

Εισαγωγή στη θεωρία Πιθανοτήτων και την Στατιστική

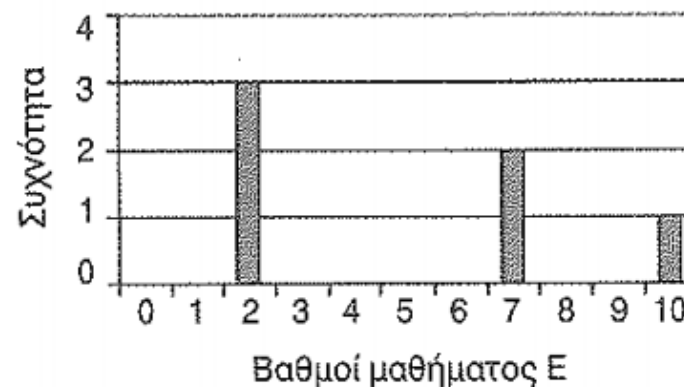
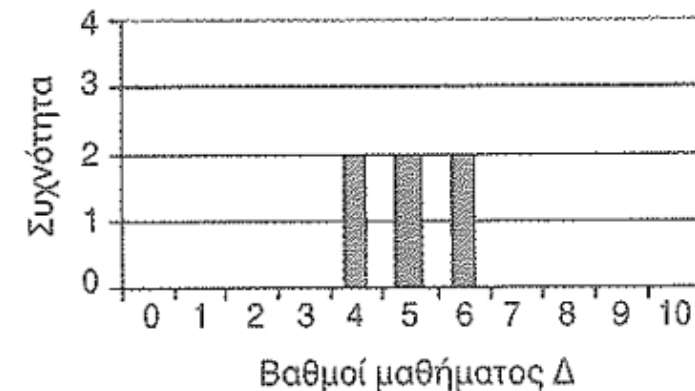
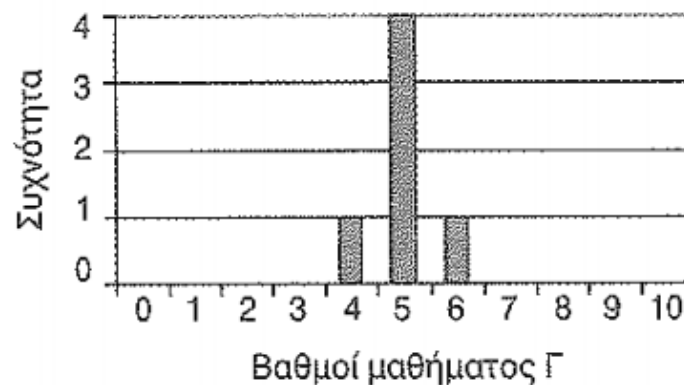
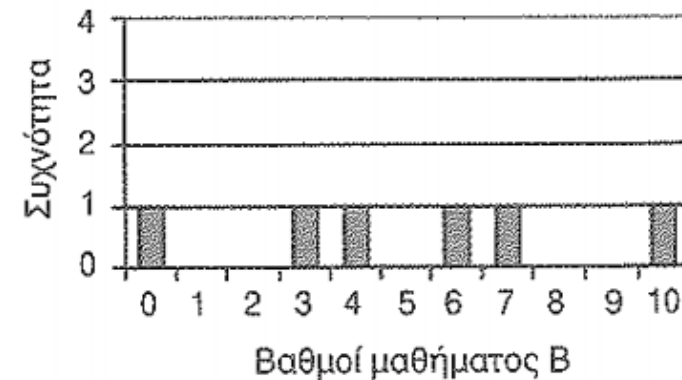
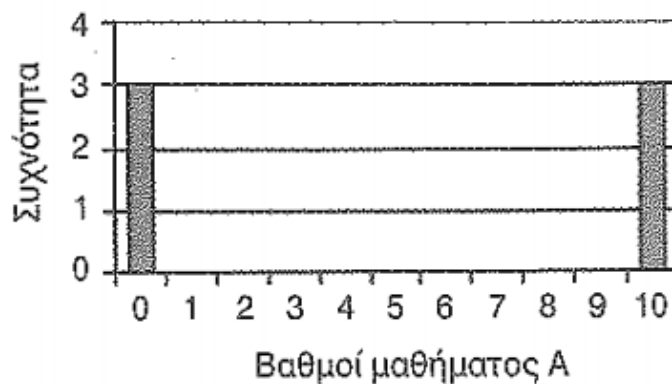
Μέτρα διασποράς

Μέτρα διασποράς

- Προσδιορισμός της διασποράς των παρατηρήσεων, δηλ. την έκταση στην οποία εκτείνονται γύρω από το κέντρο
- **Εύρος**
- **Ενδοτεταρτημοριακό εύρος**
- **Διακύμανση**
- **Τυπική απόκλιση**
- **Συντελεστής μεταβλητότητας**

Μέτρα διασποράς

Κατανομές με μέσο
όρο 5 αλλά με
διαφορετική
διασπορά



Εύρος - R

- Είναι η **διαφορά της ελάχιστης (min) από τη μεγίστη (max) μέτρηση** και συμβολίζεται με το γράμμα ***R***, δηλαδή

$$R = x_{\max} - x_{\min}$$

- Αποτελεί την απόσταση που χωρίζει τις δύο ακριανές μετρήσεις.
- Το εύρος είναι αρκετά απλό μέτρο και υπολογίζεται εύκολα, όμως δε θεωρείται αξιόπιστο μέτρο διασποράς, διότι βασίζεται μόνο στις δύο ακραίες παρατηρήσεις.

Πλεονεκτήματα

- Είναι πολύ απλό στον υπολογισμό .
- Μπορεί να χρησιμοποιηθεί για την εκτίμηση της τυπικής απόκλισης.

