Περιγραφική Στατιστική (Μέτρα διασποράς)

Έννοιες - Κλειδιά

- Μεταβλητότητα
- Εύρος (range)
- Εκατοστημόρια
- Ενδοτεταρτημοριακό εύρος
- Θηκόγραμμα (boxplot)
- Ακραίες τιμές (outliers)
- Διακύμανση
- Τυπική απόκλιση
- Εμπειρικός κανόνας
- Συντελεστής μεταβλητότητας

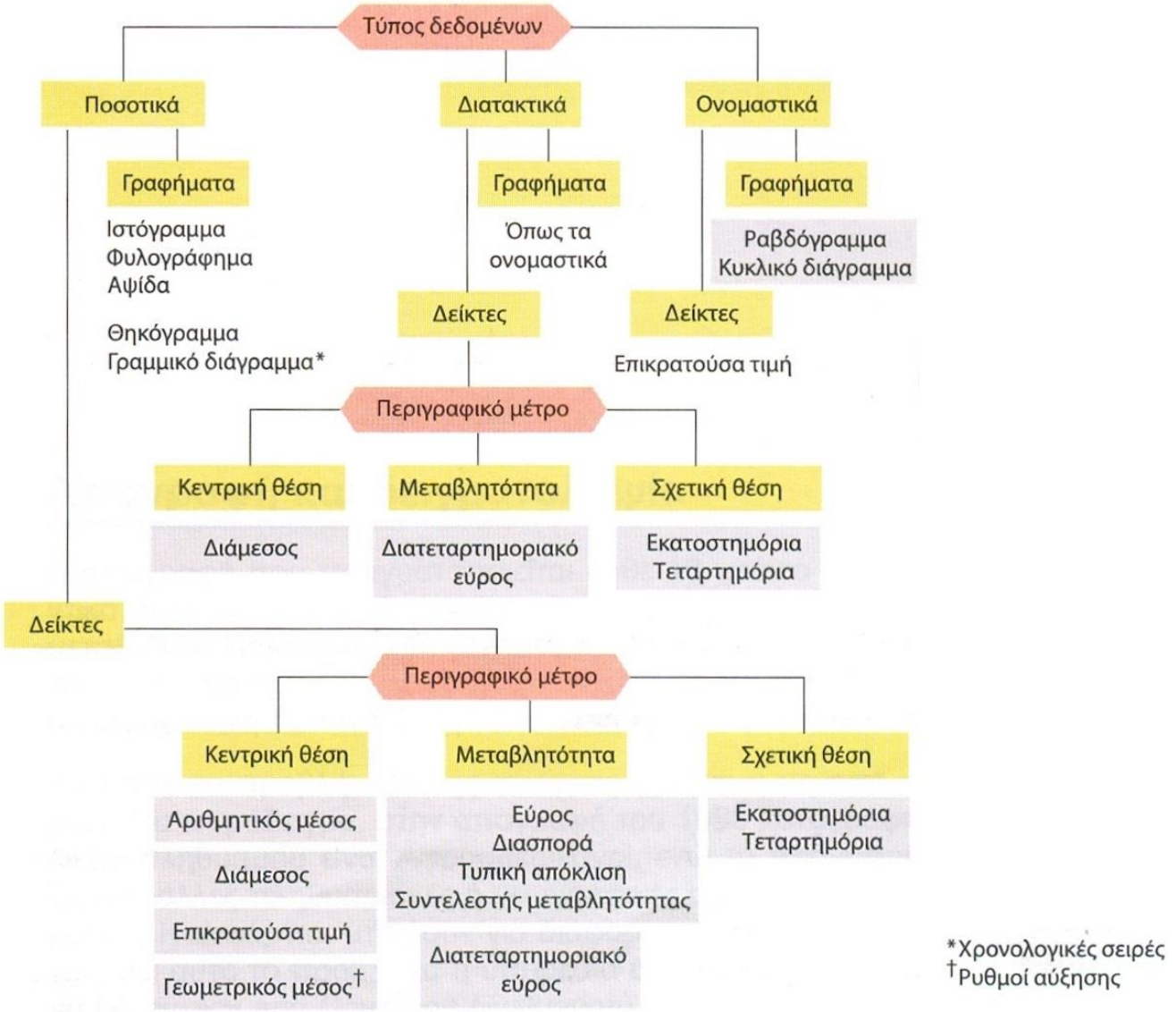
Μέτρα θέσης (σύνοψη)

Προσδιορίζουν ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα.

Τα κυριότερα μέτρα θέσης:

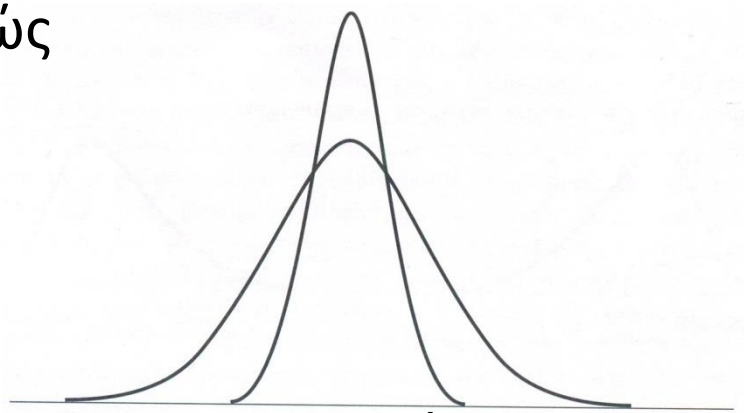
- Ο αριθμητικός μέσος *(**ποσοτικά** δεδομένα)*
- Η διάμεσος *(**ποσοτικά** ή **διατακτικά**)*
- Η επικρατούσα τιμή *(ποσοτικά , διατακτικά ή ονομαστικά)*

Περιγραφή ενός συνόλου δεδομένων



Μέτρα διασποράς – γιατί;; (1/2)

- Τα μέτρα κεντρικής τάσης δεν επαρκούν για την ακριβή περιγραφή ενός συνόλου αριθμητικών δεδομένων.
- Η αντιπροσωπευτικότητα τους **εξαρτάται** σε μεγάλο βαθμό από την ετερογένεια που παρουσιάζουν τα δεδομένα
- **Παράδειγμα**: Δύο κατανομές εντελώς διαφορετικές, οι οποίες έχουν ίδια:
  - μέση τιμή,
  - διάμεσο και
  - επικρατούσα τιμή



- Τα μέτρα διασποράς **στοχεύουν** στον προσδιορισμό της μεταβλητότητας (ή ετερογένειας) που παρουσιάζει ένα σύνολο δεδομένων.

Μέτρα διασποράς – γιατί;; (2/2)

- Ένα μέτρο διασποράς μας δίνει με τρόπο περιληπτικό και αντικειμενικό τη **μεταβλητότητα** ή **ανομοιογένεια** των παρατηρήσεων.

