



Ποιοτική & Ποσοτική Ανάλυση Δεδομένων

Εβδομάδα 5^η – 6^η

Δείκτες Κεντρικής Τάσης και Διασποράς

Παιδαγωγικό Τμήμα Δημοτικής Εκπαίδευσης

Δημοκρίτειο Πανεπιστήμιο Θράκης

Αλεξανδρούπολη,

2014 - 2015

Εμπειρικές Στατιστικές Κατανομές

- Τα προβλήματα που γεννιούνται κατά την σύγκριση δύο ή περισσότερων στατιστικών κατανομών, κάνουν φανερή την ανάγκη υπολογισμού ορισμένων **συγκεντρωτικών τιμών ή παραμέτρων** οι οποίες:
 - α) θα χαρακτηρίζουν τη θέση στην οποία τείνουν να συγκεντρωθούν οι τιμές μιας μεταβλητής,
 - β) τον τρόπο που αυτές είναι διεσπαρμένες γύρω από τη θέση αυτή, καθώς και
 - γ) τη μορφή που παρουσιάζει γραφικά η κατανομή των δεδομένων τιμών της.

Εμπειρικές Στατιστικές Κατανομές

- Με τις στατιστικές παραμέτρους των εμπειρικών κατανομών παρέχεται η δυνατότητα:
 - ◆ Να καθοριστεί η **κεντρική θέση** της κατανομής ή, καλύτερα, να προσδιοριστεί το σημείο που αντιστοιχεί στην τιμή της μεταβλητής γύρω από την οποία τείνουν να συγκεντρωθούν οι μονάδες του πληθυσμού.
 - ◆ Να καθοριστεί η **διασπορά**, που προσδιορίζει κατά πόσο οι τιμές μιας μεταβλητής είναι συγκεντρωμένες ή όχι γύρω από μία παράμετρο θέσης.

Εμπειρικές Στατιστικές Κατανομές

- Με τις στατιστικές παραμέτρους των εμπειρικών κατανομών παρέχεται η δυνατότητα (συνέχεια):
 - Να καθοριστεί η **ασυμμετρία** της κατανομής, δηλαδή η απόκλιση που παρουσιάζει η καμπύλη συχνοτήτων από μια συγκεκριμένη συμμετρική καμπύλη, και η **κυρτότητά** της, δηλαδή κατά πόσο αιχμηρή ή πεπλατισμένη είναι η καμπύλη συχνοτήτων.
 - Να καθοριστεί η **απόκλιση της κατανομής από την κατάσταση ισότητας**, πόσο δηλαδή αποκλίνει η καμπύλη συγκέντρωσης της κατανομής από αυτή της ισοκατανομής των ατόμων του πληθυσμού με την κατανομή της συγκεκριμένης μεταβλητής.

Δείκτες Κεντρικής Τάσης

- ❑ Η Επικρατούσα ή Δεσπόζουσα Τιμή [Mode]
- ❑ Ο Αριθμητικός Μέσος [Mean]
- ❑ Η Διάμεσος [Median]
- ❑ Τα Τεταρτημόρια [Quartiles]

Καθορίζουν την τιμή της μεταβλητής γύρω από την οποία τείνουν να συγκεντρωθούν οι μονάδες του πληθυσμού.

Η Επικρατούσα Τιμή (Mode)

Παράδειγμα 1 (απλή κατανομή): Τεστ γλωσσικής ικανότητας 50 μαθητών της ΣΤ' δημοτικού (αξιολόγηση σε κλίμακα από το 1 έως το 10).

Τιμές	Συχνότητες (n_i)
1	2
2	0
3	3
4	6
5	10
6	11
7	7
8	6
9	3
10	2
Σύνολο	N=50

Ποιος βαθμός συναντάται τις περισσότερες φορές;

Η επικρατούσα τιμή αυτής της κατανομής είναι το 6.

Η Επικρατούσα Τιμή (Mode)

Παράδειγμα 2 (ομαδοποιημένη κατανομή):
Ηλικίες 124 επιβατών ενός αεροπλάνου.

Διαστήματα Κλίμακας	Συχνότητες (n_i)
19-23	11
24-28	14
29-33	23
34-38	29
39-43	21
44-48	16
49-53	10
Σύνολο	N=124

Ποιά είναι η επικρατέστερη ηλικία;

Η Επικρατούσα Τιμή (Mode)

Παράδειγμα 2 (συνέχεια) (ομαδοποιημένη κατανομή):
Ηλικίες 124 επιβατών ενός αεροπλάνου.

Διαστήματα Κλίμακας	Μέσο Διαστήματος	Συχνότητες (n _i)
19-23	21	11
24-28	26	14
29-33	31	23
34-38	36	29
39-43	41	21
44-48	46	16
49-53	51	10
Σύνολο		N=124

Δημιουργούμε
τη στήλη «μέσο
διαστήματος»

Η επικρατούσα τιμή αυτής της κατανομής είναι
το **36**, η μέση τιμή του διαστήματος **34-38**

Η Επικρατούσα Τιμή (Mode)

- Είναι η τιμή που συναντάται τις *περισσότερες φορές* σε μια κατανομή. Είναι δηλαδή, η τιμή με τη **μεγαλύτερη συχνότητα**
- Χρησιμοποιείται συνήθως για την περιγραφή **κατηγορικών (ποιοτικών)** μεταβλητών
- Μπορεί να υπάρχουν δύο ή και περισσότερες δεσπόζουσες τιμές σε μια κατανομή (**δίκορφη, ή πολύκορφη κατανομή**)

Η Επικρατούσα Τιμή (Mode)

Πλεονεκτήματα	Μειονεκτήματα
Δείχνει την πιο συχνή τιμή της κατανομής	Δεν λαμβάνει υπόψη την ακριβή τιμή του κάθε στοιχείου
Μένει ανεπηρέαστη από ακραίες τιμές	Δεν μπορεί να χρησιμοποιηθεί για να υπολογιστούν παράμετροι του πληθυσμού
Μπορεί να υπολογιστεί όταν οι ακραίες τιμές είναι άγνωστες	Δεν είναι πολύ χρήσιμη για μικρό αριθμό δεδομένων

Μέσος Όρος (Mean)

Παράδειγμα 1 (απλή κατανομή):

Τιμές: 15 18 16 24 27 19 22

$$\bar{X} = \frac{15 + 18 + 16 + 24 + 27 + 19 + 22}{7} = 20.14$$

Μέσος Όρος (Mean)

Παράδειγμα 2

(ομαδοποιημένη κατανομή, κατηγορική μεταβλητή):

Μάθημα	Βαθμολογία	Βαρύτητα
Νέα Ελληνικά	12	4
Αρχαία	16	4
Μαθηματικά	18	2
Φυσική	14	1

Ποιά είναι η μέση αξιολόγηση του μαθητή;

$$\bar{X} = \frac{12 \times 4 + 16 \times 4 + 18 \times 2 + 14 \times 1}{11} = \frac{162}{11} = 14,72$$

Μέσος Όρος (Mean)

Παράδειγμα 3

(ομαδοποιημένη κατανομή, συνεχής μεταβλητή):

Ώρες απουσίας	Συχνότητες (n _i)
25-29	2
30-34	9
35-39	24
40-44	31
45-49	22
50-54	13
55-59	3
60-64	3
65-69	3
Σύνολο	N=124

Δίνονται τα διαστήματα των ωρών απουσίας 124 μαθητών κατά τη διάρκεια μιας σχολικής χρονιάς.

Πόσες απουσίες έκαναν οι μαθητές κατά μέσο όρο;

Μέσος Όρος (Mean)

Παράδειγμα 3

(ομαδοποιημένη κατανομή, συνεχής μεταβλητή):

Ώρες απουσίας	Μέσο Διαστήματος	Συχνότητες (n_i)
25-29	27	2
30-34	32	9
35-39	37	24
40-44	42	31
45-49	47	22
50-54	52	13
55-59	57	3
60-64	62	3
65-69	67	3
Σύνολο		N=124

Δημιουργούμε
τη στήλη «μέσο
διαστήματος»

Μέσος Όρος (Mean)

Παράδειγμα 3

(ομαδοποιημένη κατανομή, συνεχής μεταβλητή):

Ώρες απουσίας	Μέσο Διαστήματος	Συχνότητες (n _i)
25-29	27	2
30-34	32	9
35-39	37	24
40-44	42	31
45-49	47	22
50-54	52	13
55-59	57	3
60-64	62	3
65-69	67	3
Σύνολο		N=124

$$\bar{X} = \frac{27 \times 2 + 32 \times 9 + 37 \times 24 + \dots + 67 \times 3}{124} = \dots$$

Μέσος Όρος (Mean)

- Η **μέση τιμή** της κατανομής, που ορίζεται ως το πηλίκο του συνόλου των τιμών μιας κατανομής με τον αριθμό των ατόμων που συμμετέχουν σε αυτή.
- Είναι η πιο **αντιπροσωπευτική** τιμή της κατανομής.

Μέσος Όρος (Mean)

Και ο τύποι υπολογισμού....