Μειονεκτήματα

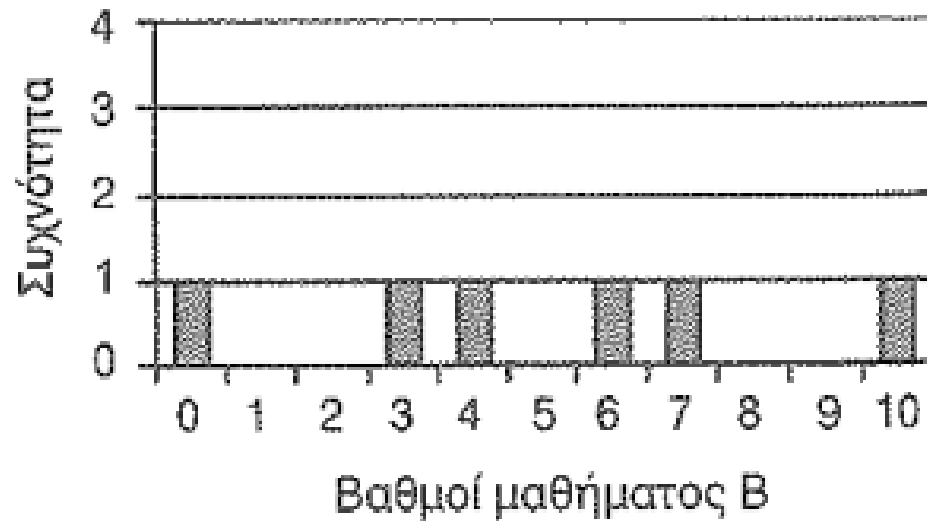
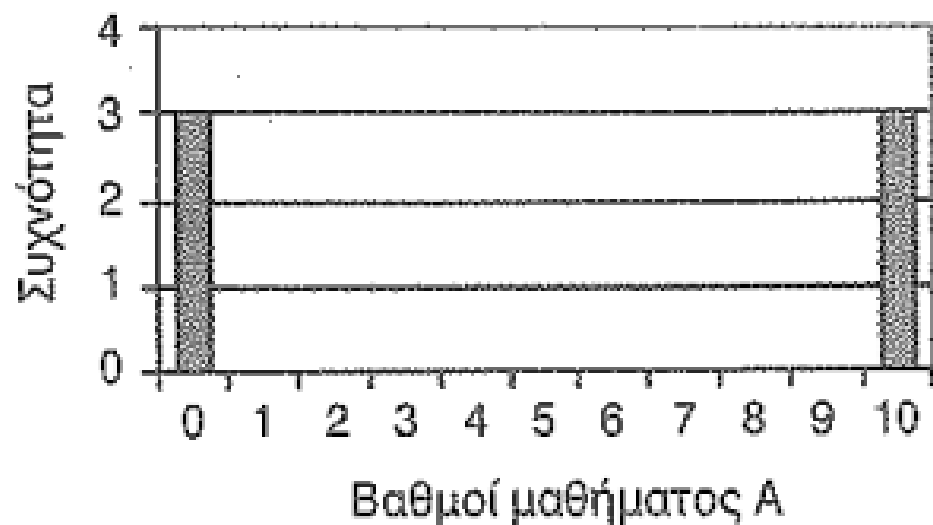
- Δε θεωρείται αξιόπιστο μέτρο διασποράς, επειδή βασίζεται μόνο στη μικρότερη και τη μεγαλύτερη παρατήρηση και συνεπώς είναι ευαίσθητο σε ακραίες τιμές.
- Δε χρησιμοποιείται για περαιτέρω στατιστική ανάλυση

Εφαρμογή 1

Να υπολογίσετε το εύρος του παρακάτω δείγματος:

3, 7, - 2, 5, 9, 13, 4

Ενδοτεταρτημοριακό εύρος - Q



Ίδιο εύρος, διαφορετική διασπορά

Ενδοτεταρτημοριακό εύρος - Q

- Το **ενδοτεταρτημοριακό εύρος** συμβολίζεται με Q και είναι η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3 .

$$Q = Q_3 - Q_1$$

- Όσες παρατηρήσεις βρίσκονται έξω από το διάστημα $[Q_1 - 1,5Q, Q_3 + 1,5Q]$ ονομάζονται **ακραίες τιμές**.

Ενδοτεταρτημοριακό εύρος - Q

- Το ενδοτεταρτημοριακό εύρος, μας δίνει το πλάτος του κεντρικού (γύρω από τη διάμεσο) διαστήματος εντός του οποίου βρίσκεται το 50% των τιμών του δείγματος, γι' αυτό θεωρείται η «καρδιά» της κατανομής.
- Όσο μικρότερο είναι αυτό το διάστημα τόσο μικρότερη εν γένει είναι η μεταβλητότητα των τιμών του δείγματος.
- Το ενδοτεταρτημοριακό εύρος είναι ο δείκτης διασποράς που συνήθως αναφέρεται, όταν η διάμεσος χρησιμοποιείται ως δείκτης κεντρικής τάσης.
- Το ενδοτεταρτημοριακό εύρος δεν επηρεάζεται από ακραίες τιμές.

Εφαρμογή 2

Οι βαθμοί 20 μαθητών σ' ένα διαγώνισμα Μαθηματικών είναι:

18, 40, 47, 53, 80, 8, 38, 42, 46, 51
79, 6, 32, 39, 48, 83, 24, 36, 44, 64

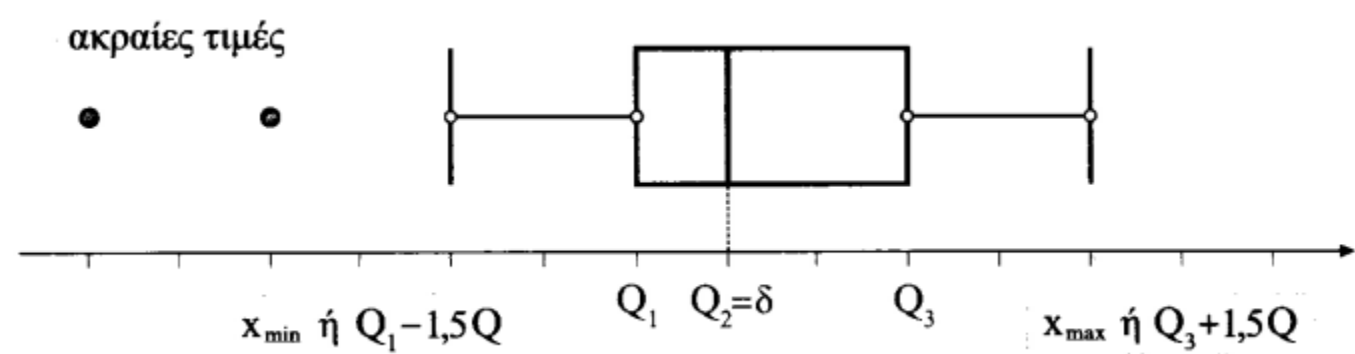
- α) Να βρείτε το ενδοτεταρτημοριακό εύρος των βαθμών.
- β) Να εξετάσετε αν υπάρχουν ακραίες τιμές στο παραπάνω δείγμα.

Εφαρμογή 2

Οι βαθμοί 20 μαθητών σ' ένα διαγώνισμα Μαθηματικών είναι:

18,	40,	47,	53,	80,	8,	38,	42,	46,	51
79,	6,	32,	39,	48,	83,	24,	36,	44,	64

Θηκόγραμμα



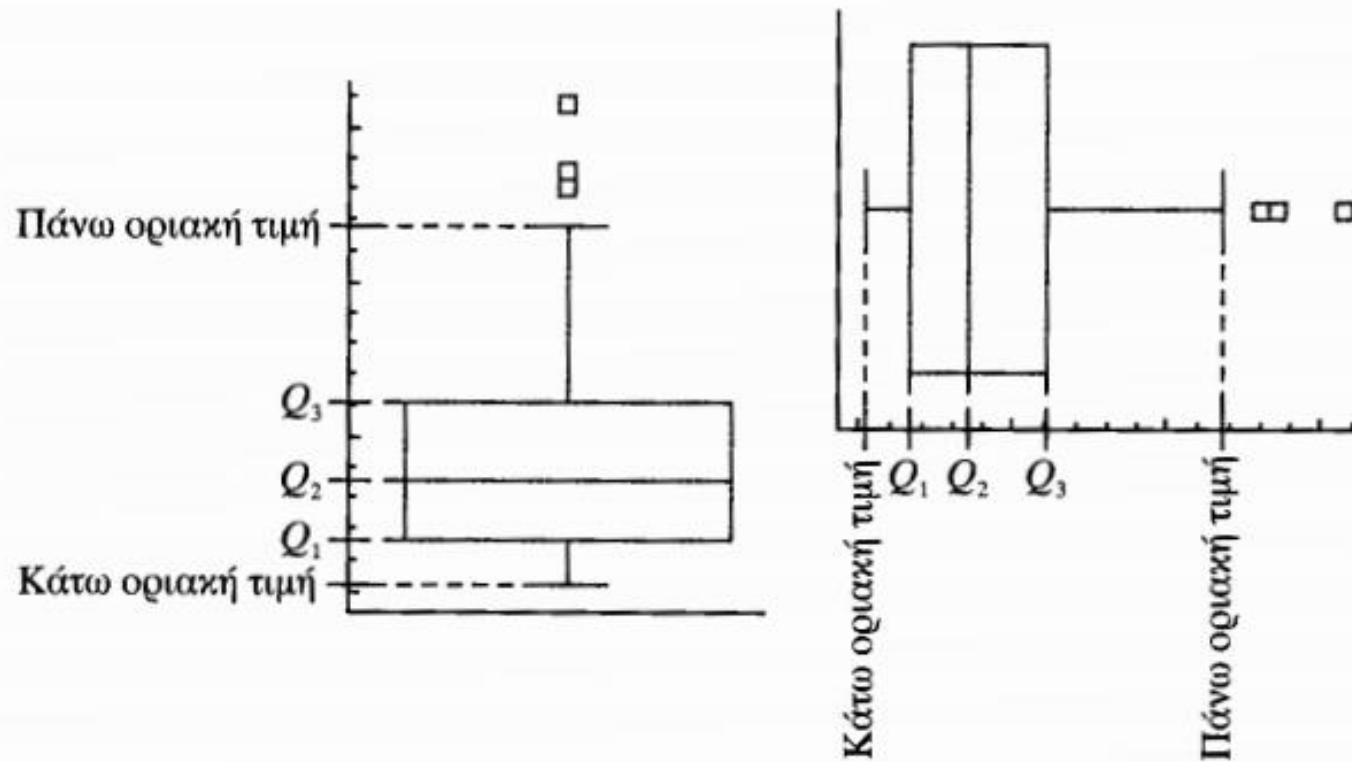
Χρήσιμο σχεδιάγραμμα, το **box-plot** (ή αλλιώς **Θηκόγραμμα**), που είναι βασισμένο σε πέντε στατιστικά στοιχεία θέσεως:

min (ή κάτω ακραία τιμή), **Q_1 ,** **Q_2 ,** **Q_3 ,** **max (ή άνω ακραία τιμή)**

- Συνοψίζει απλά και γρήγορα τα δεδομένα και δίνει, χωρίς πολύπλοκους ή χρονοβόρους υπολογισμούς, μια εύληπτη εικόνα για την κεντρική τάση και τη διασπορά της κατανομής
- Δίνει ενδείξεις για τη συμμετρία της κατανομής ανάλογα με τη θέση του «κουτιού» και τις αποστάσεις μεταξύ της διαμέσου και των άλλων δύο τεταρτημόριων