- Για να είναι ικανοποιητικό θα πρέπει να έχει τις εξής ιδιότητες:
  1. Να επηρεάζεται από τις διαφορές μεταξύ των τιμών και όχι από τη θέση τους και
  2. Να μεταβάλλεται αντίστροφα με τη συγκέντρωση των τιμών γύρω από ένα μέτρο θέσης

Μέτρα Διασποράς

- Τα κυριότερα μέτρα διασποράς είναι:
  - Το εύρος των τιμών
  - Τα εκατοστημόρια
  - Το ενδοτεταρτημοριακό εύρος
  - Η διακύμανση
  - Τυπική απόκλιση
  - Συντελεστής μεταβλητότητας CV
- Τα μέτρα αυτά χρησιμοποιούνται σε συνδυασμό με τα **μέτρα θέσης** και από κοινού περιγράφουν τις κατανομές δεδομένων με τρόπο συμπληρωματικό.

Εύρος (Range)

Range = $x_{max} - x_{min}$ (μεγαλύτερη – μικρότερη τιμή)

Παραδείγματα: {4, 4, 4, 4, 50}, Range = 46

{4, 8, 15, 24, 39, 50}, Range = 46

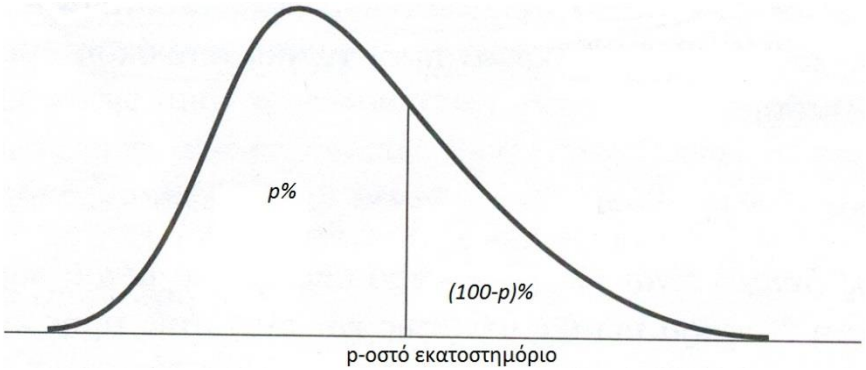
Πλεονέκτημα: Απλότητα στον υπολογισμό

Μειονέκτημα: Στον υπολογισμό του υπεισέρχονται ***μόνο δύο*** τιμές, οι πλέον ακραίες. Δεν φανερώνει την μεταβλητότητα των υπολοίπων.

>> Αντί για το **εύρος** καλύτερα να δίνεται η μέγιστη και η ελάχιστη τιμή των δεδομένων.

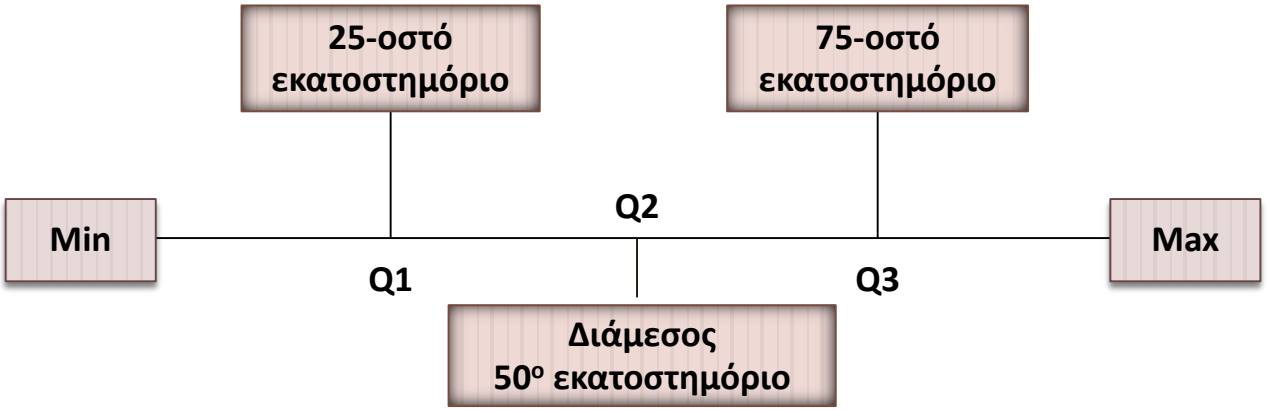
Εκατοστημόρια (ή εκατοστιαία σημεία)

- Τα **εκατοστημόρια** *(percentiles)* αποτελούν γενίκευση της έννοιας της διαμέσου *(median).*
- Το p-οστό εκατοστημόριο ενός συνόλου είναι εκείνη η τιμή, η οποία, όταν οι τιμές διαταχθούν σε **αύξουσα σειρά**, έχει από αριστερά της το p% των δεδομένων και από δεξιά της το υπόλοιπο (100-p)%



Εκατοστημόρια

- Τα συχνότερα εκατοστιαία σημεία είναι:
  - 25% : πρώτο τεταρτημόριο, Q_1
  - 50% : δεύτερο τεταρτημόριο, Q_2 (η διάμεσος)
  - 75% : τρίτο τεταρτημόριο, Q_3



Θέση εκατοστημορίου p%: $$L_p = (n+1)\frac{p}{100}$$

Quartiles = τεταρτημόρια

Εκατοστημόρια – Παράδειγμα (1/3)

Statistics

hours

N	Valid	10
	Missing	0
Percentiles	25	3,75
	50	8,50
	75	16,00

- Ώρες χρήσης διαδικτύου:

0, 7, 12, 5, 33, 14, 8, 0, 9, 22

$$L_p = (n+1)\frac{p}{100}$$

Διατάσουμε τις τιμές: 0, 0, **5**, 7, 8, 9, 12, 14, 22, 33

$$L_{25} = (10+1)\frac{25}{100} = 2.75$$

Το 25-οστό εκατοστημόριο Q_1 βρίσκεται στα ¾ (= 0.75) της απόστασης ανάμεσα στη 2η και την **3η** τιμή

Άρα: $Q_1 = 0 + ¾\ (\mathbf{5}-0) = 3.75$

Ερμηνεία: το 25% των παρατηρήσεων είναι μικρότερες από 3.75 και το 75% είναι μεγαλύτερες από 3.75

$$L_p = (n+1)\frac{p}{100}$$

Εκατοστημόρια – Παράδειγμα (2/3)

- Διατάσουμε τις τιμές: 0, 0, 5, 7, **8**, **9**, 12, 14, 22, 33

$$L_{50} = (10+1)\frac{50}{100} = 5.5$$

Το 50-οστό εκατοστημόριο Q_2 βρίσκεται στο 0.5 (=1/2) της απόστασης ανάμεσα στη **5η** και την **6η** τιμή

Άρα: $Q_2 = 8 + ½\ (9-8) = 8.5$

Statistics

hours

N	Valid	10
	Missing	0
Percentiles	25	3,75
	50	8,50
	75	16,00

Ερμηνεία: το 50% των παρατηρήσεων είναι μικρότερες από 8.5 και το 50% είναι μεγαλύτερες από 8.5

$$L_p = (n+1)\frac{p}{100}$$

Εκατοστημόρια – Παράδειγμα (3/3)

