έχουν περισσότερες καθυστερήσεις απ'ότι το καλοκαίρι



## ΚΕΦΑΛΑΙΟ 1 - Περιγραφική Στατιστική

Μελέτη συνόλου ως προς κάποιο χαρακτηριστικό

Πληθυσμός: είναι το σύνολο των μονάδων, η μελέτη των οποίων έχει ζητηθεί για την εξαγωγή συμπερασμάτων.

(πολύ μεγάλο μέγεθος) • Ολόκληρο το σύνολο που θέλουμε να μελετήσουμε

(2)

αντιπροσω-  
πευστικό

Δείγμα: Είναι το σύνολο των μονάδων που προέρχονται από τον πληθυσμό και χρησιμοποιούνται για τη συλλογή δεδομένων και την εφαρμογή συμπερασμάτων (αντικείμενο της μελέτης)  
υπόσύνολο πληθυσμού • το υπόσύνολο που επιλέξαμε τυχαία

### Μεταβλητές

\* Ποιοτικές (δεν μπορούν να πάρουν αριθμητικές τιμές)

π.χ χρώμα ματιών, επάγγελμα, ομάδα αίματος, φύλο

\* Ποσοτικές (π.χ ύψος, βαθμολογία σε ένα μάθημα) ο,τιδήποτε μετρείται

• Διακριτές (ακέραιοι αριθμοί) (συμμετρικές τιμές)

π.χ αριθμός φοιτητών κ.κ., αριθμός παιδιών, ηλικία βόλων

• Συνεχείς (τιμές σε διάστημα)

π.χ ύψος, βάρος, χρόνος, απόσταση οικίας από το ΠΚ

### Οργάνωση δεδομένων

#### Πίνακας συχνοτήτων

π.χ Ρωτήσαμε 10 φοιτητές από το ΠΚ πόσα μαθήματα χρωστάνε

• πληθυσμός: φοιτητές του ΠΚ

• μεταβλητή: αριθμός χρωστούμενων μαθημάτων

φοιτητές : 1 2 3 4 5 6 7 8 9 10

απαντήσεις : 1 0 2 2 1 3 6 2 3 3

• Μέγεθος δείγματος:  $v=10$

• Τιμές μεταβλητής:  $x_i$  ( $x_1=0, x_2=1, x_3=2, x_4=3, x_5=6$ )

#### Συχνότητα μέσης τιμής: $v_i$

Ο αριθμός εμφανίσεων της τιμής  $x_i$  στο δείγμα

$v_1=1, v_2=2, v_3=3, v_4=3, v_5=1$

3

Σχετική συχνότητα = Συχνότητα  
πληθυσμός αριθμός δείγματος

$$F_i = \frac{v_i}{v} \quad \text{π.χ. } F_1 = \frac{1}{10} = 0,1, \quad F_2 = \frac{2}{10} = 0,2$$

Πίνακας συχνοτήτων 1<sup>η</sup> στήλη: αύξοντα ρυθμό (i)  
2<sup>η</sup> στήλη: τιμή (x<sub>i</sub>)  
3<sup>η</sup> στήλη: συχνότητα (v<sub>i</sub>)  
4<sup>η</sup> στήλη: σχετική συχνότητα (f<sub>i</sub>)

i	Τιμές (x <sub>i</sub> )	Συχνότητα (v <sub>i</sub> )	Σχετική συχνότητα (F <sub>i</sub> )
1	0	1	0,1
2	1	2	0,2
3	2	3	0,3
4	3	3	0,3
5	6	1	0,1
Άθροισμα		<u>10</u>	<u>1</u>

Πρέπει  $\sum_{i=1}^5 v_i = v$  και  $\sum_{i=1}^5 F_i = 1$

Γραφικές μέθοδοι παρουσίασης

Α) Ποιοτικές μεταβλητές

π.χ ρωτήσαμε 10 άτομα το επάγγελμα του πατέρα τους

i	x <sub>i</sub>	v <sub>i</sub>
1	εργάτης	5
2	δημόσιος υπαλλ.	2
3	ιερέας	3
Άθροισμα		10

A1- Ραβδόγραμμα

κάθε τιμή  $x_i$  αντιστοιχεί σε μια ράβδο.

Υψος ράβδου: συχνότητα της αντιστοιχίας τιμής  
(τα ορθογώνια απέχουν μεταξύ τους)

\*  
συχνότητα ή  
αριθμ. συχν.