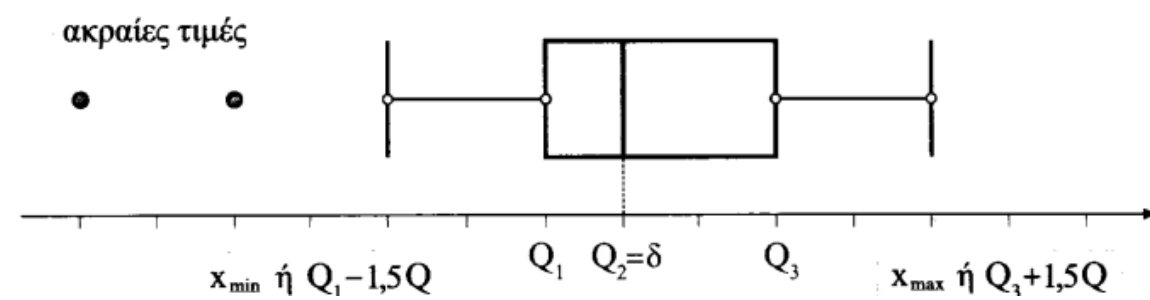
Θηκόγραμμα

Μπορεί να σχεδιαστεί οριζόντια ή κατακόρυφα



Σχήματα 9.1.17.
Θηκόγραμμα κατανομής.

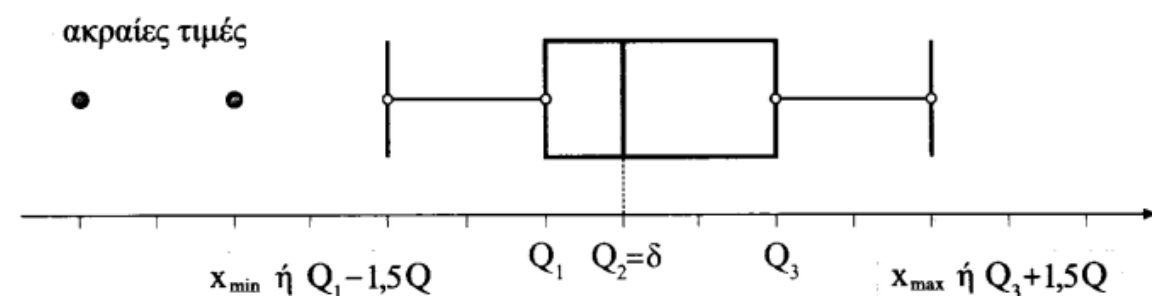
Θηκόγραμμα - Κατασκευή



1. Πάνω από έναν άξονα αριθμών κατασκευάζουμε ένα ορθογώνιο, του οποίου η βάση είναι παράλληλη στον άξονα. Η αριστερή πλευρά του ορθογωνίου βρίσκεται στη θέση που αντιστοιχεί στον αριθμό Q_1 και η δεξιά πλευρά του ορθογωνίου βρίσκεται στη θέση που αντιστοιχεί στον αριθμό Q_3 (το μήκος αυτών των πλευρών επιλέγεται αυθαίρετα).
 2. Μέσα στο ορθογώνιο σχεδιάζουμε ένα ευθύγραμμο τμήμα παράλληλο στην αριστερή και στη δεξιά πλευρά, το οποίο βρίσκεται στη θέση που αντιστοιχεί στον αριθμό Q_2 , δηλαδή στη διάμεσο δ .
 3. Από τα μέσα της αριστερής και της δεξιάς πλευράς σχεδιάζουμε ευθύγραμμα τμήματα, έξω από το ορθογώνιο, παράλληλα στον άξονα των αριθμών, τα οποία έχουν μήκος το πολύ ίσο με $1,5Q$.
- Συγκεκριμένα:

- Το ευθύγραμμο τμήμα που ξεκινά από το μέσο της αριστερής πλευράς φτάνει είτε μέχρι την ελάχιστη παρατήρηση ($x_{\min}$) είτε μέχρι τον αριθμό $Q_1 - 1,5Q$, αν η $x_{\min}$ είναι μικρότερη από τον αριθμό αυτόν.
- Το ευθύγραμμο τμήμα που ξεκινά από το μέσο της δεξιάς πλευράς φτάνει είτε μέχρι τη μέγιστη παρατήρηση ($x_{\max}$) είτε μέχρι τον αριθμό $Q_3 + 1,5Q$, αν η $x_{\max}$ είναι μεγαλύτερη από αυτόν τον αριθμό.