Απλή Κατανομή

$$\bar{X} = \frac{\sum X_i}{N}$$

Ομαδοποιημένη Κατανομή

$$\bar{X} = \frac{\sum n_i X_i}{N}$$

Αριθμητής = άθροισμα των γινομένων των μέσων τιμών
κάθε διαστήματος επί τη συχνότητά τους

N= μέγεθος δείγματος (ή άθροισμα συχνοτήτων)

Μέσος Όρος (Mean)

Πλεονεκτήματα	Μειονεκτήματα
Μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού (παραμετρικά τεστ)	Είναι ευαίσθητος στις ακραίες τιμές των δεδομένων της κατανομής
	Επειδή υπολογίζεται αλγεβρικά, η τιμή του είναι πιθανό να μην ανήκει στις τιμές της κατανομής

Η Διάμεσος ή Διχοτόμος (Median)

Παράδειγμα 1. Ποια τιμή διχοτομεί τις υπόλοιπες;
(απλή κατανομή, περιττό πλήθος παρατηρήσεων)

Τιμές:	18	25	21	4	13	15	28	17	22
Ιεραρχημένες Τιμές:	4	13	15	17	18	21	22	25	28
Θέση:	1	2	3	4	5	6	7	8	9

Διάμεσος

$$\text{Θέση} : \frac{N + 1}{2} = \frac{10}{2} = 5$$

Η Διάμεσος ή Διχοτόμος (Median)

Παράδειγμα 2. Ποια τιμή διχοτομεί τις υπόλοιπες;
(απλή κατανομή, άρτιο πλήθος παρατηρήσεων)

Ιεραρχημένες Τιμές: 4 13 13 17 18 21 22 25



Διάμεσος
17,5

Θέση Ταξινόμησης: 1 2 3 4 5 6 7 8



$$\text{Θέση} : \frac{N + 1}{2} = \frac{9}{2} = 4,5$$

Μεσαία
Θέση 4,5

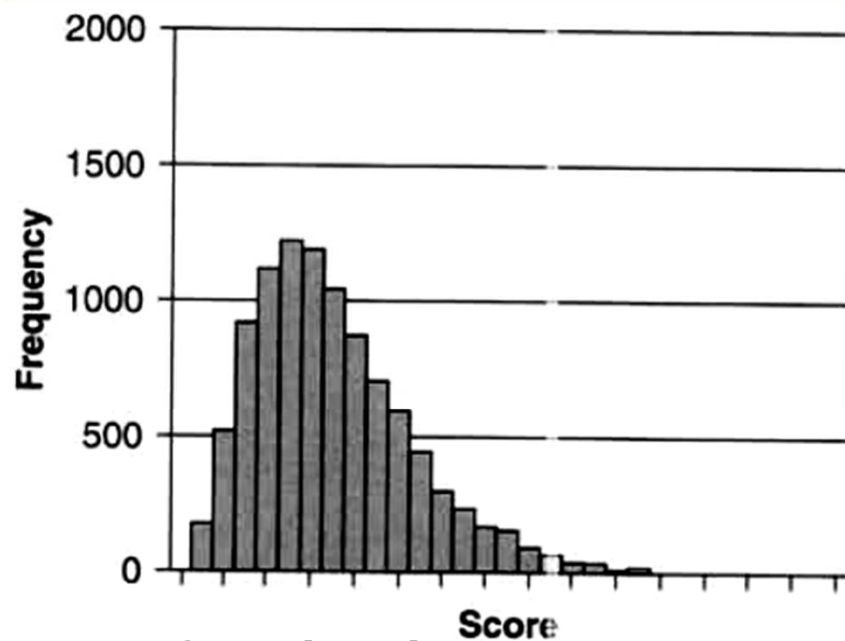
Η Διάμεσος ή Διχοτόμος (Median)

- Είναι η τιμή που χωρίζει την κατανομή σε **δύο ίσα** τμήματα
- Για να υπολογίσουμε τη διάμεσο μιας απλής κατανομής θα πρέπει πρώτα να **ιεραρχήσουμε** τις τιμές της (αρχίζοντας από τη μικρότερη)
- Στη συνέχεια επιλέγουμε την τιμή που βρίσκεται στη **μεσαία θέση** της κατανομής

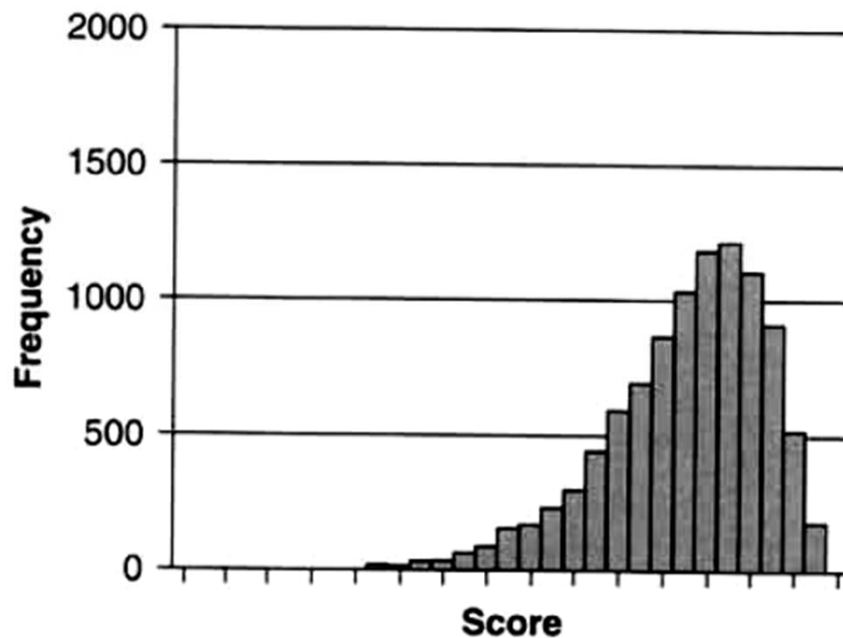
Η Διάμεσος ή Διχοτόμος (Median)

Πλεονεκτήματα	Μειονεκτήματα
Δεν επηρεάζεται από τις ακραίες τιμές, οπότε είναι ο καλύτερος δείκτης κεντρικής τάσης σε ασύμμετρη κατανομή	Δεν λαμβάνει υπόψη την ακριβή τιμή του κάθε στοιχείου
Μπορεί να υπολογιστεί ακόμη και όταν δεν γνωρίζουμε τις ακραίες τιμές	Δεν μπορεί να χρησιμοποιηθεί για να υπολογιστούν παράμετροι του πληθυσμού
	Εάν οι τιμές της κατανομής είναι λίγες τότε η διάμεσος δεν τις αντιπροσωπεύει με ακρίβεια

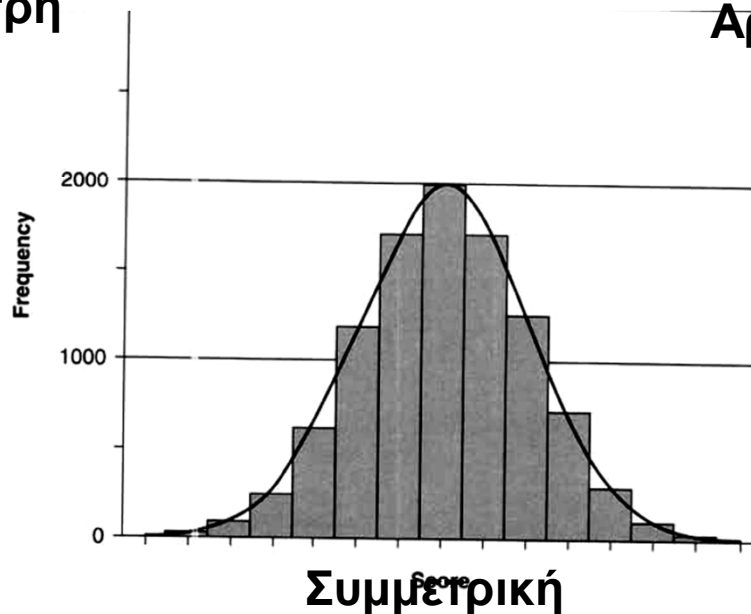
Σχέση μέσης, διάμεσης και επικρατέστερης