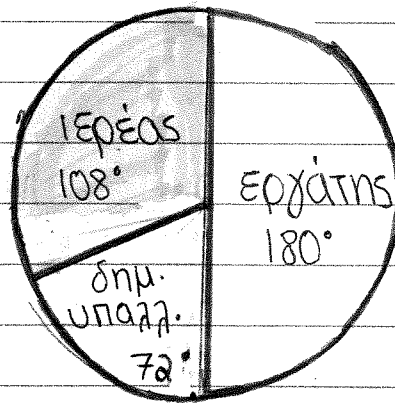
A2- κυκλικό διάγραμμα (pie chart)

κάθε τιμή  $x_i$  αντιστοιχεί σε τόσο κώνο του  $Q_i = \frac{v_i}{v} \cdot 360^\circ = F_i \cdot 360^\circ$

•  $Q_1 = \frac{5}{10} \cdot 360^\circ = 180^\circ$

•  $Q_2 = \frac{2}{10} \cdot 360^\circ = 72^\circ$

•  $Q_3 = \frac{3}{10} \cdot 360^\circ = 108^\circ$



$\frac{5}{10} \cdot 100 = 50\%$  Σχετική Συχν.

20%

30%

B) Ποσοτικές μεταβλητές

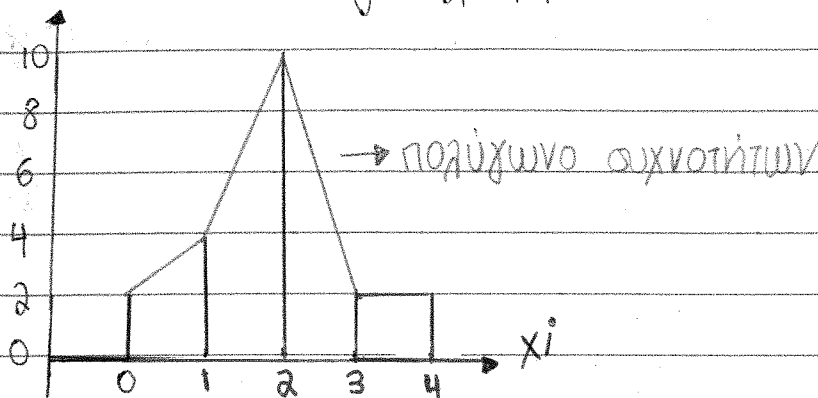
B- Διάγραμμα συχνοτήτων ή γραμμικό διάγραμμα

π.χ. Ρωτήσαμε 20 οικογένειες για τον αριθμό των ανήλικων μελών

\*  
συχν ή  
ΣΣ

i	$x_i$	$v_i$
1	0	2
2	1	4
3	2	10
4	3	2
5	11	2

Γραμμικό διάγραμμα: Κάθε τιμή  $x_i \Rightarrow$  κατακόρυφη γραμμή  
 ύψος γραμμής  $\Rightarrow$  συχνότητα της αντιστοίχης τιμής



Πολύγωνο συχνοτήτων: Πολυγωνική γραμμή που συνδέει τις κορυφές των γραμμών

B2 - Κυκλικό Διάγραμμα

Αριθμός όπως και στο A2 (διαμοιρές)

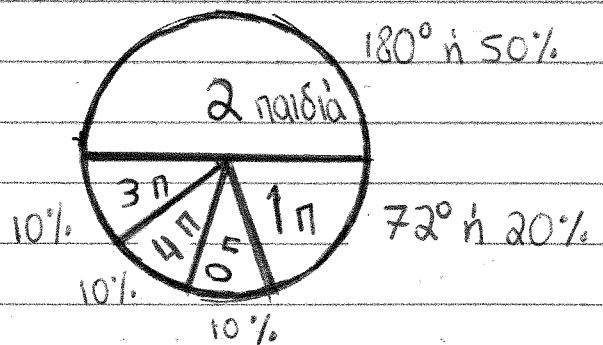
$Q_1 = \frac{2}{20} \cdot 360^\circ = 36^\circ$  10%