- Διατάσουμε τις τιμές: 0, 0, 5, 7, 8, 9, 12, **14**, **22**, 33

$$L_{75} = (10+1)\frac{75}{100} = 8.25$$

Το 75-οστό εκατοστημόριο Q_3 βρίσκεται στο 0.25 (=1/4) της απόστασης ανάμεσα στη **8η** και την **9η** τιμή

Άρα: $Q_3 = 14 + ¼\ (22-14) = 16$

Statistics

hours

N	Valid	10
	Missing	0
Percentiles	25	3,75
	50	8,50
	75	16,00

Ερμηνεία: το 75% των παρατηρήσεων είναι μικρότερες από 16 και το 25% είναι μεγαλύτερες από 16

Εκατοστημόρια - *Περίπτωση κατανομών συχνοτήτων*

- Τα εκατοστιαία σημεία για δεδομένα που είναι ομαδοποιημένα σε μια κατανομή συχνοτήτων υπολογίζονται προσεγγιστικά όπως και η διάμεσος.
- Ο τύπος γενικεύεται για το p-οστό εκατοστιαίο σημείο ως εξής:

$$X_p = L + \frac{c}{f_i}\,(p\,n - F_{i-1})$$

όπου $L =$ το κάτω όριο της τάξης που περιέχει το p-οστό εκατοστιαίο σημείο

$c =$ το εύρος της τάξης

$f_i =$ η συχνότητα του

$n =$ το πλήθος των δεδομένων

$F_{i-1} =$ η αθροιστική συχνότητα της προηγούμενης τάξης

Άσκηση

Ο παρακάτω πίνακας δίνει την κατανομή συχνότητας των μισθών 30 υπαλλήλων μιας δημόσιας υπηρεσίας.

Μισθός σε ευρώ	Αριθμός υπαλλήλων
600-700	7
700-800	14
800-900	5
900-1000	3
1000-1100	1

Από την ανωτέρω κατανομή να υπολογιστούν προσεγγιστικά:

α) το πρώτο τεταρτημόριο Q_1

β) το δεύτερο τεταρτημόριο Q_2 (δηλ. η διάμεσος)

γ) το τρίτο τεταρτημόριο Q_3.

Απάντηση (1/3)

Τάξεις	f_i	F_i
600-700	7	7
700-800	14	21
800-900	5	26
900-1000	3	29
1000-1100	1	30
Άθροισμα	30	

- Αρχικά, στην 3η στήλη του πίνακα υπολογίζουμε τις αθροιστικές συχνότητες $\boldsymbol{F_i}$.

α) Για τον υπολογισμό του Q_1 θα πρέπει να προσδιορίζουμε το ταξικό διάστημα που το περιέχει.

1. Προσδιορίζουμε την τιμή: $p\,n = \frac{25}{100}30 = 7{,}5$
2. Η τιμή 7,5 βρίσκεται ανάμεσα σε διαδοχικούς όρους της αθροιστικής σειράς F_i (εδώ ανάμεσα στο 7 και 21). Ο προηγούμενος όρος, δηλ. ο 7, είναι ο F_{i-1}. $F_{i-1}= 7$
3. Παρατηρούμε ότι ο επόμενος όρος, δηλ. ο 21, ανήκει στο ταξικό διάστημα 700-800, το κατώτατο όριο του οποίου το συμβολίζουμε με L, δηλ. $L = 700$
4. Πηγαίνουμε στην τάξη από την οποία προσδιορίσαμε την τιμή L και παρατηρούμε πόσες συχνότητες έχει. Αυτή είναι η τιμή του f_i, εδώ $f_i = 14$
5. $c = 100$ είναι το πλάτος της τάξης στην οποία ανήκει το L.

$$\boldsymbol{X_p = L + \frac{c}{f_i}(p\,n - F_{i-1})}$$

$$\boldsymbol{X_{25} = 700 + \frac{100}{14}(7{,}5 - 7) = 700 + \frac{50}{14} = 703{,}57}$$

Απάντηση (2/3)

Τάξεις	f_i	F_i
600-700	7	7
700-800	14	21
800-900	5	26
900-1000	3	29
1000-1100	1	30
Άθροισμα	30	

β) Για τον υπολογισμό του Q_2 έχουμε:

1. Προσδιορίζουμε την τιμή: $p\,n = \frac{50}{100}30 = 15$
2. Η τιμή 15 βρίσκεται ανάμεσα σε διαδοχικούς όρους της αθροιστικής σειράς F_i (εδώ ανάμεσα στο 7 και 21). Ο προηγούμενος όρος, δηλ. ο 7, είναι ο F_{i-1}. $F_{i-1}= 7$

$$X_p = L + \frac{c}{f_i}(p\,n - F_{i-1})$$

3. Παρατηρούμε ότι ο επόμενος όρος, δηλ. ο 21, ανήκει στο ταξικό διάστημα 700-800, το κατώτατο όριο του οποίου το συμβολίζουμε με L, δηλ. $L = 700$
4. Πηγαίνουμε στην τάξη από την οποία προσδιορίσαμε την τιμή L και παρατηρούμε πόσες συχνότητες έχει. Αυτή είναι η τιμή του f_i, εδώ $f_i = 14$
5. $c = 100$ είναι το πλάτος της τάξης στην οποία ανήκει το L.

$$X_{50} = 700 + \frac{100}{14}(15 - 7) = 700 + \frac{800}{14} = 757,14$$

Απάντηση (3/3)

Τάξεις	f_i	F_i
600-700	7	7
700-800	14	21
800-900	5	26
900-1000	3	29
1000-1100	1	30
Άθροισμα	30	

β) Για τον υπολογισμό του Q_3 έχουμε:

1. Προσδιορίζουμε την τιμή: $p\,n = \frac{75}{100}30 = 22{,}5$
2. Η τιμή 22,5 βρίσκεται ανάμεσα σε διαδοχικούς όρους της αθροιστικής σειράς F_i (εδώ ανάμεσα στο 21 και 26). Ο προηγούμενος όρος, δηλ. ο 21, είναι ο F_{i-1}. $F_{i-1}= 21$