Θετικά Ασύμμετρη

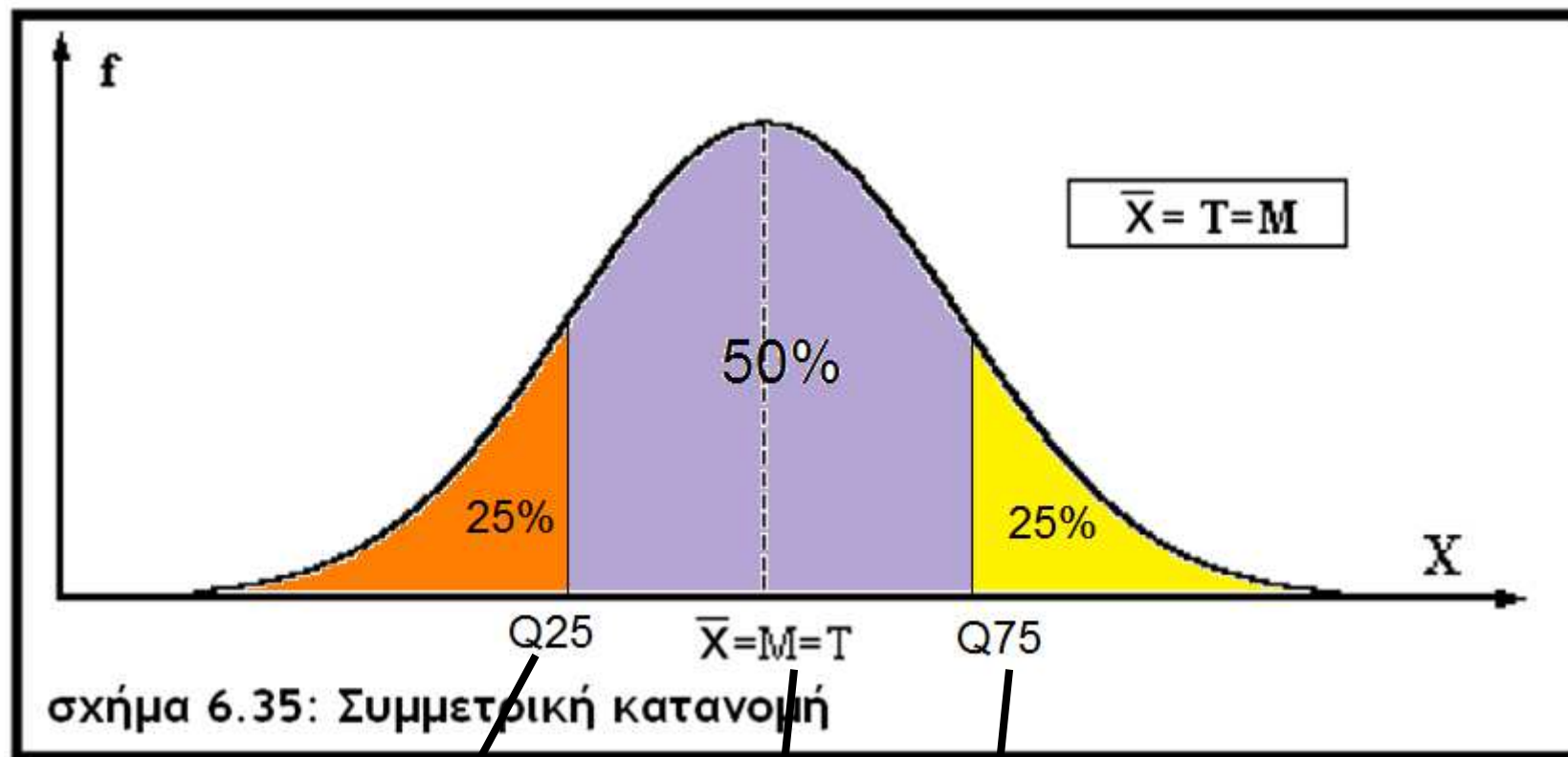


Αρνητικά Ασύμμετρη



Συμμετρική

Σχέση μέσης, διάμεσης και επικρατέστερης



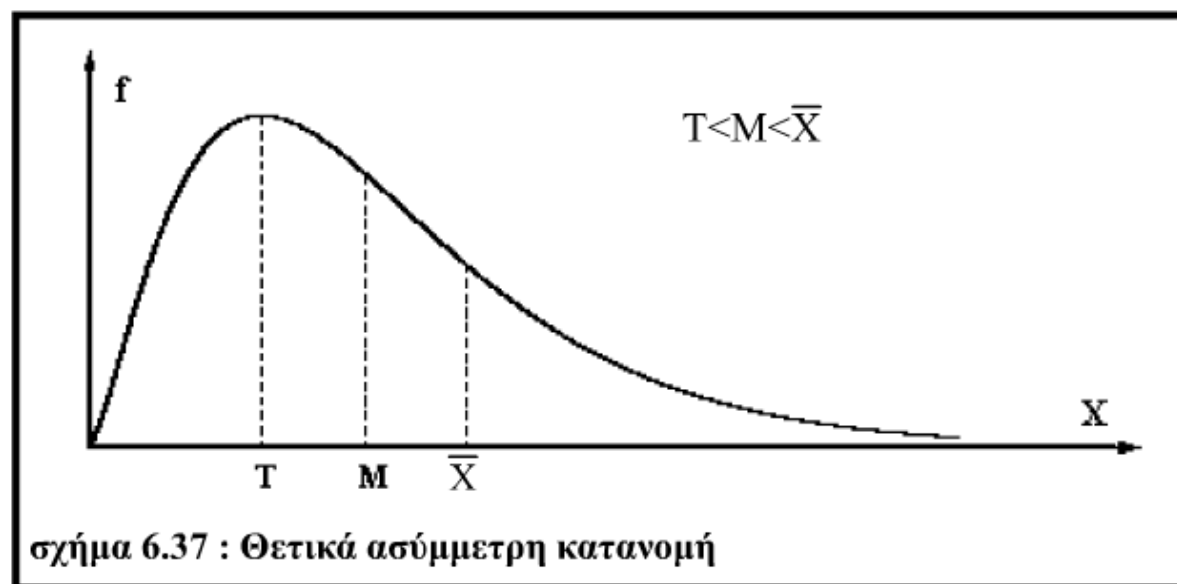
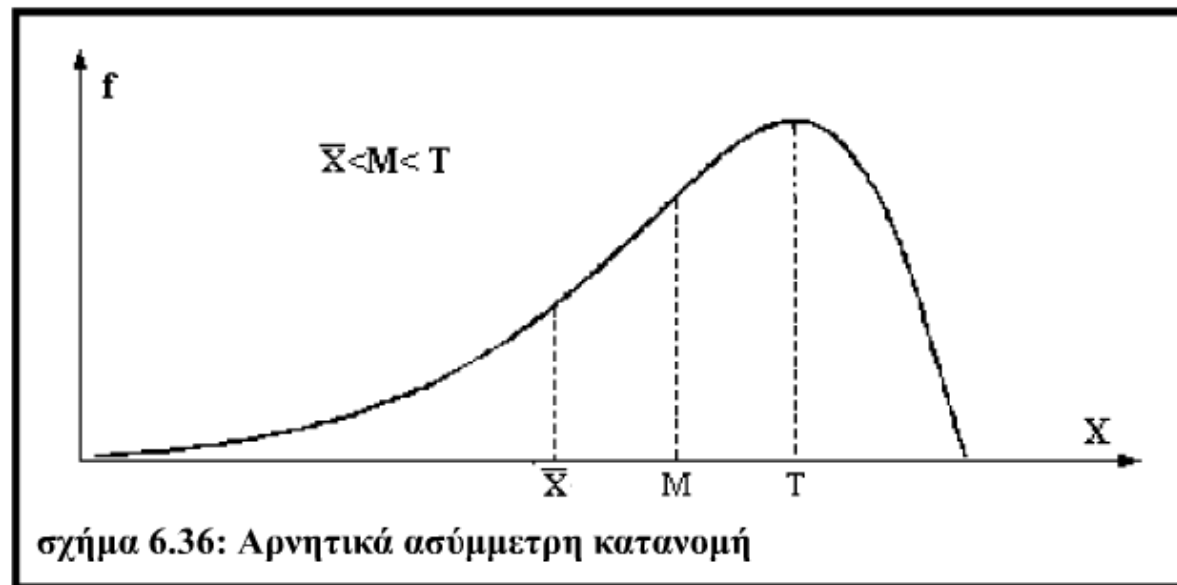
Βαθμολογία

12,5

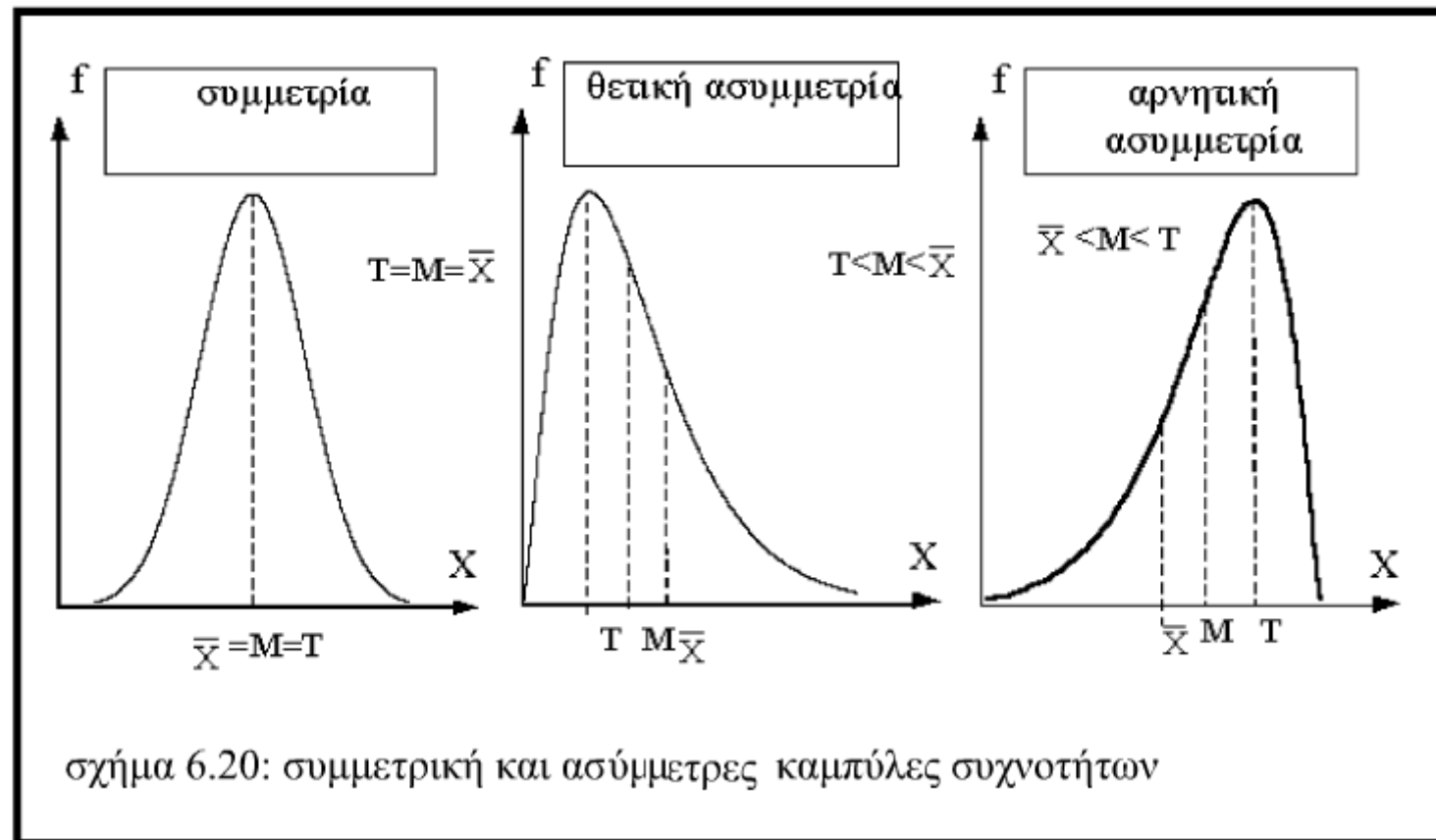
15

18

Σχέση μέσης, διάμεσης και επικρατέστερης



Σχέση μέσης, διάμεσης και επικρατέστερης



Ποιος δείκτης είναι ο πιο κατάλληλος;

- Όταν η κατανομή των δεδομένων μας είναι **συμμετρική**, τότε οι τιμές και των τριών δεικτών είναι ίδιες, αλλά προτιμάμε τον **μέσο όρο**.
- Όταν η κατανομή των δεδομένων μας είναι **ασύμμετρη** τότε καταλληλότερος δείκτης είναι η **διάμεσος**.
- Όταν υπάρχει ανάγκη για μια γρήγορη και κατά προσέγγιση **εκτίμηση** της κεντρικής τάσης, τότε μπορούμε να χρησιμοποιήσουμε τη **επικρατούσα τιμή**.

Τεταρτημόρια (Quartiles)

Τεταρτημόρια: Τα σημεία που χωρίζουν την κατανομή σε τέσσερα ίσα μέρη:

- ❑ 1ο τεταρτημόριο ή Q_1 ή Q_{25}
- ❑ 2ο τεταρτημόριο ή Q_1 ή Q_{25} ή διάμεσος
- ❑ 3ο τεταρτημόριο ή Q_3 ή Q_{75}

Τα τεταρτημόρια υπολογίζονται με παρόμοιο τρόπο όπως η διάμεσος.