$Q_2 = \frac{4}{20} \cdot 360^\circ = 72^\circ$  20%

$Q_3 = \frac{10}{20} \cdot 360^\circ = 180^\circ$  50%

$Q_4 = \frac{2}{20} \cdot 360^\circ = 36^\circ$  10%

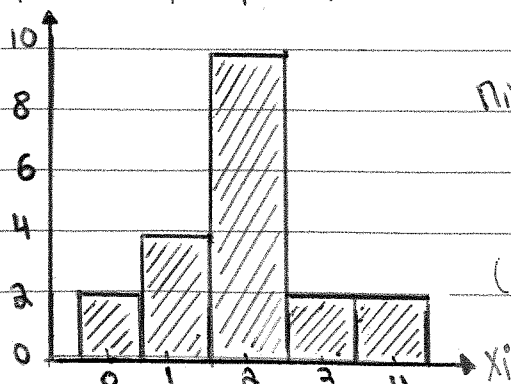
$Q_5 = \frac{2}{20} \cdot 360^\circ = 36^\circ$  10%



B3 - Ιστόγραμμα (για ομαδοποιημένες παρατηρήσεις)

\* Συχνότητα ή σχετική συχνότητα. Κάθε τιμή αντιπροσωπεύεται από 1 ορθογώνιο με εμβαδόν ίσο με τη συχνότητα ή σχετική συχνότητα. (Διαδοχικά ορθογώνια.)

• Το ιστογράμμο είναι χρήσιμο όταν έχουμε πολλές τιμές  $x_i$  ή όταν έχουμε συνεχή μεταβλητή.



Πηχ το ίδιο παράδειγμα

(με βάση = 1) όλα

6

79

ΣΕ ΑΥΤΕΣ ΤΙΣ ΠΕΡΙΠΤΩΣΕΙΣ ΚΑΝΟΥΜΕ ΟΜΑΔΟΠΟΙΗΣΗ ΤΩΝ ΠΑΡΑΤΗΡΗΣΕΩΝ  
 Π.χ. Ρυθμίσαμε 20 φοιτητές για τη βαθμολογία τους σε ένα μάθημα  
 79, 63, 80, 67, 97, 83, 78, 71, 84, 77, 76, 64, 82, 66, 99, 81, 79, 72, 83, 75

- 1) Αριθμός κλάσεων
- 2) Μέγεθος ή πλάτος κλάσεων
- 3) Πίνακας συχνοτήτων κλάσεων

πλάτος κλάσεων: 10 (ομαδοποιούμε)

μικρότερη τιμή: 63

μεγαλύτερη τιμή: 99

$$[a, b] = \frac{a+b}{2}$$

i	κλάση	κέντρο κλάσης	$V_i$ (Συχνότητα)	Σχ. Συχνότητα
1	[60, 70)	65	4	0,2
2	[70, 80)	75	8	0,4
3	[80, 90)	85	6	0,3
4	[90, 100]	95	2	0,1
Άθροισμα			20	1

• Ομαδοποιούμε σε κλάσεις

1) Κάθε κλάση είναι ένα διάστημα  $[a, b]$

2) Πλάτος κλάσης: το μήκος του διαστήματος (δηλ.  $b - a$ )

3) Κέντρο κλάσης: η μεσοία τιμή του διαστήματος

• Κάθε κλάση αντιστοιχεί σε ένα ορθογώνιο

• Η βάση του θα είναι η κλάση (θα έχουν το ίδιο πλάτος)

• Για το ύψος θα πρέπει να ισχύει:  $\text{βάση} \cdot \text{ύψος} = \text{συχνότητα}$

$E = \text{συχνότητα}$

(7)

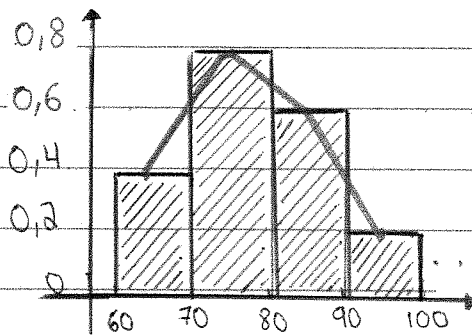
• [60,70] → ορθογώνιο με βάση 10  
 πλάτος υψους · ύψος = συχνότητα  
 $10 \cdot \text{ύψος} = 4 \Rightarrow \text{ύψος} = 0,4$

• [70,80] → ορθογ. με βάση 10  $\Rightarrow 10 \cdot \text{ύψος} = 8 \Rightarrow \text{ύψος} = 0,8$

• [80,90] → ορθ. με βάση 10  $\Rightarrow 10 \cdot \text{ύψος} = 6 \Rightarrow \text{ύψος} = 0,6$

• [90,100] → ορθ. με βάση 10  $\Rightarrow 10 \cdot \text{ύψος} = 2 \Rightarrow \text{ύψος} = 0,2$

Ιστόγραμμα  
Συχνοτήτων



• Για ιστόγραμμα σχετιωών συχνοτήτων το εμβαδόν κάθε ορθογωνίου είναι ίσο με την αντίστοιχη σχετιωή συχνότητα

• Το πολύγωνο συχνοτήτων είναι η πολυγωνική γραμμή που συνδέει τα μέσα των άκρων βάσεων.