$$X_p = L + \frac{c}{f_i}(p\,n - F_{i-1})$$

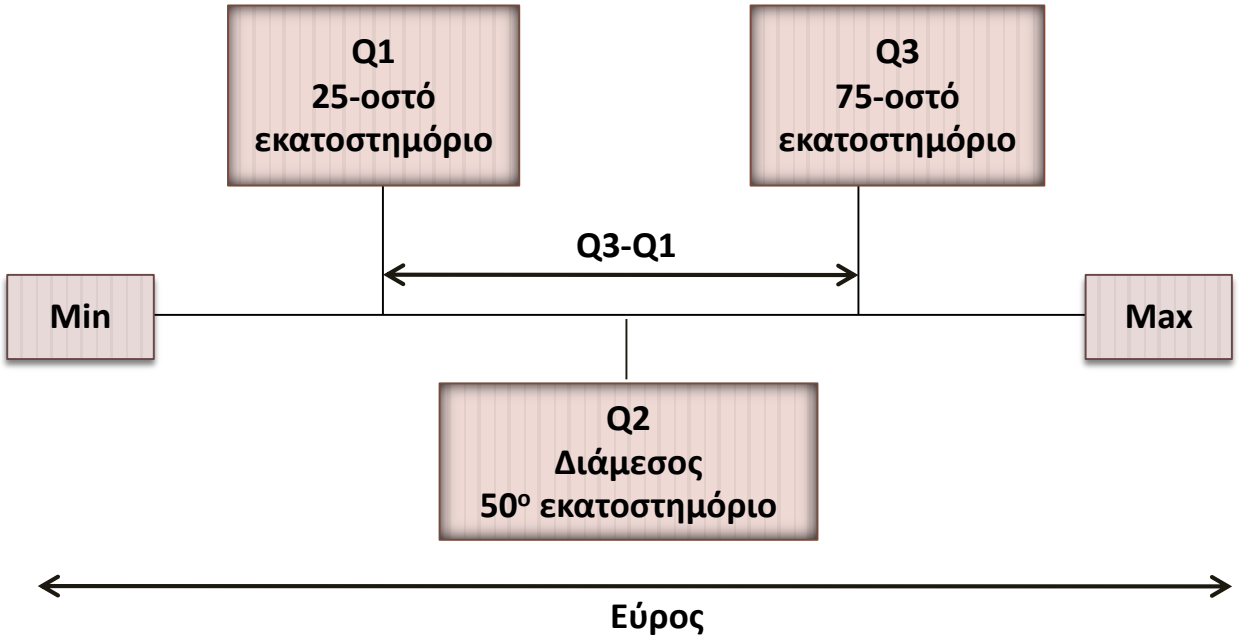
3. Παρατηρούμε ότι ο επόμενος όρος, δηλ. ο 26, ανήκει στο ταξικό διάστημα 800-900, το κατώτατο όριο του οποίου το συμβολίζουμε με L, δηλ. $L = 800$
4. Πηγαίνουμε στην τάξη από την οποία προσδιορίσαμε την τιμή L και παρατηρούμε πόσες συχνότητες έχει. Αυτή είναι η τιμή του f_i, εδώ $f_i = 5$
5. $c = 100$ είναι το πλάτος της τάξης στην οποία ανήκει το L.

$$X_{75} = 800 + \frac{100}{5}(22,5 - 21) = 800 + \frac{150}{5} = 830$$

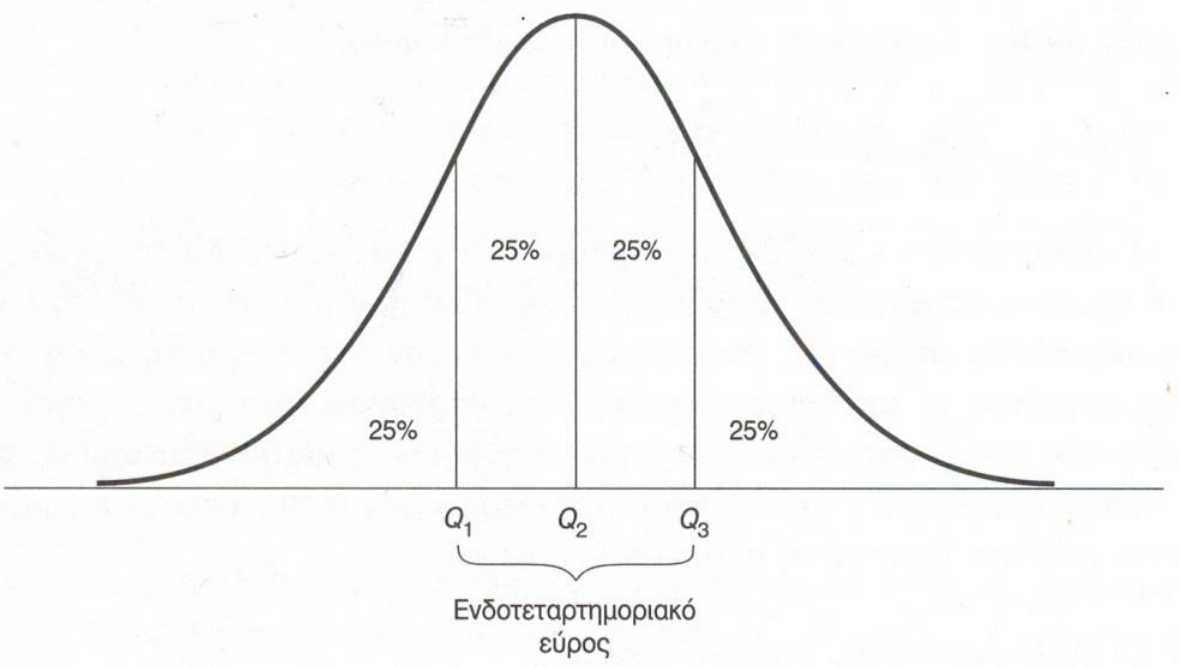
Ενδοτεταρτημοριακό εύρος

Τα τεταρτημόρια βοηθούν στον ορισμό ενός νέου *δείκτη μεταβλητότητας*: **ενδοτεταρτημοριακό εύρος** *(interquartile range IQR):*

$$IQR = Q_3 - Q_1$$



Ενδοτεταρτημοριακό εύρος (2/2)



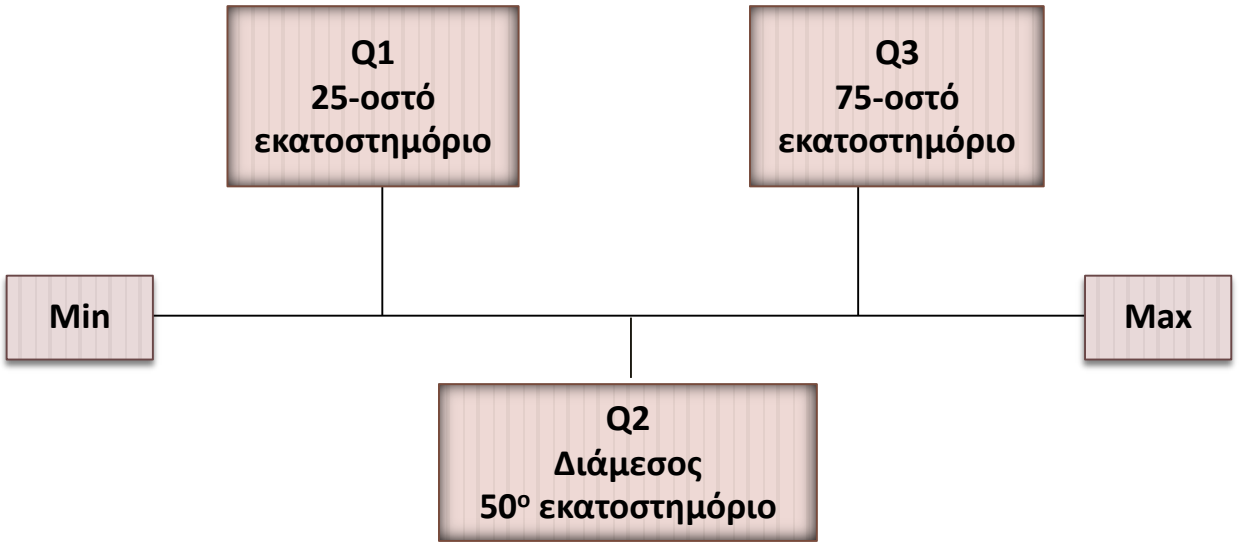
- Στο μεταξύ τους διάστημα περιέχεται το 50% των τιμών του δείγματος
- Μικρό διάστημα → μεγάλη συγκέντρωση τιμών → μικρή διασπορά τιμών