Τεταρτημόρια (Quartiles) – Q25

Παράδειγμα 1. Κάτω από ποιά τιμή βρίσκεται το 25% των τιμών (απλή κατανομή)

Τιμές: 18 25 21 4 13 15 28 17 22

Ιεραρχημένες Τιμές: 4 13 15 17 18 21 22 25 28

Q25
14

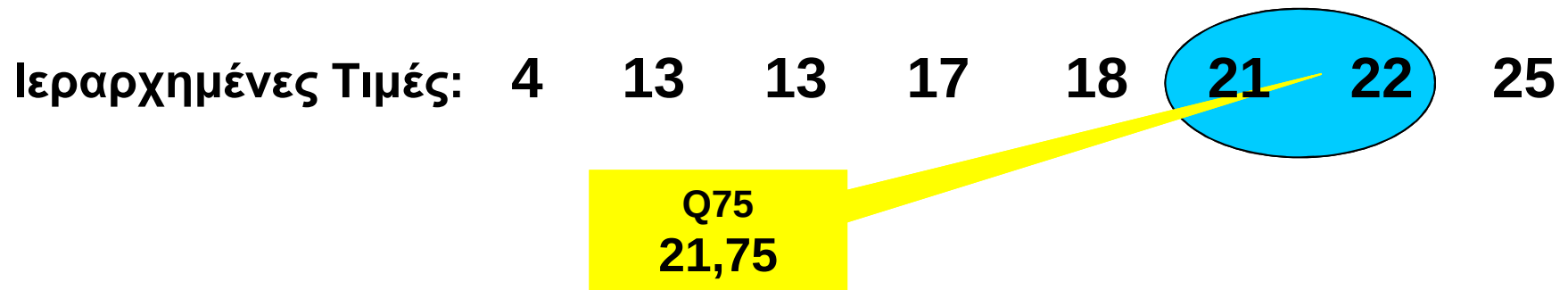
Θέση: 1 2 3 4 5 6 7 8 9

$$\text{Θέση} : \frac{N + 1}{4} = \frac{10}{4} = 2,5$$

Θέση 2,5

Τεταρτημόρια (Quartiles) – Q75

Παράδειγμα 2. Πάνω από ποιά τιμή βρίσκεται το 25% των τιμών (απλή κατανομή)



$$\text{Θέση: } 3 \frac{(N+1)}{4} = \frac{27}{4} = 6,75$$

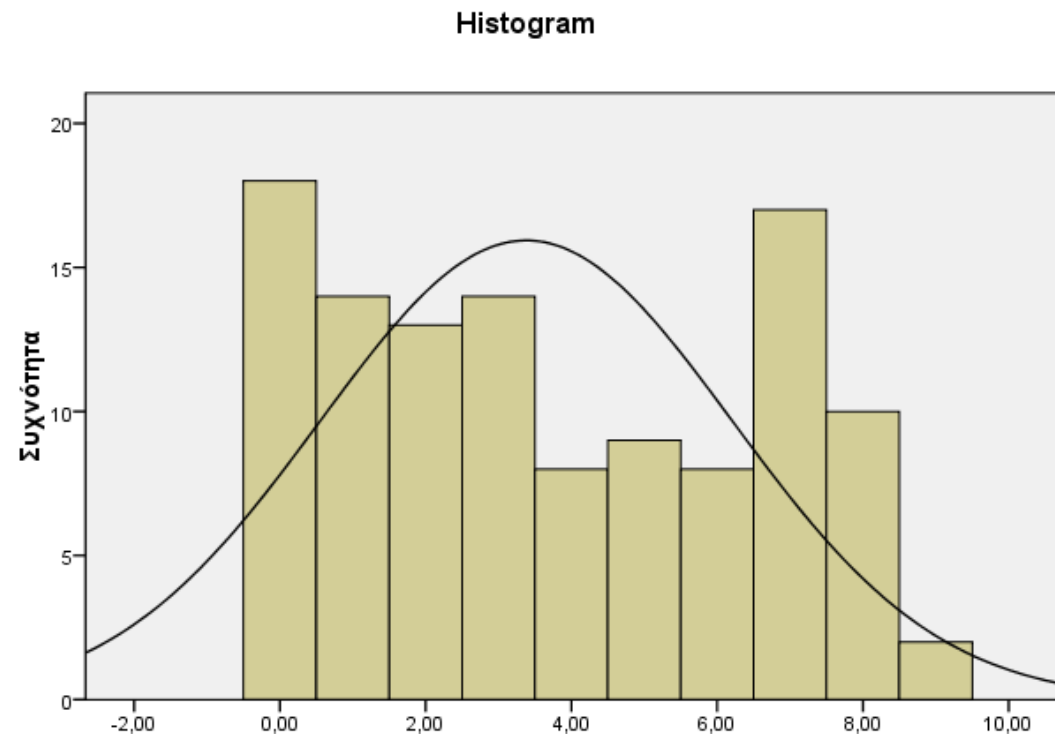
Παράδειγμα

■ Εύκολο ή Δύσκολο

Μερικά Στατιστικά από τις εξετάσεις ενός μαθήματος (κλίμακα αξιολόγησης 0-10):

- $N = 113$
- μέσος όρος, $\mu = 3,75$
- διάμεσος, $M = 3$
- επικρατούσα τιμή, $T = 0$
- $Q25 = 1$
- $Q75 = 7$

Σε ποια συμπεράσματα καταλήγετε;



Δείκτες Διασποράς

- ❑ **Εύρος** (range),
- ❑ **Ενδοτεταρτημοριακό εύρος** (interquartile range),
- ❑ **Μέση αριθμητική απόκλιση** (mean deviation)
- ❑ **Διακύμανση** (variance)
- ❑ **Τυπική απόκλιση** (standard deviation), και
- ❑ **Συντελεστής Μεταβλητότητας** (Coefficient of Variation)

Καθορίζουν τον τρόπο που αυτές είναι διεσπαρμένες γύρω από τις παραμέτρους κεντρικής θέσης.

Χρησιμότητα - Ένα παράδειγμα

Ένας καθηγητής Ιστορίας έδωσε σε δύο τμήματα μιας τάξης του σχολείου του το ίδιο τεστ. Η επίδοση των μαθητών του κάθε τμήματος (όπως μετρήθηκε με τη χρήση μιας εικοσαβάθμιας κλίμακας) παρουσιάζεται στον Πίνακα:

Τμήμα Α				Τμήμα Β			
17	18	19	14	20	16	14	17
18	17	13	16	17	20	9	16
20	13	17	17	19	16	17	20
14	17	15	18	15	17	9	16
16	12	19	17	17	10	20	17
14				19			

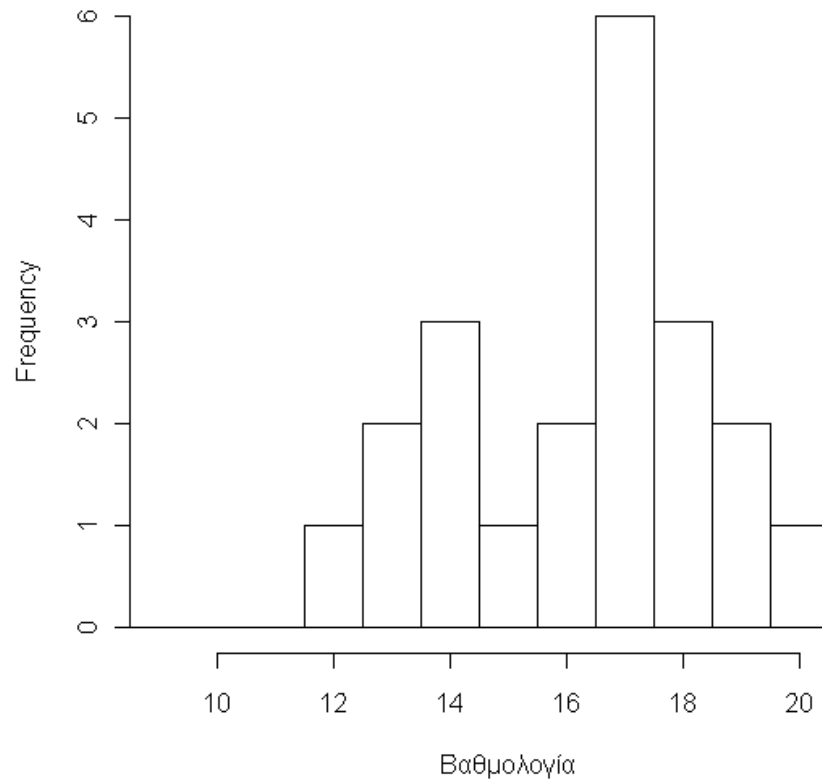
Ποιο από τα δύο τμήματα είχε καλύτερη επίδοση;

Χρησιμότητα - Ένα παράδειγμα

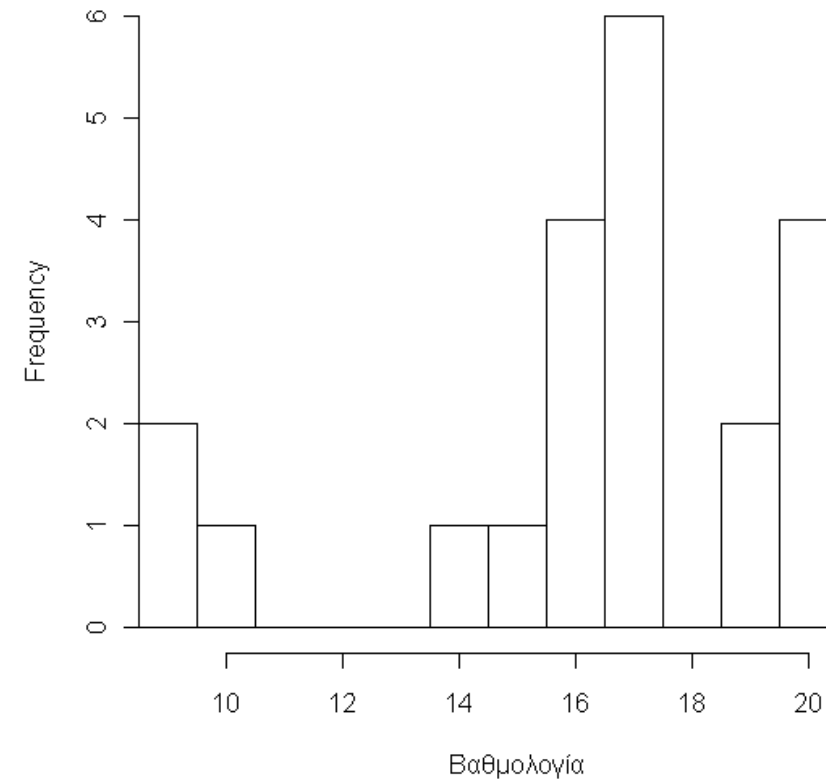
- ❑ Οι επιδόσεις των μαθητών των δύο τμημάτων δεν είναι οι ίδιες...
- ❑ Ωστόσο, οι τιμές και των τριών δεικτών κεντρικής τάσης είναι ακριβώς οι ίδιες μεταξύ των δύο ομάδων!
- ❑ Η επικρατούσα τιμή και για τα δύο τμήματα είναι η τιμή 17 (η συχνότητά της είναι 6 και στα δύο τμήματα)
- ❑ Η διάμεσος και στις δύο ομάδες είναι 16,75
- ❑ Τέλος, ο αριθμητικός μέσος όρος και των δύο ομάδων είναι επίσης ο ίδιος: 16,24
- ❑ Άρα;...