B4) Δευδρόγραμμα ή Φυλλογράφημα

Για δεδομένα με δύο ή περισσότερα και είναι πίνακας με:

- 1) Πρώτη στήλη τα πρώτα ψηφία των τιμών (εξτός του τελευταίου) - κορμιά
- 2) Δεύτερη στήλη τα τελευταία ψηφία σε αύξουσα σειρά

π.χ (ίδια δεδομένα)

Δευάδες	Μονάδες
6	3,4,6,7
7	1,2,5,6,7,8,9,9
8	0,1,2,3,3,4
9	7,9

(Βρίσκω την ελάχιστη + μέγιστη τιμή)

(8)

## Μέτρα ή δείκτες θέσης

1) Δειγματικός μέσος (ή μέση τιμή) mean,  $\bar{x}$

Αν έχουμε δείγμα μεγέθους "n"  $\Rightarrow \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$

ή  $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n v_i x_i$ , αν έχουμε πίνακα συχνοτήτων  
( $v_i$  η συχνότητα της τιμής  $x_i$ )

π.χ Ρωτήσαμε 15 φοιτητές πόσα σουζαρίκια φερούν  
0, 1, 1, 3, 5, 2, 2, 2, 0, 1, 2, 4, 2, 3, 2

$$\bar{x} = \frac{0+1+1+3+5+2+2+2+0+1+2+4+2+3+2}{15} = 2$$

$$\bar{x} = \frac{2 \cdot 0 + 3 \cdot 1 + 6 \cdot 2 + 2 \cdot 3 + 1 \cdot 4 + 1 \cdot 5}{15} = 2$$

2) Διάμεσος (median), m

Αν βάλουμε τις τιμές σε αύξουσα σειρά, η μεσαία παρατήρηση (δηλ. αυτή που έχει 50% μικρότερες και 50% μεγαλύτερες) είναι η διάμεσος.

Αν υπάρχουν 2 μεσαίες, η διάμεσος είναι η μέση τιμή τους (ημιάθροισμα)

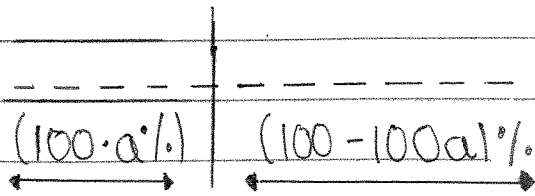
π.χ 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5. Αν π.χ ήταν 14:  $\frac{7n+8n}{2}$

3) 100·α<sup>ο</sup>-ποσοστημόριο (percentile) ( $0 \leq \alpha \leq 1$ )

(σε αύξουσα σειρά) είναι η τιμή που έχει 100·α<sup>ο</sup>% των τιμών αριστερά (μικρότερες) και το (100-100·α<sup>ο</sup>%) δεξιά (μεγαλύτερες)



9



- Υπολογίζουμε  $a \cdot n$
- Αν δεν είναι αμέσως επόμενος αμέσως αντίστροφα πίνει τον ζητούμενο δείκτη
- Αν είναι αμέσως τότε το ποσοστμόριο είναι το  $\frac{X_{a_n} + X_{a_{n+1}}}{2}$
- Για  $a = 0,25$  (25° ποσοστμόριο) συμβολίζεται με  $Q_1$
- Για  $a = 0,5$  (50° ποσοστμόριο) συμβολίζεται με  $Q_2$
- Για  $a = 0,75$  (75° ποσοστμόριο) συμβολίζεται με  $Q_3$

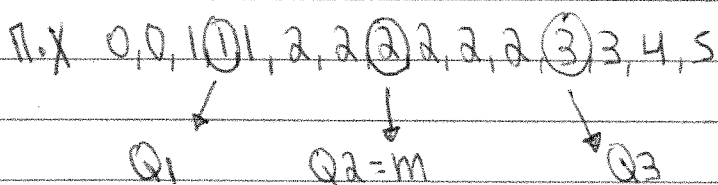
\* Για το  $Q_1$ :  $a = 0,25$

$$Q \cdot n = 15 \cdot 0,25 = 3,75 \rightarrow 4 \text{ (αμέσως επόμενος αμέσως)}$$