Θηκόγραμμα - Κατασκευή

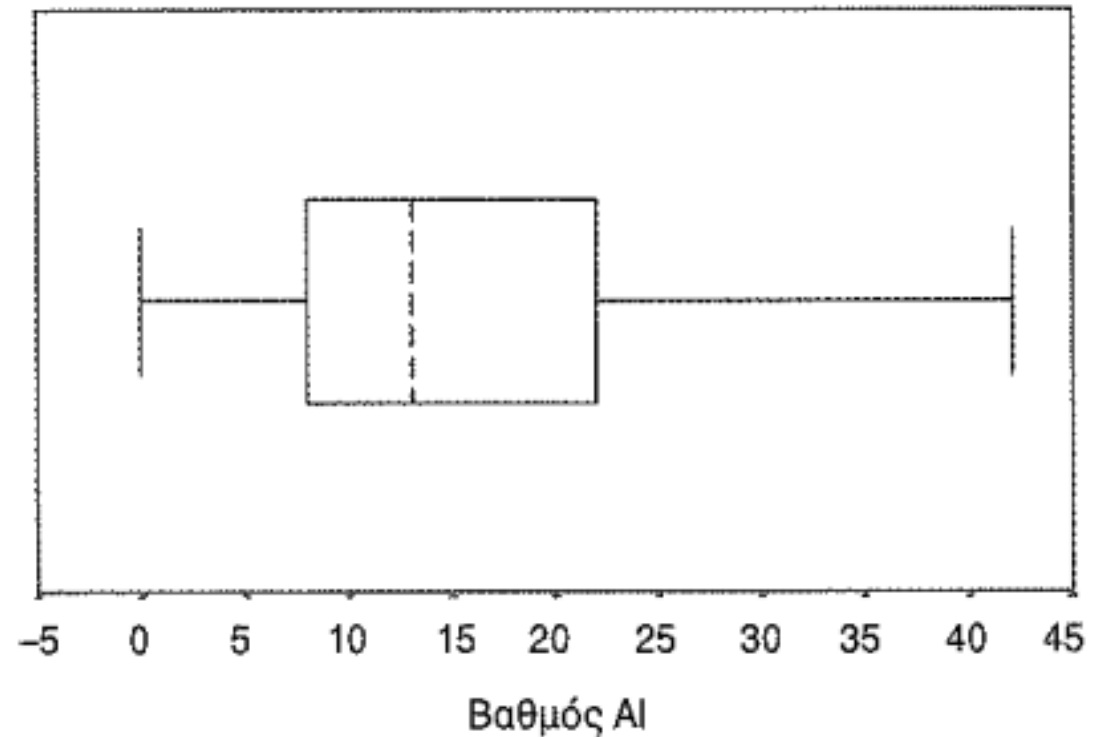


4. Στο τέλος των παραπάνω ευθύγραμμων τμημάτων φέρουμε κάθετα ευθύγραμμα τμήματα.
5. Αν υπάρχουν παρατηρήσεις έξω από το διάστημα $[Q_1 - 1,5Q, Q_3 + 1,5Q]$, τότε αυτές παριστάνονται με τελείες στις αντίστοιχες θέσεις.

Θηκόγραμμα

- Δίνει ενδείξεις για τη συμμετρία της κατανομής

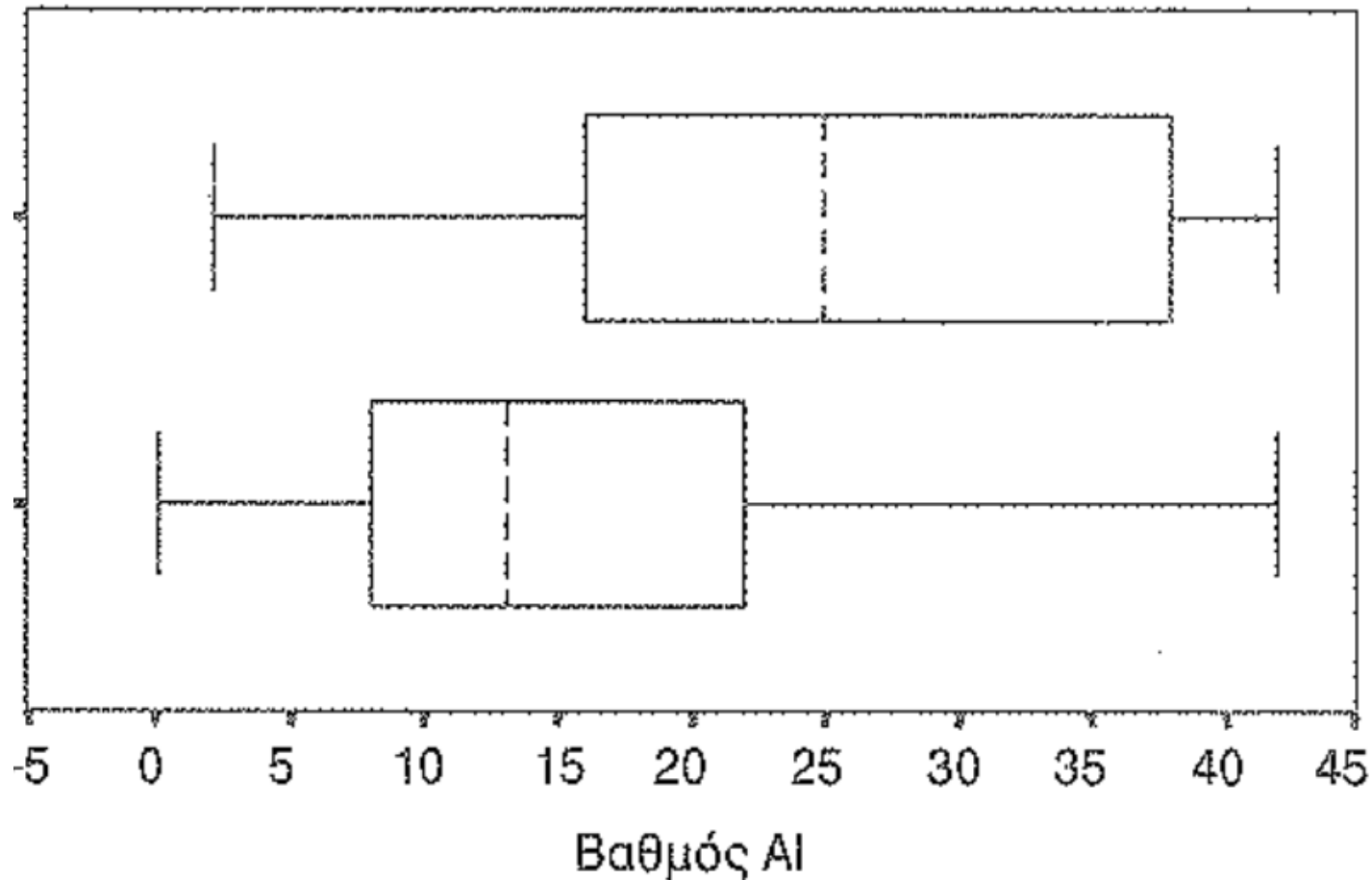
Η έλλειψη συμμετρίας είναι φανερή. Το μεσαίο μισό βρίσκεται σαφώς προς το αριστερό άκρο της κλίμακας (δηλαδή τα περισσότερα παιδιά με ΜΔ έλαβαν χαμηλό βαθμό) και η διάμεσος είναι πλησιέστερα στο πρώτο τεταρτημόριο.