Χρησιμότητα - Ένα παράδειγμα

Τμήμα Α



Τμήμα Β



$$\bar{X}_1 = \bar{X}_2 = 16,24 \quad M_1 = M_2 = 16,75 \quad T_1 = T_2 = 17$$

Σε τι διαφέρουν τα δύο ιστογράμματα;

Εύρος

Η διαφορά μεταξύ της μέγιστης και της ελάχιστης τιμής της κατανομής

Πλεονεκτήματα:

- ❑ Είναι πολύ εύκολο στον υπολογισμό του
- ❑ Περιλαμβάνει και τις ακραίες τιμές της κατανομής

Μειονεκτήματα:

- ❑ Αλλοιώνεται από τις ακραίες τιμές με αποτέλεσμα, σε πολλές περιπτώσεις, να μην παρουσιάζει μια αντιπροσωπευτική εικόνα της διασποράς της κατανομής
- ❑ Δεν παρέχει καμιά πληροφορία σχετικά με τη διασπορά των τιμών μεταξύ των άκρων της κατανομής. Για παράδειγμα, δεν μας λέει τίποτα για τη διασπορά των τιμών της κατανομής γύρω από το μέσο όρο

Ενδοτεταρτημοριακό εύρος

Ενδοτεταρτημοριακό εύρος: Το εύρος του κεντρικού 50% των τιμών μιας κατανομής

- ❑ Το ενδοτεταρτημοριακό εύρος ισούται με τη διαφορά μεταξύ του 1ου και του 3ου τεταρτημορίου.
- ❑ Υπολογίζουμε την τιμή των δύο τεταρτημορίων:
 - ❑ 1ο τεταρτημόριο = 14,26
 - ❑ 3ο τεταρτημόριο = 17,75

Επομένως:

$$\text{3ο τεταρτημόριο} - \text{1ο τεταρτημόριο} = 17,75 - 14,26 = 3,49$$

Ενδοτεταρτημοριακό εύρος

Πλεονεκτήματα:

- ☐ Δεν επηρεάζεται από τις ακραίες τιμές
- ☐ Είναι σχετικά εύκολο στον υπολογισμό του
- ☐ Είναι αντιπροσωπευτικό των κεντρικών τιμών της κατανομής

Μειονεκτήματα:

- ☐ Δεν λαμβάνει υπόψη τις ακραίες τιμές της κατανομής
- ☐ Όπως και το εύρος, δεν επιτρέπει την ακριβή ερμηνεία μιας συγκεκριμένης τιμής της κατανομής
- ☐ Δεν είναι ακριβές όταν τα δεδομένα είναι ομαδοποιημένα κατά μεγάλα διαστήματα τιμών
- ☐ Όπως και η διάμεσος, δεν περιγράφει καμιά από τις παραμέτρους, οι οποίες είναι βασικές για την επαγωγική στατιστική

Διακύμανση και Τυπική Απόκλιση

Διακύμανση ή διασπορά (Variance) της κατανομής της μεταβλητής X είναι ο αριθμητικός μέσος των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από τη μέση τιμή της.

Ο τύπος υπολογισμού της πληθυσμιακής διακύμανσης

Απλή Κατανομή

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Ομαδοποιημένη Κατανομή

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n n_i (X_i - \bar{X})^2$$

Διακύμανση και Τυπική Απόκλιση

Τυπική απόκλιση (Standard Deviation): Δείκτης διασποράς αντιπροσωπευτικός των αποκλίσεων μιας ομάδας τιμών από τον μέσο όρο τους, πόσο απέχουν οι τιμές από τον μ.ο.

Χαμηλή τυπική απόκλιση σημαίνει ότι οι τιμές έχουν την τάση να βρίσκονται κοντά στον μ.ο., ενώ υψηλή τυπική απόκλιση σημαίνει ότι οι τιμές είναι διεσπαρμένες (υπάρχει μεγάλο εύρος τιμών).

Η τυπική απόκλιση ισούται με την τετραγωνική ρίζα της διασποράς.

$$\sigma = \sqrt{\sigma^2}$$

Διακύμανση και Τυπική Απόκλιση

Δίνονται οι τιμές 2, 4, 4, 4, 5, 5, 7, 9.

Ο μέσος όρος τους είναι $\frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5$

Υπολογίζουμε τη διαφορά κάθε τιμής από τη μέση τιμή και στη συνέχεια υψώνουμε στο τετράγωνο:

$$(2-5)^2 = (-3)^2 = 9 \quad (5-5)^2 = (0)^2 = 0$$

$$(4-5)^2 = (-1)^2 = 1 \quad (5-5)^2 = (0)^2 = 0$$

$$(4-5)^2 = (-1)^2 = 1 \quad (7-5)^2 = (2)^2 = 4$$

$$(4-5)^2 = (-1)^2 = 1 \quad (9-5)^2 = (4)^2 = 16$$

Έπειτα διαιρούμε το άθροισμά τους με το πλήθος των τιμών:

$$\frac{9 + 1 + 1 + 1 + 0 + 0 + 4 + 16}{8} = 4 \quad \text{Η διακύμανση είναι ίση με 4 και η τυπική απόκλιση ρίζα(4) = 2}$$

Διακύμανση και Τυπική Απόκλιση

Παράδειγμα. Επιδόσεις 21 μαθητών Γυμνασίου σε ένα διαγώνισμα. Υπολογισμός **διακύμανσης** και **τυπικής απόκλισης**.

Βαθμοί	$d = X_i - \bar{X}$	n_i	$n_i (X_i - \bar{X})^2$
20	20-16=4	1	16
19	19-16=3	2	18
18	18-16=2	3	12
17	17-16=1	6	6
16	16-16=0	2	0
15	15-16=-1	1	1
14	14-16=-2	3	12
13	13-16=-3	2	18
12	12-16=-4	1	16
Σύνολο		N=21	99

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n n_i (X_i - \bar{X})^2$$

$$= \frac{99}{21} = 4,71$$

$$\sigma = \sqrt{4,71} = 2,17$$

Διακύμανση και Τυπική Απόκλιση

Παράδειγμα. Επιδόσεις 21 μαθητών Γυμνασίου σε ένα διαγώνισμα. Υπολογισμός **διακύμανσης** και **τυπικής απόκλισης**.

α. Να υπολογιστεί η διακύμανση της βαθμολογίας.
(ομαδοποιημένη
διακριτή μεταβλητή)

κατανομή,

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n n_i (X_i - \bar{X})^2$$
$$= \frac{99}{21} = 4,71$$

β. Πόσο απέχουν οι βαθμολογίες κατά μ.ο. από τον μέσο βαθμό;

$$\sigma = \sqrt{4,71} = 2,17 \text{ βαθμοί}$$

Διακύμανση και Τυπική Απόκλιση

Παράδειγμα. Ηλικία 124 επιβατών ενός αεροπλάνου. Ποιά είναι η διάμεση ηλικία αυτών; (ομαδοποιημένη κατανομή, συνεχής μεταβλητή)

Διαστήματα Κλίμακας	Συχνότητα (n_i)
19-23	11
24-28	14
29-33	23
34-38	29
39-43	21
44-48	16
49-53	10
Σύνολο	N=124

Διακύμανση και Τυπική Απόκλιση

Παράδειγμα. Ηλικία 124 επιβατών ενός αεροπλάνου. Ποιά είναι η διάμεση ηλικία αυτών; (ομαδοποιημένη κατανομή, συνεχής μεταβλητή)

Διαστήματα Κλίμακας	Συχνότητα (n_i)
19-23 (21)	11
24-28 (26)	14
29-33 (31)	23
34-38 (36)	29
39-43 (41)	21
44-48 (46)	16
49-53 (51)	10
Σύνολο	N=124

Παίρνουμε το μέσο του διαστήματος...

Διακύμανση και Τυπική Απόκλιση

Παράδειγμα. Ηλικία 124 επιβατών ενός αεροπλάνου. Ποιά είναι η διάμεση ηλικία αυτών; (ομαδοποιημένη κατανομή, συνεχής μεταβλητή)

Διαστήματα Κλίμακας	Συχνότητα (n _i)
19-23 (21)	11
24-28 (26)	14
29-33 (31)	23
34-38 (36)	29
39-43 (41)	21
44-48 (46)	16
49-53 (51)	10
Σύνολο	N=124

...και εφαρμόζουμε τους τύπους της διακύμανσης, τυπικής απόκλισης, όπως προηγούμενα

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n n_i (X_i - \bar{X})^2$$

$$\sigma = \sqrt{\sigma^2}$$

Τυπική απόκλιση

Πλεονεκτήματα:

- ❑ Μπορεί να χρησιμοποιηθεί για τον υπολογισμό των παραμέτρων του πληθυσμού
- ❑ Λαμβάνει υπόψη όλες τις τιμές της κατανομής
- ❑ Είναι ο πιο ευαίσθητος από τους δείκτες διασποράς

Μειονεκτήματα:

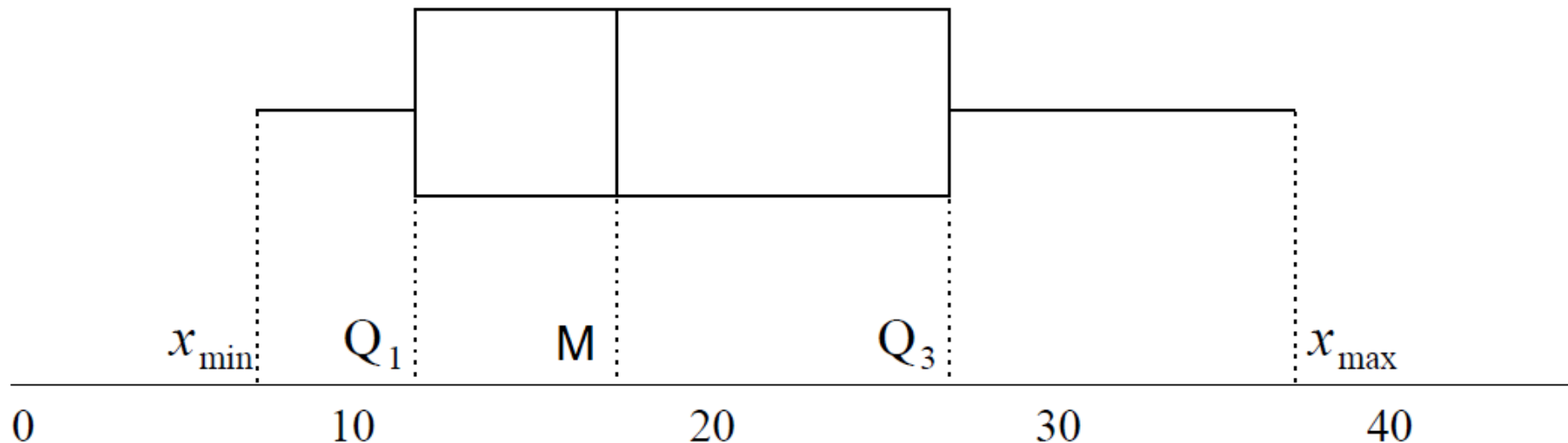
- ❑ Ο υπολογισμός της είναι σχετικά πιο περίπλοκος σε σχέση με τους υπόλοιπους δείκτες διασποράς
- ❑ Είναι πολύ ευαίσθητη στις ακραίες τιμές της κατανομής

Θηκόγραμμα - Boxplot (1)

- Το Boxplot είναι ένας βολικός τρόπος γραφικής αναπαράστασης μιας μεταβλητής, ως προς πέντε βασικές παραμέτρους που συνοψίζουν την κατανομή της:
 - την ελάχιστη τιμή (min)
 - το 1^ο τεταρτημόριο (Q25 ή Q1)
 - τη διάμεσο (Q50 ή Q2)
 - το 3^ο τεταρτημόριο (Q75 ή Q3)
 - και τη μέγιστη τιμή (max).