\* Για το  $Q_2$ :  $a = 0,5$  ( $Q_2 = m$ )  $\rightarrow 8$

\* Για το  $Q_3$ :  $a = 0,75$

$$Q \cdot n = 15 \cdot 0,75 = 11,25 \rightarrow 12$$



Για το 40° ποσοστμόριο,  $a = 0,40$

$$a \cdot n = 0,40 \cdot 15 = 6 \text{ άρα είναι το } \frac{X_6 + X_7}{2} = \frac{2 + 2}{2} = 2$$

4) Επικρατούσα τιμή (mode),  $E$

Είναι η τιμή με τη μεγαλύτερη συχνότητα π.χ στα ίδια δεδομένα  $E = 2$  (που βρίσκεται η κορυφή)

(10)

5)  $k\%$  - ξαυρισμένος μέσος (trimmed mean)

$k/2\%$	Μέση τιμή	$k/2\%$
---------	-----------	---------

• Είναι ο μέσος των παρατηρήσεων που μένουν αν αφαιρέσουν οι  $k/2\%$  μικρότερες και οι  $k/2\%$  μεγαλύτερες παρατηρήσεις

π.χ. ιδία δεδομένα

40/2

Για το 40% ξαυρισμένο μέσο, 20% του 15 = 3

=> Αφαιρούμε τις 3 μικρότερες και τις 3 μεγαλύτερες τιμές

$$= \frac{1+1+2+2+2+2+2+2+3}{9} = \frac{17}{9}$$

Τα μέτρα ή δείκτες διασποράς ή μεταβλητότητας

1) Δειγματική διασπορά (sample variance)

$$S^2 = \frac{1}{n-1} (\sum x_i^2 - n \cdot \bar{x}^2) \quad \text{ή} \quad S^2 = \frac{1}{n-1} (\sum n_i x_i^2 - n \cdot \bar{x}^2) \quad \text{αν έχουμε πίνακα συχνοτήτων}$$

π.χ. ίδιες παρατηρήσεις:  $S^2 = \frac{2 \cdot 0^2 + 3 \cdot 1^2 + 6 \cdot 2^2 + 2 \cdot 3^2 + 1 \cdot 4^2 + 1 \cdot 5^2 - 15 \cdot 2^2}{14} = \frac{13}{7}$

2) Δειγματική τυπική απόκλιση (standard deviation)

$$S = \sqrt{S^2} \quad \text{π.χ.} \quad S = \sqrt{13/7}$$

Μέτρα θέσης: μέσος  $\bar{x}$

διαμεσός

ποσοστημόριο

επιπαρατούσα τιμή

κ% φαυλοπλενός μέσος

$Q_1 = 10$  τεταρτημόριο

$Q_2 = 20$

$Q_3 = 30$

Μέτρα μεταβλητότητας 1) Δειγμ. διασπορά 2) τυπιή απόκλιση

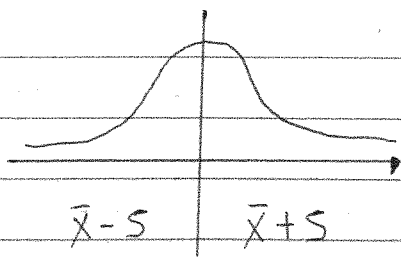
3) Μέση τυπιή απόκλιση (MAD - mean absolute deviation)

$$MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \quad \text{ή} \quad \frac{\sum_{i=1}^n n_i |x_i - \bar{x}|}{n}$$

πληθος

π.χ 0,0,1,1,1,2,2,2,2,2,2,3,3,4,5  $n=15, \bar{x}=2$

$$MAD = \frac{2 \cdot |0-2| + 3 \cdot |1-2| + 6 \cdot |2-2| + 2 \cdot |3-2| + |4-2| + |5-2|}{15} = \frac{14}{15}$$



$$\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad \frac{|x_i - \bar{x}|}{n}$$

4) Δειγματικός εύρος (sample range)

Μεγαλύτερη Παρατήρηση - Μικρότερη Παρατήρηση

π.χ Εύρος =  $5 - 0 = 5$

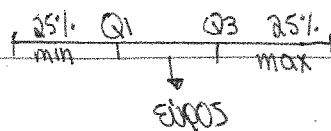
5) Διατετακτισμένο εύρος

$Q_3 - Q_1$

$Q_1: 10$  τεταρτημόριο  $\alpha = 0,25$

$Q_3: 30$  τεταρτημόριο  $\alpha = 0,75$

π.χ  $Q_3 - Q_1 = 3 - 1 = 2$



(12)