Θηκόγραμμα

- Βοηθά στη σύγκριση της κατανομής 2 ή περισσότερων δειγμάτων

Στα παιδιά με ΜΔ το «κουτί» είναι μετατοπισμένο προς τους χαμηλούς βαθμούς της κλίμακας, ενώ, αντίθετα, το μεσαίο μισό των μαθητών χωρίς ΜΔ βρίσκεται προς τους υψηλούς βαθμούς της κλίμακας.



Εφαρμογή 2α

Οι βαθμοί 20 μαθητών σ' ένα διαγώνισμα Μαθηματικών (με άριστα το 100) είναι:

18, 40, 47, 53, 80, 8, 38, 42, 46, 51
79, 6, 32, 39, 48, 83, 24, 36, 44, 64

α) Να κατασκευάσετε το θηκόγραμμα

Εφαρμογή 3

Οι βαθμοί των παιδιών μιας τάξης σε ένα διαγώνισμα είναι:

8 6 15 17 12 6 5 20 10 7 20 6

Να υπολογίσετε:

α) το εύρος,

β) την επικρατούσα τιμή,

γ) τον μέσο όρο

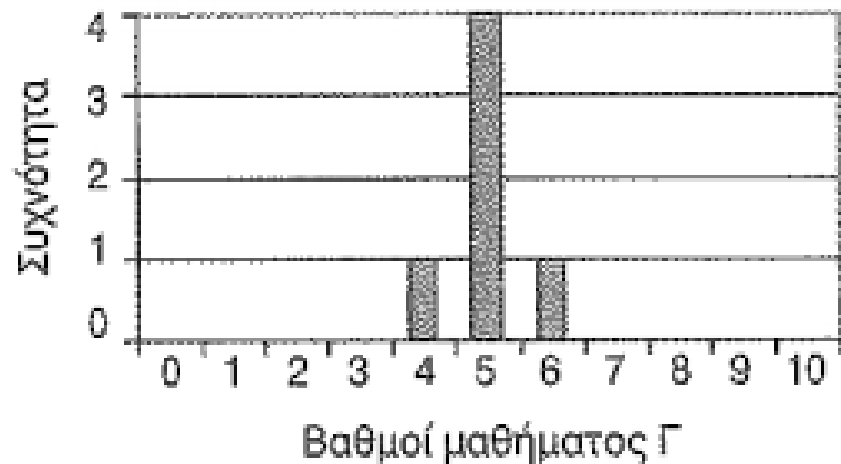
δ) τα τεταρτημόρια Q_1 , Q_2 , Q_3

ε) το ενδοτεταρτημοριακό εύρος

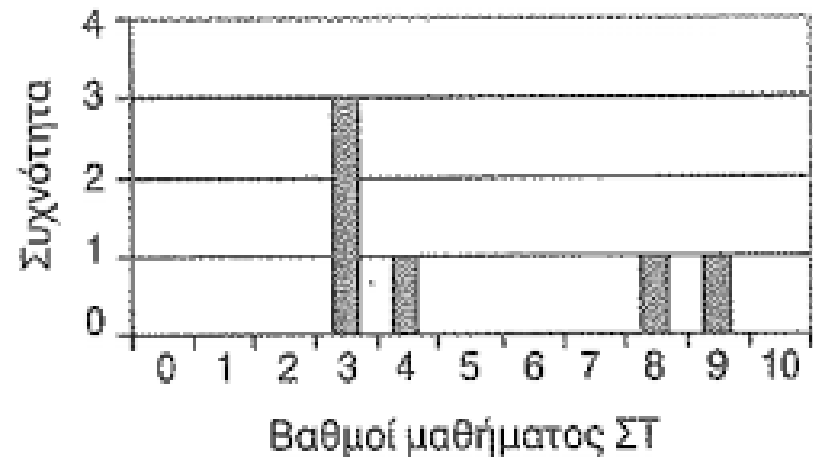
στ) να κατασκευάσετε το θηκόγραμμα των βαθμών

Διακύμανση – s^2

Παρατηρούνται
διαφορετικές
αποστάσεις των
βαθμών από τον
μέσο όρο, δηλ.
διαφορετικές
αποκλίσεις



A/A	Μάθημα Γ	Απόκλιση = $X - \bar{X}$
1	4	$4 - 5 = -1$
2	5	$5 - 5 = 0$
3	5	$5 - 5 = 0$
4	5	$5 - 5 = 0$
5	5	$5 - 5 = 0$
6	6	$6 - 5 = 1$
$\bar{X} = 5$		$\sum (X - \bar{X}) = 0$



A/A	Μάθημα ΣΤ	Απόκλιση = $X - \bar{X}$
1	3	$3 - 5 = -2$
2	3	$3 - 5 = -2$
3	3	$3 - 5 = -2$
4	4	$4 - 5 = -1$
5	8	$8 - 5 = 3$
6	9	$9 - 5 = 4$
$\bar{X} = 5$		$\sum (X - \bar{X}) = 0$

Διακύμανση – s^2

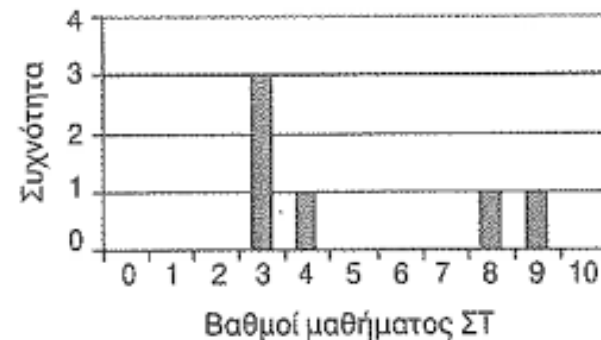
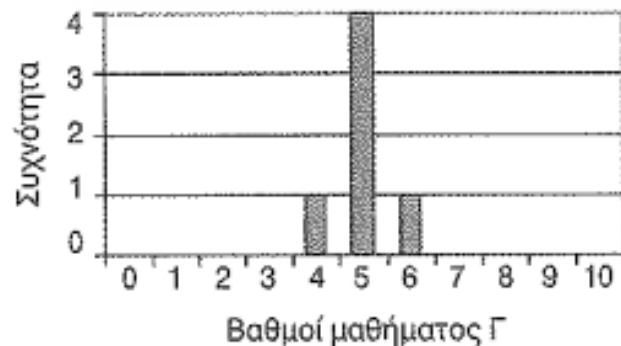
- Το άθροισμα των αποκλίσεων των παρατηρήσεων από τον μέσο όρο είναι πάντα μηδέν.