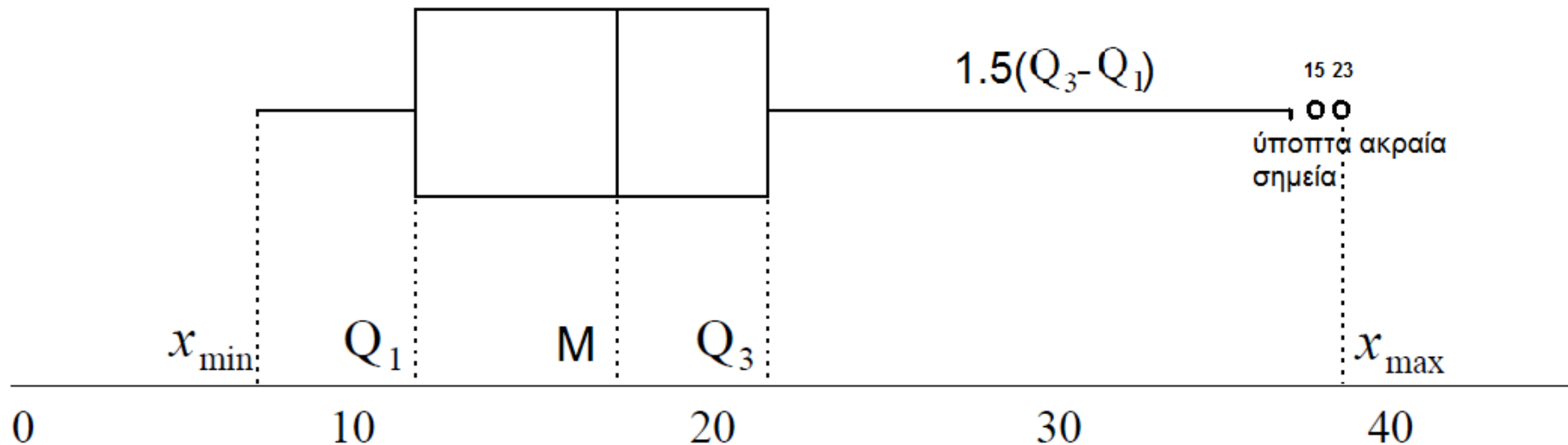
Με το boxplot, μπορούμε επιπλέον να εντοπίσουμε τυχόν ακραίες τιμές (outliers) μιας μεταβλητής.

Θηκόγραμμα - Boxplot (2)



- Κάθε πλαίσιο-κουτί απεικονίζει το Q_1 , τη διάμεσο (M) και το Q_3 .
- Οι απολήξεις (μουστάκια) υποδεικνύουν τα όρια των τιμών $\min$ και $\max$ όταν το μήκος τους είναι μικρότερο από $(1.5 \cdot (Q_3 - Q_1))$
- Τιμές που βρίσκονται πάνω από $(1.5 \cdot (Q_3 - Q_1))$ ονομάζονται **ύποπτα ακραίες (outlier)**.
- Τιμές που βρίσκονται πάνω από $(3 \cdot (Q_3 - Q_1))$ ονομάζονται **ακραίες (extreme)**.
- Κατανομές με συμμετρικά θηκογράμματα πλησιάζουν την κανονική.

Θηκόγραμμα - Boxplot (3)



- Κάθε πλαίσιο-κουτί απεικονίζει το Q_1 , τη διάμεσο (M) και το Q_3 .
- Οι απολήξεις (μουστάκια) υποδεικνύουν τα όρια των τιμών $\min$ και $\max$ όταν το μήκος τους είναι μικρότερο από $(1.5 \cdot (Q_3 - Q_1))$
- Τιμές που βρίσκονται πάνω από $(1.5 \cdot (Q_3 - Q_1))$ ονομάζονται **ύποπτα ακραίες (outlier)**.
- Τιμές που βρίσκονται πάνω από $(3 \cdot (Q_3 - Q_1))$ ονομάζονται **ακραίες (extreme)**.
- Κατανομές με συμμετρικά θηκογράμματα πλησιάζουν την κανονική.

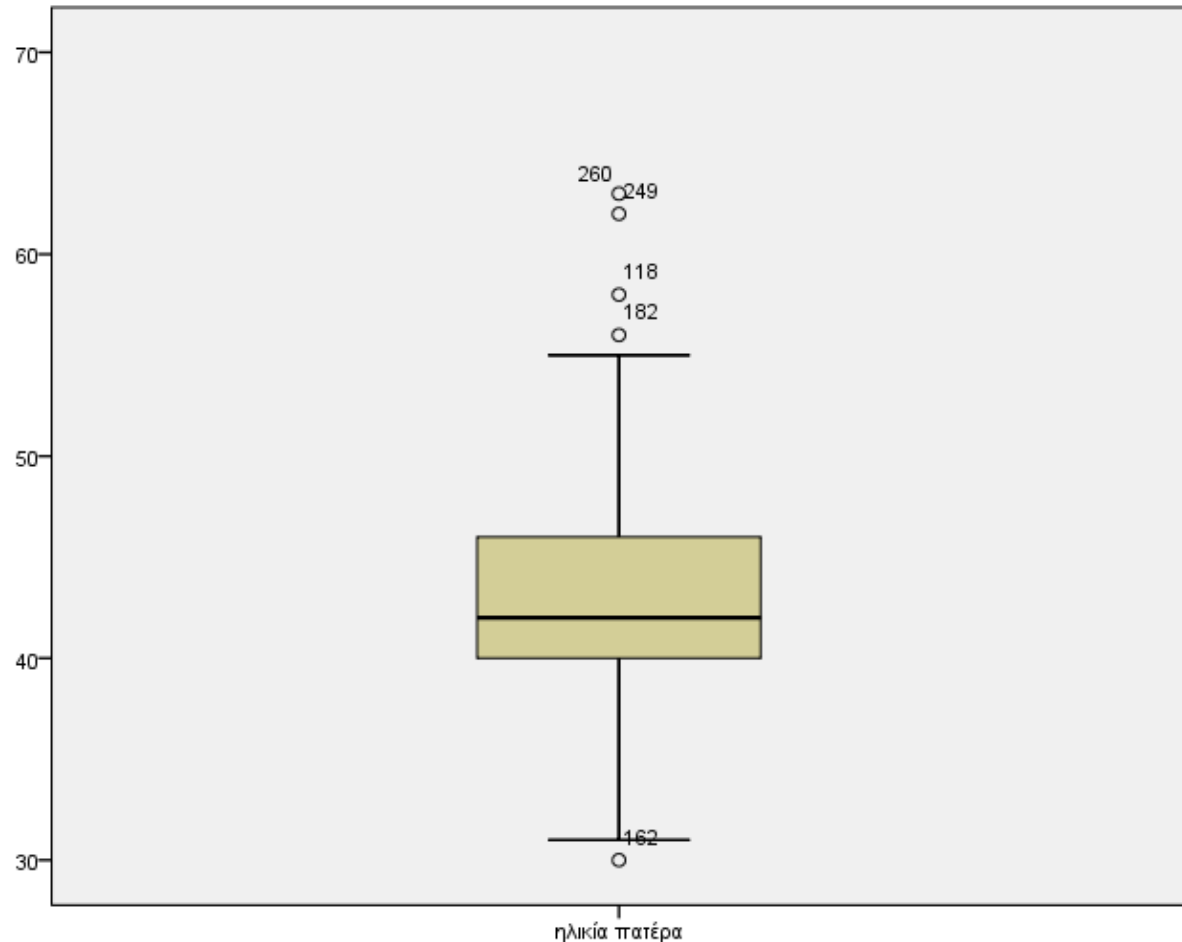
Θηκόγραμμα - Boxplot (4)

- Το boxplot, όπως και το ιστόγραμμα, μας επιτρέπει να κρίνουμε αν η κατανομή μιας συνεχούς μεταβλητής είναι κανονική. Για να κάνουμε αυτήν την παραδοχή θα πρέπει:
- η διάμεσος να μην αποκλίνει σημαντικά προς το πρώτο ή το τρίτο τεταρτομόριο, δηλαδή η γραμμή που αντιστοιχεί στη διάμεσο να μην πλησιάζει σε κάποιο από τα δύο άκρα του κουτιού (γιατί αλλιώς αυτό θα σήμαινε πως η κατανομή δεν πλησιάζει τη συμμετρική),
- το εύρος των τιμών στα δύο ακραία τεταρτημόρια να μη διαφέρει σημαντικά, δηλαδή τα μήκη των δύο απολήξεων να είναι συγκρίσιμα (για τη διατήρηση της συμμετρίας).
- να μην υπάρχουν ακραίες τιμές, δηλαδή να μην υπάρχουν σημεία μακριά από τις δύο απολήξεις.

Θηκόγραμμα - Boxplot (5)

➤ Οι παρατηρήσεις 260, 249, 118 και 182 (ο) είναι πιθανές ακραίες τιμές (μεγάλες ηλικίες) γιατί βρίσκονται πάνω από $1.5(Q3 - Q1)$.