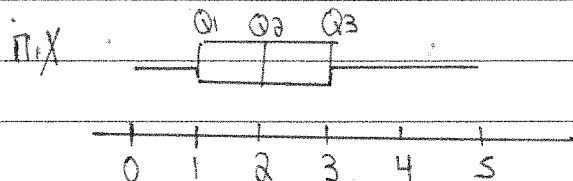
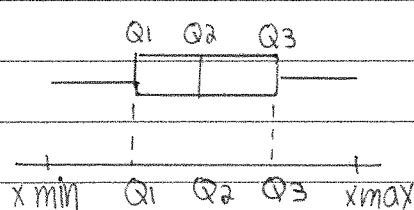
Εξυτρεφής Μεταβλητότητας (CV = coefficient of variation)

$$CV = \frac{s}{\bar{x}} \quad \text{π.χ.} \quad CV = \frac{\sqrt{\frac{13}{7}}}{2} = 0,68$$

Θηλογράμμα (Boxplot)

(συνή με μυστοάμα)

Διογράμμα που περιλαμβάνει  $Q_2$  = διομέος,  $Q_1$ ,  $Q_3$ , εύρος



Άσκηση

Δίνεται η βαθμολογία 50 φοιτητών σε ένα διοώνισμα μαθηματικών

55, 75, 89, 98, 71, 93, 81, 73, 72, 78, 74, 59, 81, 92, 51, 84, 97, 86, 95, 66, 61,  
72, 65, 70, 80, 57, 65, 53, 64, 65, 69, 76, 78, 74, 95, 85, 80, 73, 85, 75,  
65, 84, 86, 76, 61, 69, 90, 77, 72, 78

α) Να ομαδοποιηθούν σε κλάσεις πλάτους 10 και να κατασκευαστεί ο πίνακας συχνοτήτων

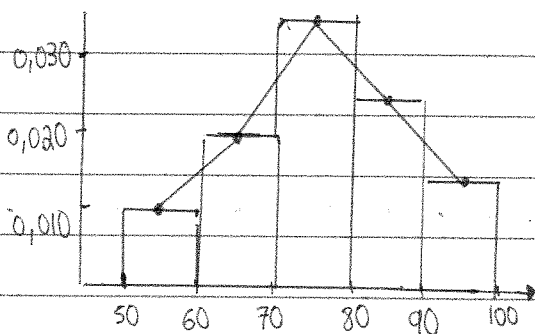
β) Να κατασκευαστεί το ιστόγραμμα σχετιών συχνοτήτων και το αντίστοιχο παύλεο συχνοτήτων

γ) Να υπολογιστούν η μέση τιμή και η τυπική απόκλιση των βαθμολογιών

α) κλάσεις:  $[50, 60)$ ,  $[60, 70)$ ,  $[70, 80)$ ,  $[80, 90)$ ,  $[90, 100)$

Κλάσεις	Κέντρο μάζας	Συχνότητα	Σχετική Συχνότητα	Σχ. Συχνότητα Πλάτος Κλάσης	Κέντρο μάζας x Συχνότητα	(Κέντρο) <sup>2</sup> / (κλάσης)	*
[50,60)	55	5	5/50 = 0,1	0,1/10 = 0,01	275	3025	15125
[60,70)	65	10	10/50 = 0,2	0,2/10 = 0,02	650	4225	42250
[70,80)	75	17	17/50 = 0,34	0,34/10 = 0,034	1275	5625	95625
[80,90)	85	11	11/50 = 0,22	0,22/10 = 0,022	935	7225	79475
[90,100]	95	7	7/50 = 0,14	0,14/10 = 0,014	665	9025	63175
Αθροισμα		50	1				

Βασ. ύψος = σχετική Συχνότητα ⇒ ύψος = σχετική συχνότητα  
Βάση



8) Για μέση τιμή και διασπορά  $10,7/50$  τα κέντρα των κλάσεων με τις αντίστοιχες συχνότητες.

$$\bar{x} = \frac{275 + 650 + 1275 + 935 + 665}{50} = \frac{3800}{50} = 76$$

$$S^2 = \frac{1}{n-1} (\sum n_i x_i^2 - n \cdot \bar{x}^2) = \frac{1}{49} (15125 + 42250 + 42250 + 95625 + 79475 + 63175 - 50 \cdot 76^2)$$

$$= 139,8$$

$$S = \sqrt{S^2} = \sqrt{139,8} \approx 11,85$$