Μεγάλη τιμή του IQR δείχνει μεγάλη μεταβλητότητα

Σύνοψη των 5 αριθμών

Οι 5 αριθμοί αποτελούν τη λεγόμενη σύνοψη των 5 αριθμών *(five numbers summary)* και αποτελούν τη βάση για το θηκόγραμμα *(boxplot).*

1. minimum
2. Q1
3. Q2 (διάμεσος)
4. Q3
5. Maximum



Ακραίες τιμές (outliers)

Ασυνήθιστα **μικρές** ή **μεγάλες τιμές**, απομακρυσμένες από το κύριο σώμα των δεδομένων

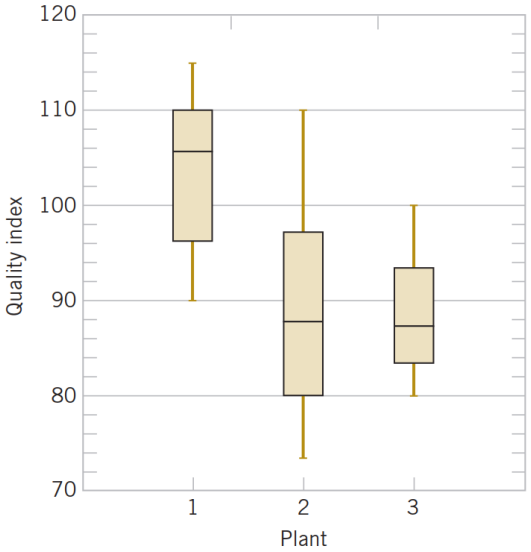
- Ίσως να οφείλονται σε λάθος καταγραφή, ή να κρύβουν χρήσιμες πληροφορίες
- Π.χ. ακραία τιμή (θετική ή αρνητική) στην απόδοση ενός πωλητή μιας επιχείρησης

Το ενδοτεταρτημοριακό εύρος (IQR) δεν επηρεάζεται από πιθανές ακραίες τιμές που μπορεί να υπάρχουν στα δεδομένα.

Θηκογράμματα

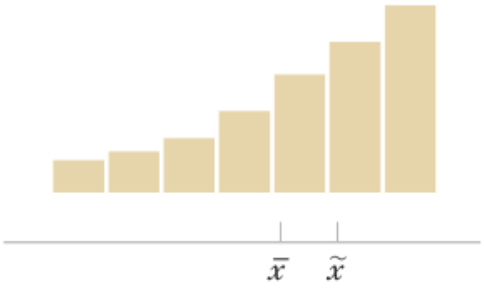
Τα θηκογράμματα (box plots) είναι γραφήματα τα οποία συνοψίζουν **βασικά περιγραφικά μέτρα**, όπως:

- η διάμεσος
- τα τεταρτημόρια
- το ενδοτεταρτημοριακό εύρος
- καθώς και τις ακραίες τιμές

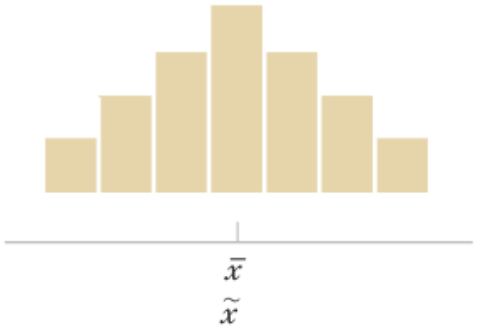


✓ Επίσης, μπορούν να προϊδεάσουν για τη σχηματική μορφή της κατανομής ως προς την **ασυμμετρία** που πιθανώς αυτή εμφανίζει.

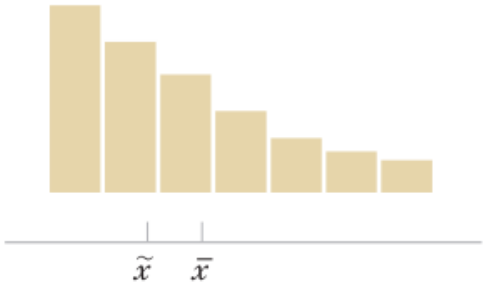
Ασυμμετρία κατανομής δεδομένων



Αρνητική ή
αριστερή λόξευση
(left skew)
(a)

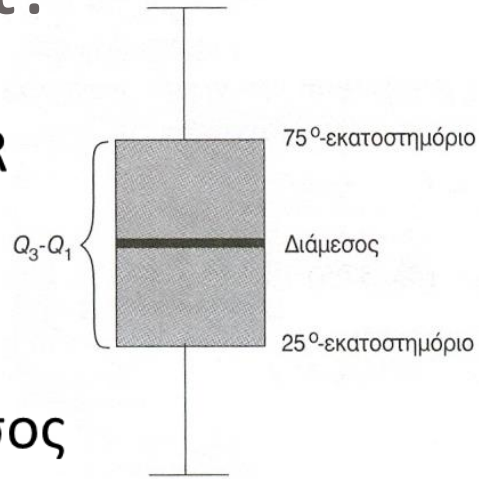


Συμμετρική
(b)



Θετική ή δεξιά
λόξευση (right skew)
(c)

Θηκόγραμμα – από τι αποτελείται?

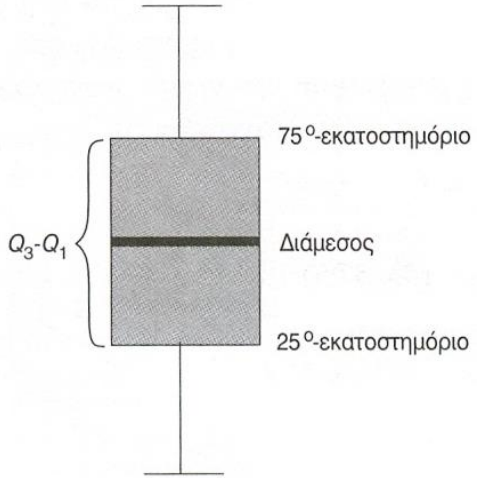


- Από ένα ορθογώνιο παραλληλόγραμο με ύψος IQR
- Κάτω οριζόντια πλευρά → 25ο εκατοστημόριο
- Πάνω οριζόντια πλευρά→ 75ο εκατοστημόριο
- Στο εσωτερικό του μια οριζόντια γραμμή → διάμεσος
- Οριζόντιες γραμμές (**φράκτες**) σε αποστάσεις ίσες το πολύ με $1{,}5(Q_3 - Q_1)$. Αν η μικρότερη ή μεγαλύτερη τιμή βρίσκονται εντός των περιοχών αυτών, τότε οι φράκτες φέρονται ακριβώς στο ύψος των τιμών αυτών.
- Τιμές που βρίσκονται εκτός των φρακτών ονομάζονται ακραία σημεία (outliers).