➤ Η παρατήρηση 162 (ο) πιθανόν να είναι ακραία ηλικία (μικρή ηλικία) γιατί βρίσκεται κάτω από $1.5(Q3 - Q1)$.

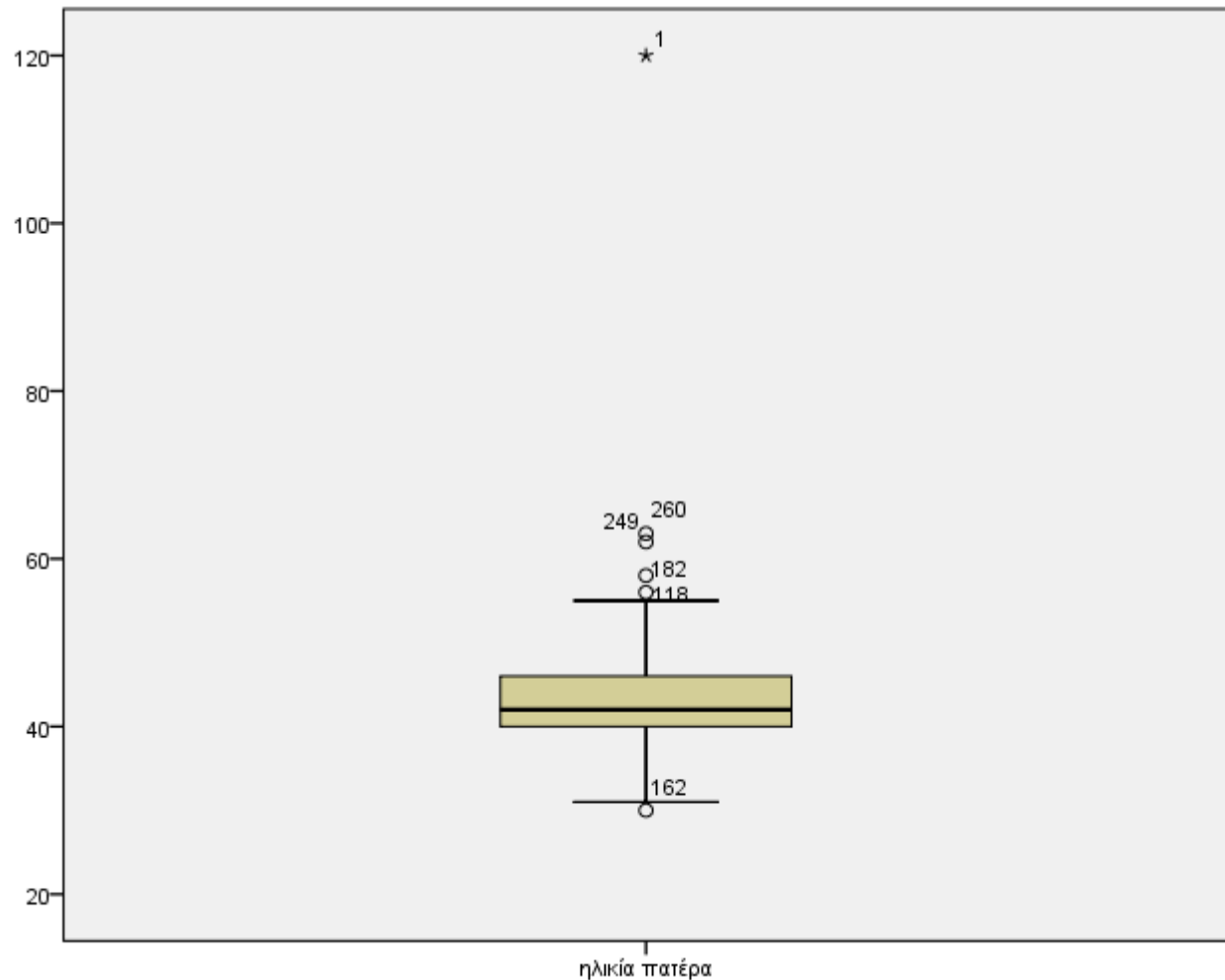


Θηκόγραμμα - Boxplot (6)

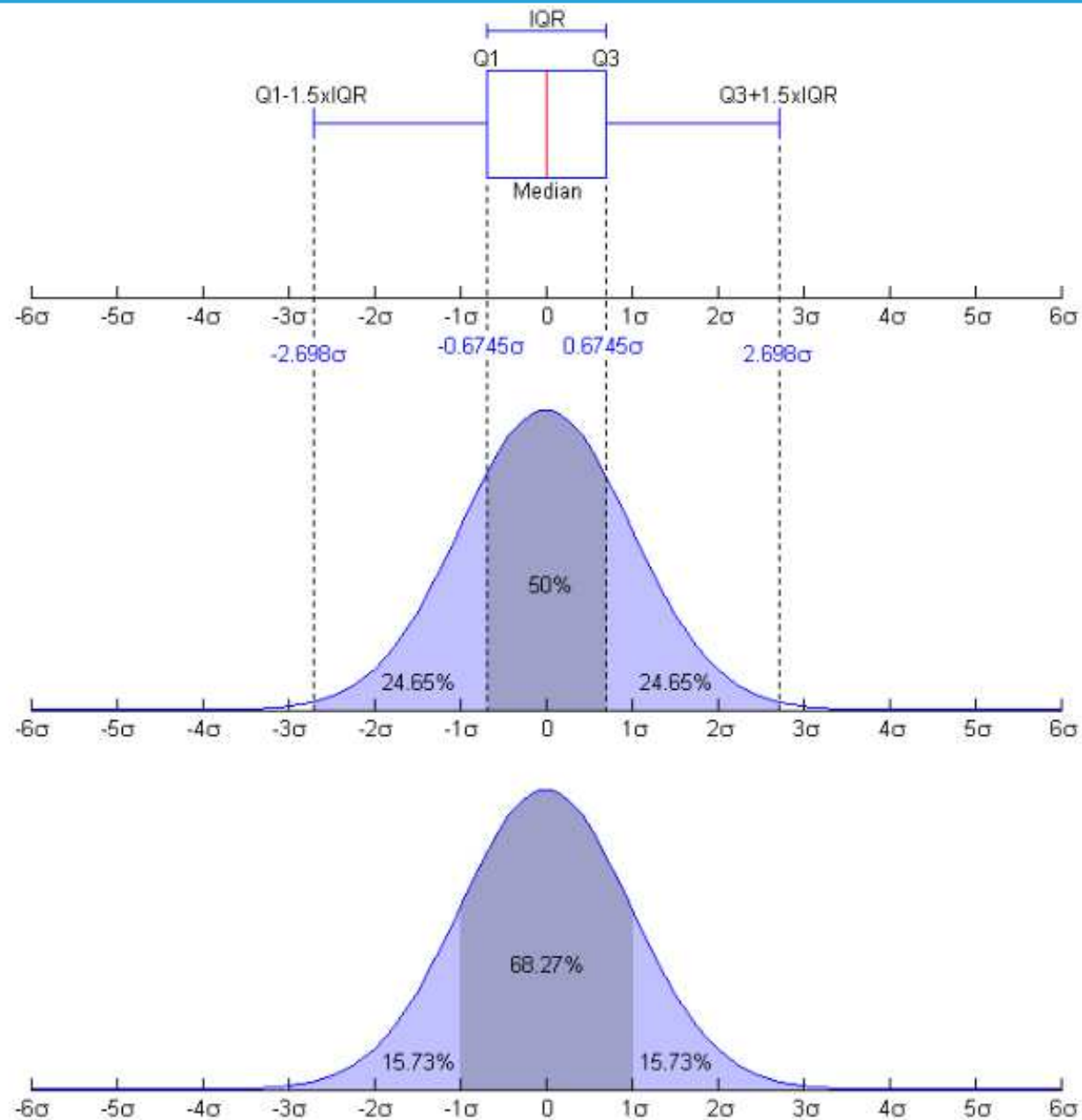
➤ Αλλάζουμε την ηλικία του πατέρα του 1^{ου} μαθητή σε μια πολύ υψηλή τιμή.

➤ Η παρατήρηση 1 (*) αποτελεί ακραία παρατήρηση της κατανομής (πολύ μεγάλη ηλικία) γιατί βρίσκεται πάνω από $3(Q3-Q1)$.

➤ Αυτή η παρατήρηση θα ξεχωρίζει επίσης και στο Ιστόγραμμα και το QQ-Plot.

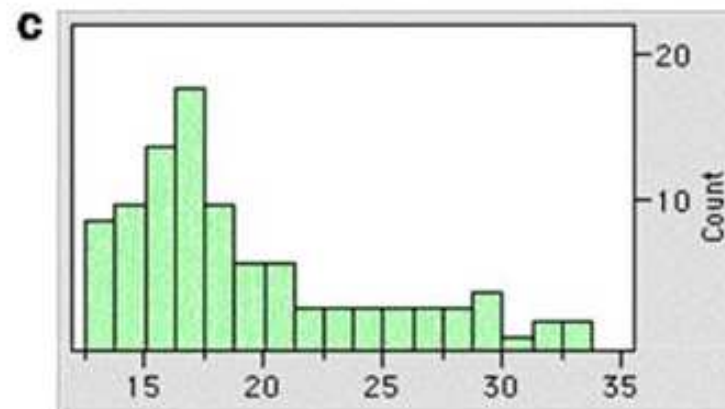
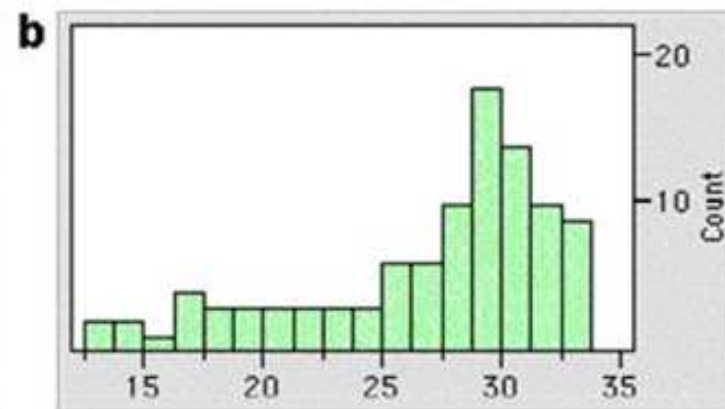
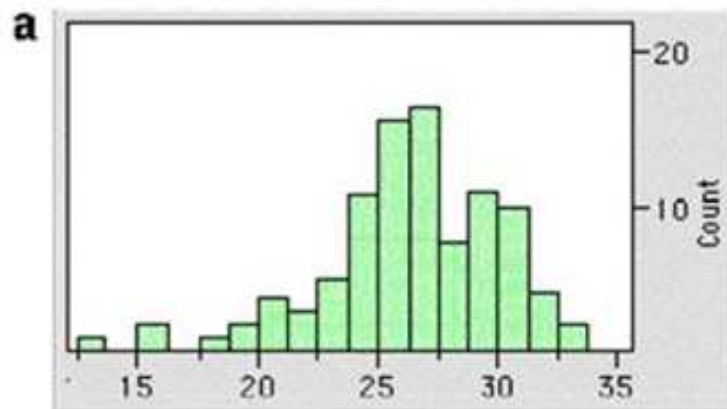