$$\sum (x - \bar{X}) = 0$$

- Εφόσον ο μέσος όρος, αποτελεί το κέντρο βάρους της κατανομής, είναι επόμενο οι αποστάσεις κάποιων σημείων από τον μέσο όρο να είναι αρνητικές (όταν οι μετρήσεις είναι μικρότερες του μέσου όρου) και κάποιων άλλων θετικές (όταν οι μετρήσεις είναι μεγαλύτερες του μέσου όρου). Οι θετικές αποκλίσεις πρέπει να εξουδετερώνουν τις αρνητικές αποκλίσεις, αφού ο μέσος όρος είναι το σημείο ισορροπίας. Αυτό συνεπάγεται το άθροισμα των αποκλίσεων να ισούται με το μηδέν.

Διακύμανση – s^2

- 1^η λύση ο υπολογισμός της απόλυτης απόκλισης



A/A	Μάθημα Γ	Απόλυτη απόκλιση = $ X - \bar{X} $
1	4	$ 4 - 5 = 1$
2	5	$ 5 - 5 = 0$
3	5	$ 5 - 5 = 0$
4	5	$ 5 - 5 = 0$
5	5	$ 5 - 5 = 0$
6	6	$ 6 - 5 = 1$
$\bar{X} = 5$		$\sum X - \bar{X} = 2$

A/A	Μάθημα ΣΤ	Απόλυτη απόκλιση = $ X - \bar{X} $
1	3	$ 3 - 5 = 2$
2	3	$ 3 - 5 = 2$
3	3	$ 3 - 5 = 2$
4	4	$ 4 - 5 = 1$
5	8	$ 8 - 5 = 3$
6	9	$ 9 - 5 = 4$
$\bar{X} = 5$		$\sum X - \bar{X} = 14$

Δεν έχει καλές
στατιστικές
ιδιότητες

Διακύμανση – s^2

Αν υπολογίσουμε, αντί για τον μέσο όρο των απόλυτων αποκλίσεων, τον μέσο όρο των αποκλίσεων υψωμένων στο τετράγωνο, έχουμε τη **μέση τετραγωνική απόκλιση** των μετρήσεων.

Η μέση τετραγωνική απόκλιση ονομάζεται **διακύμανση** και συμβολίζεται με S^2

$$s^2 = \frac{(t_1 - \bar{x})^2 + (t_2 - \bar{x})^2 + (t_3 - \bar{x})^2 + \dots + (t_v - \bar{x})^2}{v} = \frac{\sum_{i=1}^v (x_i - \bar{x})^2}{v}$$

Διακύμανση – s^2

Η **διακύμανση** ενός δείγματος τείνει συστηματικά να υποεκτιμά τη διακύμανση του πληθυσμού, μια υποεκτίμηση που είναι εντονότερη σε δείγματα μικρού μεγέθους.

Για τον λόγο αυτό για την εκτίμηση της διακύμανσης του πληθυσμού εισάγεται μια διόρθωση προκειμένου να έχουμε αμερόληπτη εκτίμηση της διακύμανσης του πληθυσμού

$$s^2 = \frac{(t_1 - \bar{x})^2 + (t_2 - \bar{x})^2 + (t_3 - \bar{x})^2 + \dots + (t_v - \bar{x})^2}{v-1} = \frac{\sum_{i=1}^v (x_i - \bar{x})^2}{v-1}$$

Διακύμανση – s^2

A/A	Μάθημα Γ	Τετραγωνική απόκλιση = $(X - \bar{X})^2$
1	4	$(4-5)^2 = 1$
2	5	$(5-5)^2 = 0$
3	5	$(5-5)^2 = 0$
4	5	$(5-5)^2 = 0$
5	5	$(5-5)^2 = 0$
6	6	$(6-5)^2 = 1$
	$\bar{X} = 5$	$\sum (X - \bar{X})^2 = 2$

A/A	Μάθημα ΣΤ	Τετραγωνική απόκλιση = $(X - \bar{X})^2$
1	3	$(3-5)^2 = 4$
2	3	$(3-5)^2 = 4$
3	3	$(3-5)^2 = 4$
4	4	$(4-5)^2 = 1$
5	8	$(8-5)^2 = 9$
6	9	$(9-5)^2 = 16$
	$\bar{X} = 5$	$\sum (X - \bar{X})^2 = 38$

η διακύμανση για το μάθημα Γ είναι $S^2 = 2/5 = 0,4$ τετραγωνικές μονάδες

και για το μάθημα ΣΤ είναι $S^2 = 38/5 = 7,6$ τετραγωνικές μονάδες

Διακύμανση – s^2

- Δηλώνει πόσο μακριά από τη μέση τιμή απέχουν οι παρατηρήσεις
- Μέτρο της απόστασης των παρατηρήσεων από το μέσο
 - Όταν οι τιμές απέχουν πολύ από τη μέση τιμή η διακύμανση είναι μεγάλη
 - Όταν οι τιμές δεν διαφέρουν πολύ από τη μέση τιμή, η διακύμανση είναι μικρή
- Έχει αξία όταν συγκρίνουμε μεταξύ τους δύο διαφορετικά σύνολα δεδομένων:
 - Αν η διακύμανση του πρώτου συνόλου είναι μικρότερη από τη διακύμανση του δεύτερου, τότε οι τιμές του πρώτου είναι σε μεγαλύτερο ποσοστό συγκεντρωμένες γύρω από τον αριθμητικό μέσο σε σχέση με το δεύτερο σύνολο.