Θηκόγραμμα – από τι αποτελείται?

- Αν τα ακραία σημεία βρίσκονται σε απόσταση (από την άνω ή κάτω πλευρά) μικρότερη του $3(Q_3 - Q_1)$, δηλ. μεταξύ $1{,}5(Q_3 - Q_1)$ και $3(Q_3 - Q_1)$, συμβολίζονται με έναν μικρό κύκλο (ο)
- Διαφορετικά, συμβολίζονται με έναν αστερίσκο (*)

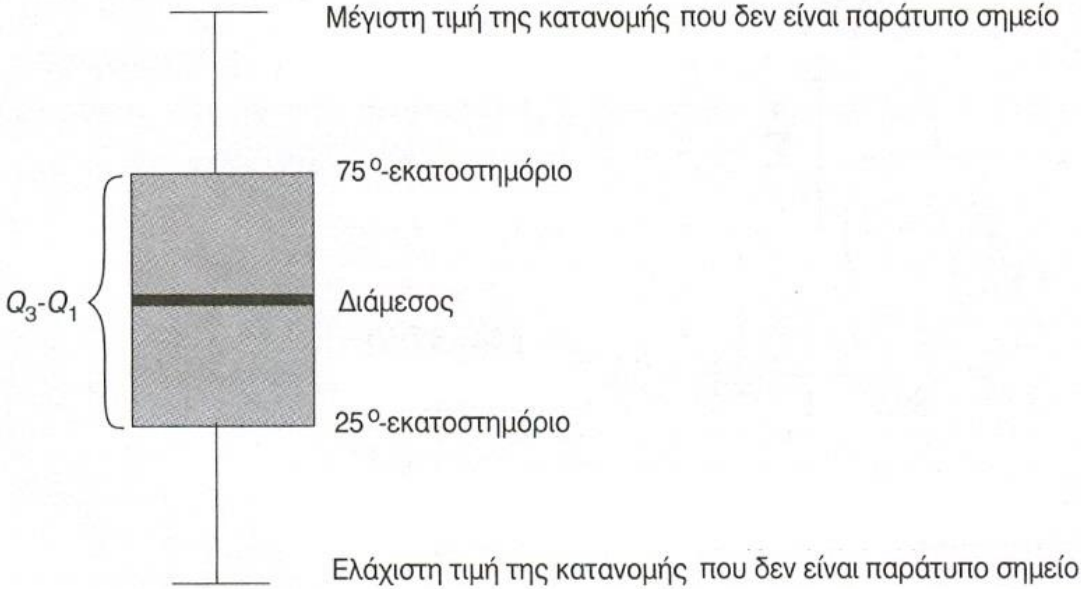
- Εσωτερικός φράκτης → $\pm 1{,}5(Q_3 - Q_1)$
- Εξωτερικός φράκτης → $\pm 3(Q_3 - Q_1)$

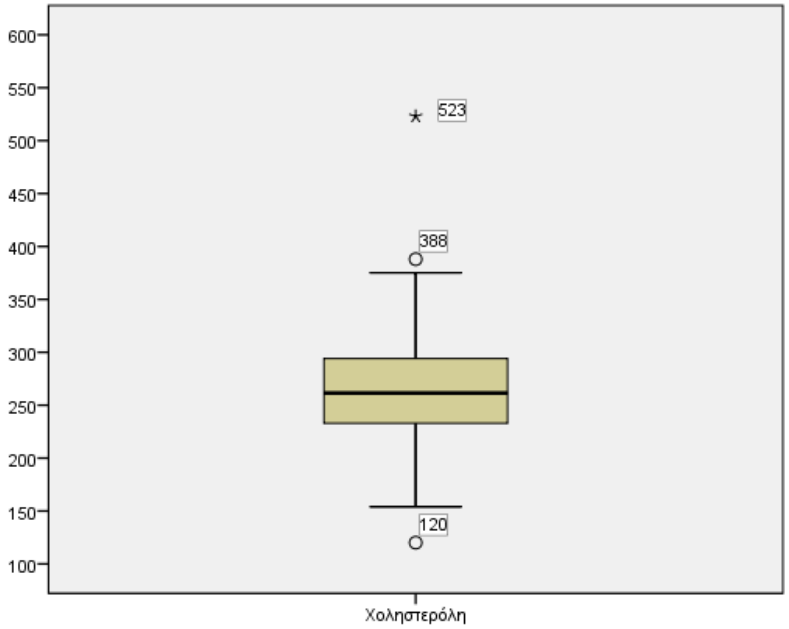


Παράδειγμα boxplot

* Τιμές μεγαλύτερες κατά $3(Q_3\text{-}Q_1)$ τουλάχιστον από το 75ο-εκατοστημόριο

• Τιμές μεγαλύτερες κατά $1{,}5(Q_3\text{-}Q_1)$ τουλάχιστον από το 75ο-εκατοστημόριο





• Τιμές μικρότερες κατά $1{,}5(Q_3\text{-}Q_1)$ τουλάχιστον από το 25ο-εκατοστημόριο

* Τιμές μικρότερες κατά $3(Q_3\text{-}Q_1)$ τουλάχιστον από το 25ο-εκατοστημόριο

Άσκηση

- Δίνεται ο αριθμός x_i των απασχολούμενων σε τυχαίο δείγμα $n = 25$ βιοτεχνιών:

35	12	23	18	5	58	11	14	53	29	61	45	10
6	11	32	92	17	17	44	9	15	12	38	28	

Να κατασκευαστεί το θηκόγραμμα με βάση τη σύνοψη των 5 αριθμών:

$$x_{min} = 5, \mathrm{Q}_1 = 11.5, Q_2 = 18,\ Q_3 = 41,\ x_{max} = 92$$

Statistics

workers

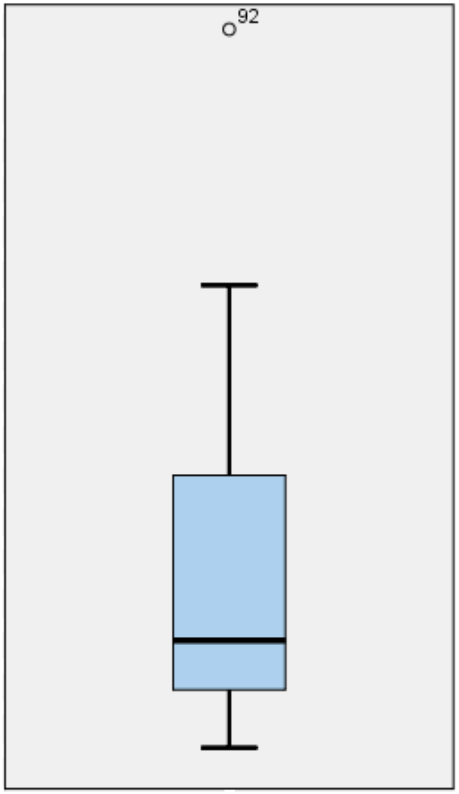
N	Valid	25
	Missing	0
Mean		27,80
Median		18,00
Mode		11[a]
Minimum		5
Maximum		92
Percentiles	25	11,50
	50	18,00
	75	41,00