Θηκόγραμμα - Boxplot (7)



Άσκηση Κατανόησης #1

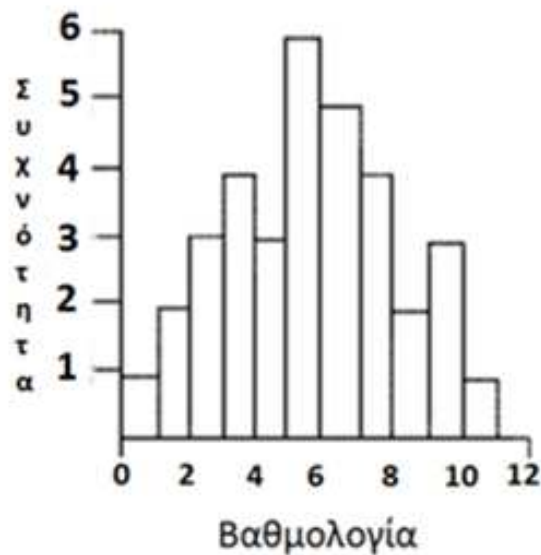
1. Στο πλαίσιο μιας έρευνας, επιλέχθηκε τυχαίο δείγμα 100 ψαριών ενός είδους που ζει σε μια λίμνη. Στη συνέχεια, μετρήθηκε το μήκος κάθε ψαριού σε χιλιοστά (mm). Τα παρακάτω στατιστικά συνοψίζουν την κατανομή του μήκους: μέσος όρος = 26,8mm, διάμεσος = 29,4mm, τυπική απόκλιση = 5mm, ελάχιστο = 12mm, μέγιστο = 33,4mm. Ποιο από τα παρακάτω ιστογράμματα είναι πιο πιθανό να αντιστοιχεί στα παραπάνω στατιστικά;



Άσκηση Κατανόησης #2

8. Στο παρακάτω ζευγάρι γραφημάτων, να χαρακτηρίσετε ποια μεταβλητή από αυτές που απεικονίζονται έχει την υψηλότερη τυπική απόκλιση (δεν είναι απαραίτητο να κάνετε υπολογισμούς)

A



B

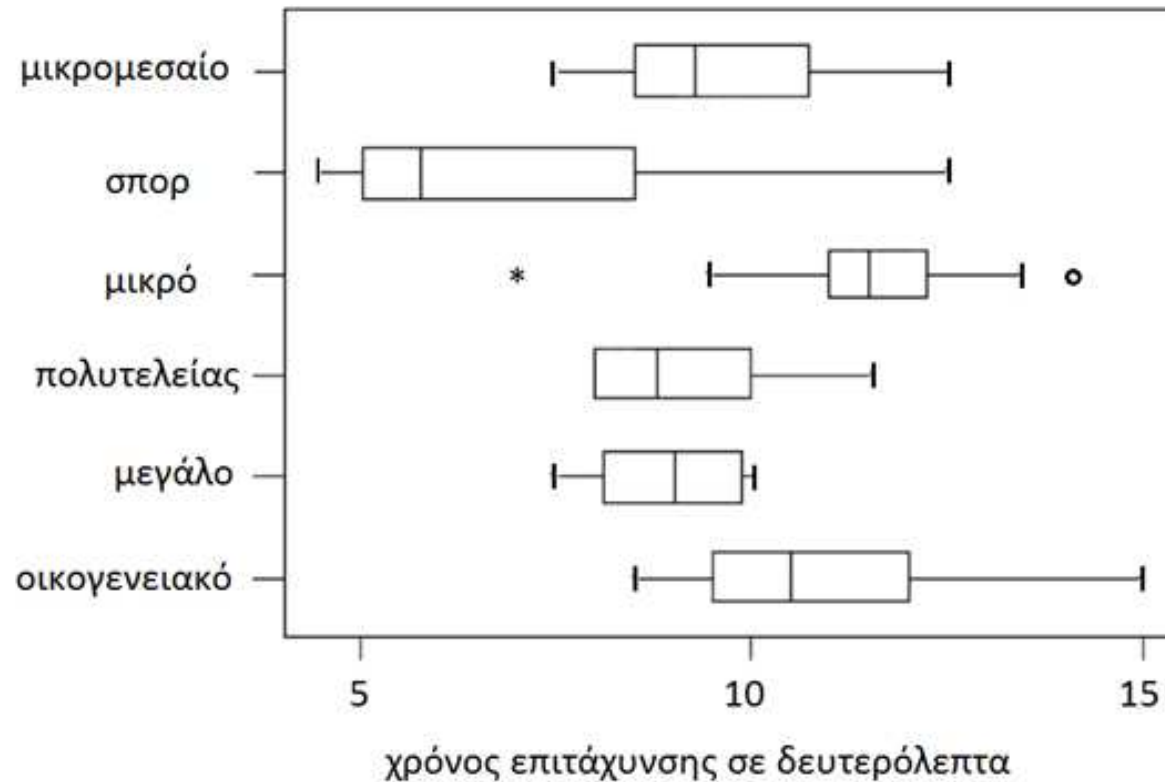


- α. Οι δύο μεταβλητές έχουν την ίδια τυπική απόκλιση
- β. Στο B η τυπική απόκλιση είναι μεγαλύτερη απ' ό,τι στο A
- γ. Στο A η τυπική απόκλιση είναι μεγαλύτερη απ' ό,τι στο B
- δ. Δεν μπορούμε να αποφανθούμε μόνο από το γράφημα

Εξηγήστε την απάντησή σας.

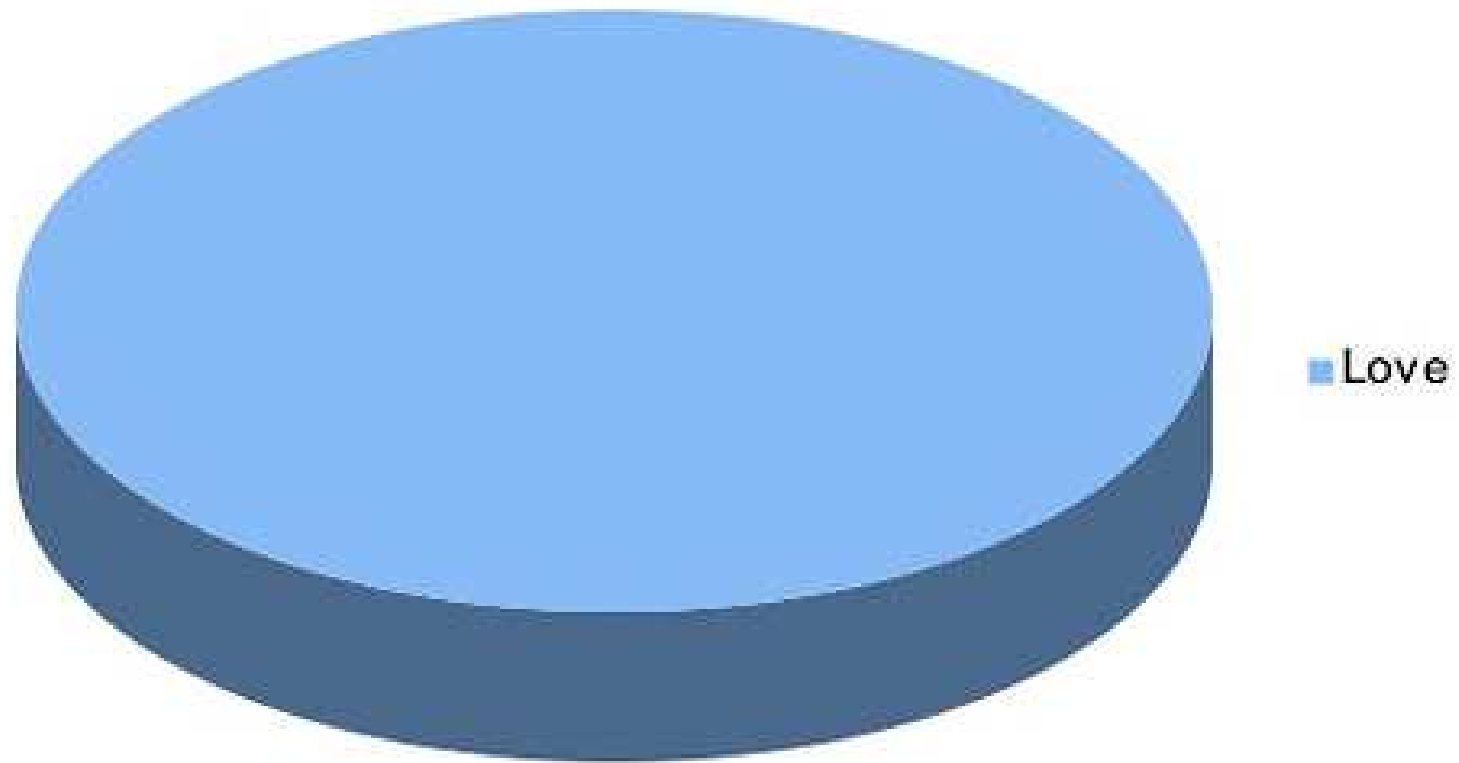
Άσκηση Κατανόησης #3

13. Το 2009 ένα περιοδικό για το αυτοκίνητο δημοσίευσε μία έρευνα για το χρόνο που χρειάζονται 250 διαφορετικά αυτοκίνητα να επιταχύνουν από τα 0 στα 100km/h. Τα αυτοκίνητα ταξινομήθηκαν σε έξι κατηγορίες ως προς τον τύπο τους. Το παρακάτω σύνθετο θηκόγραμμα αναπαριστά την κατανομή του χρόνου επιτάχυνσης για κάθε τύπο αυτοκινήτου.



Ποιες πληροφορίες αποκομίζετε από το γράφημα σχετικά με τις διαφορές που υπάρχουν ανάμεσα στις έξι κατηγορίες ως προς το χρόνο επιτάχυνσης (max 100 λέξεις);

Things you need



#end