Πρόβλημα: οι μονάδες είναι υψωμένες στο τετράγωνο, π.χ. 33,2 (αιτήσεις)²

Διακύμανση – s^2

Βήματα υπολογισμού:

1. Υπολογίζουμε τον μέσο όρο των παρατηρήσεων
2. Υπολογίζουμε την απόκλιση κάθε τιμής από τον μέσο όρο
3. Υψώνουμε τις αποκλίσεις στο τετράγωνο
4. Αθροίζουμε τα τετράγωνα των αποκλίσεων
5. Διαιρούμε με το (πλήθος των παρατηρήσεων - **1**)

Εφαρμογή 4

Να υπολογίσετε τη μέση τιμή και τη διακύμανση των παρακάτω παρατηρήσεων:

1 4 5 7 8

Διακύμανση – s^2

Βήματα υπολογισμού από πίνακα κατανομής συχνοτήτων:

1. Υπολογίζουμε τον μέσο όρο των παρατηρήσεων
2. Υπολογίζουμε την απόκλιση κάθε τιμής από τον μέσο όρο
3. Υψώνουμε τις αποκλίσεις στο τετράγωνο
4. Πολλαπλασιάζουμε με την αντίστοιχη συχνότητα εμφάνιση της τιμής
5. Αθροίζουμε τα τετράγωνα των αποκλίσεων επί την απόλυτη συχνότητα
6. Διαιρούμε με το (πλήθος των παρατηρήσεων -1)

$$s^2 = \frac{(x_1 - \bar{x})^2 v_1 + (x_2 - \bar{x})^2 v_2 + \dots + (x_k - \bar{x})^2 v_k}{v_k - 1}$$

Εφαρμογή 5

Δίνεται ο διπλανός πίνακας κατανομής συχνοτήτων, στον οποίο φαίνεται η βαθμολογία (1 - 5 αστέρια) που έβαλαν σ' ένα εστιατόριο 20 άτομα. Να υπολογίσετε τη διακύμανση του παραπάνω δείγματος.

Βαθμολογία x_i	v_i
1	4
2	2
3	7
4	4
5	3
Σύνολο	20

Εφαρμογή 5

x_i	v_i	$x_i v_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 v_i$
1	4				
2	2				
3	7				
4	4				
5	3				
Σύνολο	20				

Τυπική απόκλιση

Αν πάρουμε την τετραγωνική ρίζα του s^2 , τότε έχουμε έναν δείκτη διασποράς με μονάδες ίδιες με αυτές του χαρακτηριστικού.

Η εξαγωγή της ρίζας «αποκαθιστά» τον τετραγωνισμό των αποκλίσεων και επαναφέρει τις τετραγωνικές μονάδες στις αρχικές.

Η θετική τετραγωνική ρίζα του εκτιμητή διακύμανσης ονομάζεται **τυπική απόκλιση** και συμβολίζεται με s .

Συντελεστής μεταβλητότητας

Όταν θέλουμε να συγκρίνουμε ως προς τη **μεταβλητότητα ή ομοιογένεια** δύο κατανομές, οι οποίες είτε εκφράζονται σε διαφορετικές μονάδες μέτρησης (πχ 'ύψος και βάρος) είτε εκφράζονται στην ίδια μονάδα μέτρησης αλλά έχουν σημαντικά διαφορετικές μέσες τιμές, τότε μας χρειάζεται ένας τρόπος να μετράμε τη διασπορά ενός δείγματος με έναν **δείκτη απαλλαγμένο από μονάδες**.

Αυτός ο δείκτης είναι ο **συντελεστής μεταβλητότητας**, συμβολίζεται με **cv** και εκφράζει την τυπική απόκλιση ως ποσοστό του μέσου όρου.

$$CV = \frac{s}{|\bar{x}|}$$

Συντελεστής μεταβλητότητας

Ο συντελεστής μεταβλητότητας CV έχει τις εξής ιδιότητες:

- Συνήθως εκφράζεται επί τοις εκατό και είναι ανεξάρτητος από τις μονάδες μέτρησης (είναι δηλαδή καθαρός αριθμός).
- Αποτελεί μέτρο **σχετικής διασποράς** και εκφράζει την τυπική απόκλιση στο ποσοστό της μέσης τιμής.
- Όταν για δύο δείγματα **A** και **B** ισχύει ότι $CV_A < CV_B$, θα λέμε ότι **το δείγμα A έχει μεγαλύτερη ομοιογένεια από το δείγμα B**

Εφαρμογή 7

Δίνονται τα παρακάτω δείγματα τιμών:

- Δείγμα A: 1, 3, 4, 5, 7
- Δείγμα B: 3, 7, 7, 9, 9, 13

Να συγκρίνετε τα δείγματα A και B ως προς την ομοιογένεια.

Εφαρμογή 8

Οι βαθμοί του πρώτου τετραμήνου 30 μαθητών της Γ' Λυκείου στην Ιστορία είναι

15	17	19	18	16
20	16	18	15	20
19	15	17	20	16
17	16	15	15	18
16	19	18	17	15
15	17	15	20	16

- A) Να βρείτε τη διάμεσο των βαθμών,
- B) Να υπολογίσετε τη μέση τιμή των βαθμών,
- Γ) Να υπολογίσετε τη διακύμανση των βαθμών

Εφαρμογή 11

Οι βαθμοί 8 μαθητών στο πρώτο τετράμηνο στα Μαθηματικά είναι:

18, 10, 16, 14, 10, 20, α , 2α

και η μέση βαθμολογία τους είναι 14.

α) Να βρείτε το α .

β) Να βρείτε τη διάμεσο των παραπάνω βαθμών.

γ) Να υπολογίσετε τη διακύμανση των παραπάνω βαθμών