a. Multiple modes exist. The smallest value is shown

Απάντηση

1. Σχηματίζουμε το ορθογώνιο, κάτω πλευρά στο Q1=11,5, πάνω πλευρά στο Q3=41 και αναπαριστούμε με παχιά γραμμή την διάμεσο (Q2=18).
2. Υπολογίζουμε το: IRQ =Q3-Q1=29.5
3. Συνεπώς: 1,5*IRQ = 44.25 και 3*IRQ=118.
4. **Εσωτερικός φράκτης**→ $\pm 1{,}5(Q_3 - Q_1) = \pm 44.25$ Κάτω όριο = 0 (δεν επιτρέπονται αρνητικές τιμές), Άνω όριο = 41+44.25=85.25

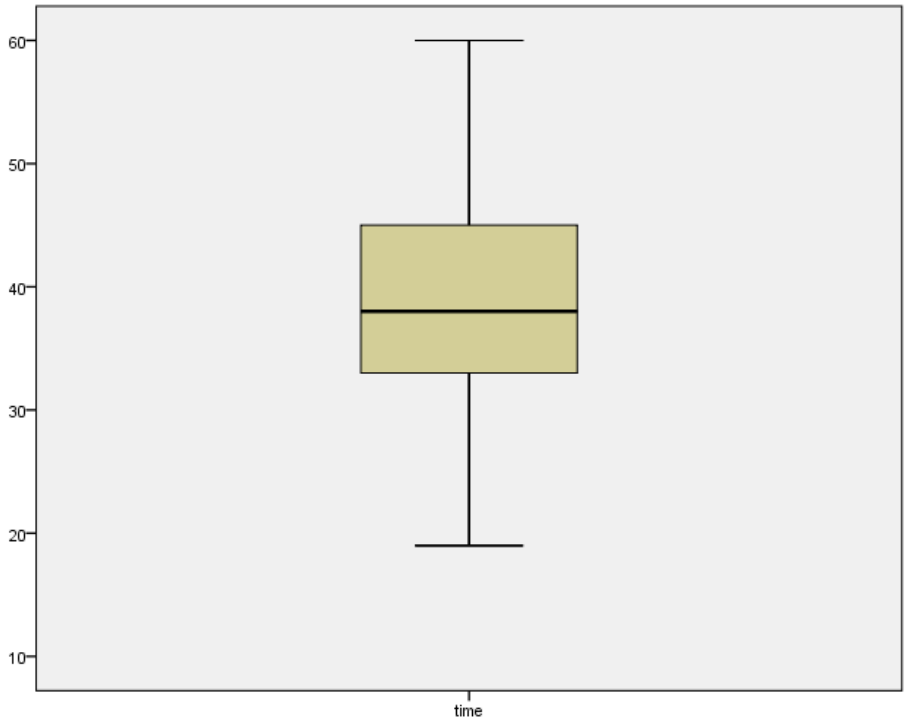
   Επειδή $x_{min} = 5 > 0$ κάτω μύστακας στο 5, ενώ επειδή $x_{max} = 92 > 85.25$ ο άνω μύστακας στο 85.25 και η 17η παρατήρηση με τιμή 92 είναι outlier. Το σημειώνουμε με κύκλο (**o**)
5. **Εξωτερικός φράκτης**→ $\pm 3(Q_3 - Q_1) = \pm 88{,}5$ Κάτω όριο=0 , άνω όριο =41+88,5=129,5. Δεν υπάρχουν παρατηρήσεις μεγαλύτερες από 129,5 για να τις σημειώσουμε με αστερίσκο (*)



$x_{min} = 5$
$Q_1 = 11.5$
$Q_2 = 18$
$Q_3 = 41$
$x_{max} = 92$

Άσκηση

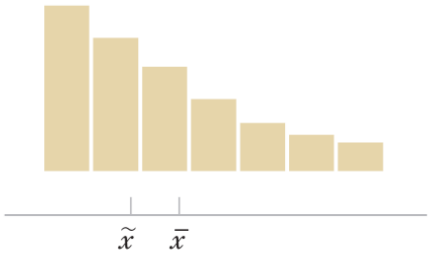
33	29	45	60	42	19	52	38	36



Mean	39,33	
Median	38,00	
Mode	19[a]	
Std. Deviation		12,247
Variance	150,000	

a. Multiple modes exist. The smallest value is shown

Minimum	19	
Maximum	60	
Percentiles	25	31,00
	50	38,00
	75	48,50



Θετική ή δεξιά λόξευση (right skew)

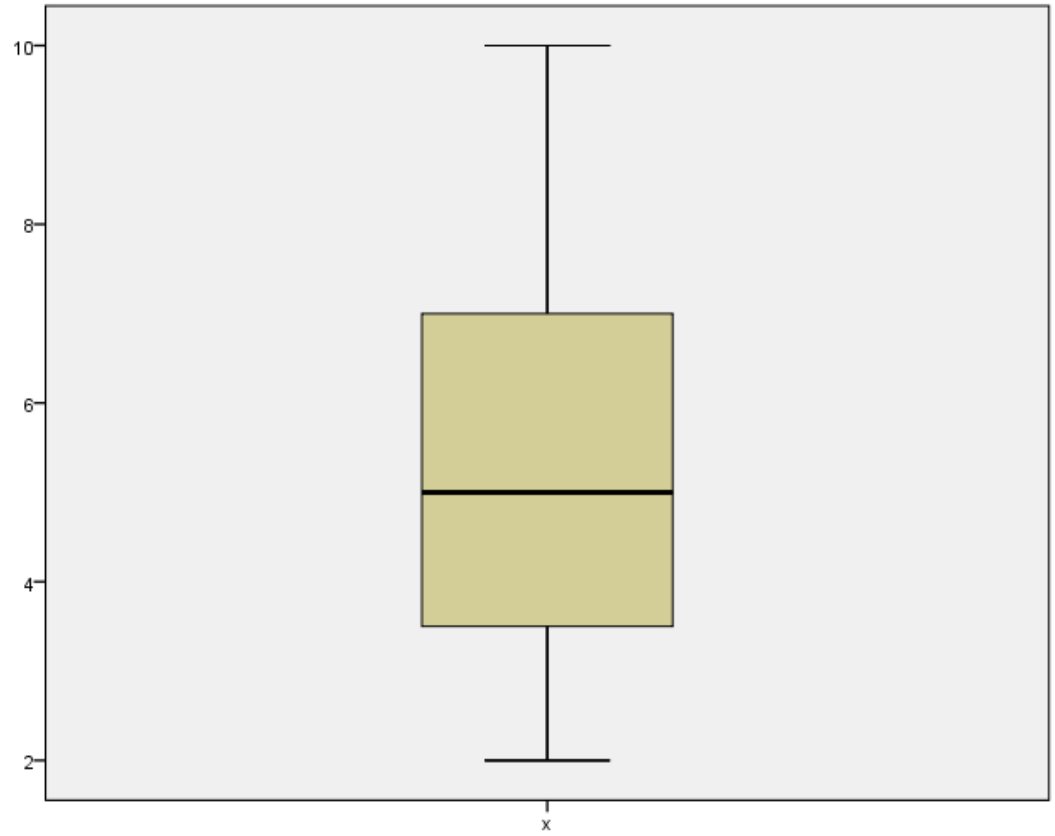
Άσκηση

Για τα ακόλουθα δεδομένα (n=15):

5	8	2	9	5	3	7	4	2	7	4	10	4	3	5

1. Να υπολογίσετε το πρώτο (Q_1), δεύτερο (Q_2) και τρίτο (Q_3) τεταρτημόριο.
2. Να υπολογίσετε το διατεραρτημοριακό εύρος ($Q_3 - Q_1$)
3. Να σχεδιάσετε το θηκόγραμμα (boxplot) των ανωτέρω δεδομένων.

Minimum	2	
Maximum	10	
Percentiles	25	3,00
	50	5,00
	75	7,00



Διασπορά (Variance)

- Ο σημαντικότερος δείκτης μεταβλητότητας
- Παίζει κεντρικό ρόλο στην επαγωγική στατιστική

Διασπορά Πληθυσμού: $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Διασπορά Δείγματος: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Συντομευμένη Μέθοδος: $$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$$

Διασπορά

- Δηλώνει πόσο μακριά από τη μέση τιμή απέχουν οι παρατηρήσεις
  - Μέτρο της απόστασης των παρατηρήσεων από το μέσο

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  - Όταν οι τιμές απέχουν πολύ από τη μέση τιμή η διασπορά είναι μεγάλη
  - Όταν οι τιμές δεν διαφέρουν πολύ από τη μέση τιμή, η διασπορά είναι μικρή

Διασπορά

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Παράδειγμα: Πως η δειγματική διασπορά μετρά τη μεταβλητότητα μέσω των αποκλίσεων.

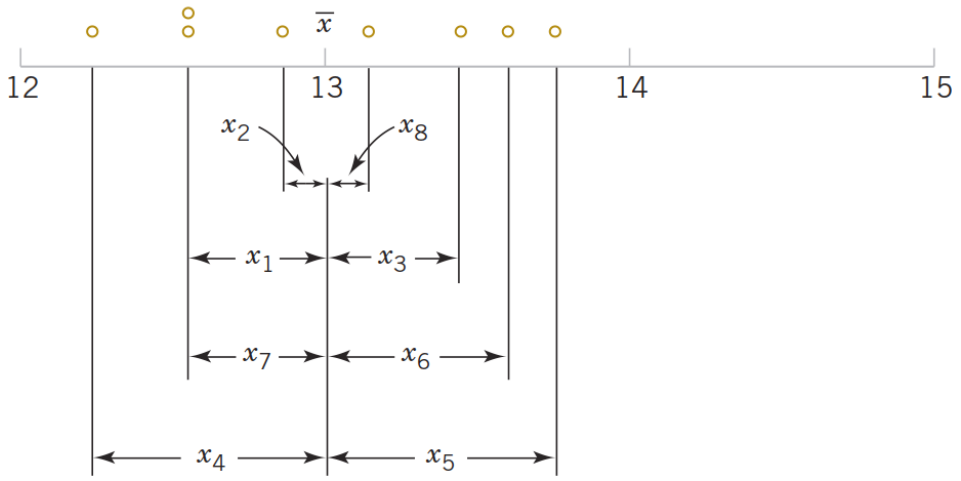
Το <u>άθροισμα των αποκλίσεων ισούται πάντα με μηδέν</u>! Πρέπει να χρησιμοποιήσουμε ένα μέτρο το οποίο θα μετατρέπει τις αρνητικές αποκλίσεις σε μη αρνητικές ποσότητες, υψώνουμε τις αποκλίσεις στο τετράγωνο.

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	12.6	–0.4	0.16
2	12.9	–0.1	0.01
3	13.4	0.4	0.16
4	12.3	–0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	–0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

$$\sum_{i=1}^{8}(x_i - \bar{x})^2 = 1.60$$

$$s^2 = \frac{1.60}{8-1} = \frac{1.60}{7} = 0.2286$$

$$s = \sqrt{0.2286} = 0.48$$



Βήματα που ακολουθούμε:

- Υπολογίζουμε τον αριθμητικό μέσο $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{\mathrm{n}}$
- Υπολογίζουμε την **απόκλιση** *(deviation)* κάθε τιμής που είναι η διαφορά της τιμής x_i από τον αριθμητικό μέσο $\bar{x}$
- Οι αποκλίσεις υψώνονται στο τετράγωνο και αθροίζονται
- Τέλος το άθροισμα των τετραγώνων διαιρείται δια (n-1)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Διασπορά - Παράδειγμα

- Το πλήθος των αιτήσεων για θερινή εργασία ενός δείγματος $n = 6$ φοιτητών: 17, 15, 23, 7, 9, 13

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots (13-14)^2\right] = 33.2$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

Διασπορά: Ερμηνεία

- Φανερώνει πόσο **απομακρυσμένες** είναι οι τιμές από τον αριθμητικό μέσο

- Έχει αξία όταν συγκρίνουμε μεταξύ τους **δύο** διαφορετικά σύνολα δεδομένων:
  - Αν η διασπορά του πρώτου συνόλου είναι μικρότερη από τη διασπορά 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 $(\text{αιτήσεις})^2$

Τυπική Απόκλιση (standard deviation)

- Τυπική Απόκλιση Πληθυσμού: $= \sqrt{\sigma^2}$
- Τυπική Απόκλιση Δείγματος: $= \sqrt{s^2}$

Παράδειγμα: $s = \sqrt{33,2} = 5,76$ αιτήσεις

Ερμηνεία: Αποτελεί δείκτη αξιοπιστίας. Γνωρίζοντας την τυπική απόκλιση και τον αριθμητικό μέσο μπορούμε να εξάγουμε χρήσιμα συμπεράσματα που εξαρτώνται επίσης από το ιστόγραμμα.

Α. Τυποποιημένες τιμές (Z-score)

- Οι z-τιμές είναι ένα ακόμα μέτρο σχετικής θέσης των τιμών των παρατηρήσεων.

- Ως τυποποιημένη τιμή μιας παρατήρησης ορίζεται η απόσταση της από τον αριθμητικό μέσο του συνόλου των παρατηρήσεων στο οποίο ανήκει, εκφρασμένη σε μονάδες τυπικής απόκλισης.

- Υπολογίζονται ως: $z_i = \frac{x_i - \bar{x}}{s}$

Τυποποιημένες τιμές (Z-score)

Παράδειγμα: Η επίδοση ενός τυχαίου δείγματος **100** φοιτητών στα **Μαθηματικά** που αποτελείται από δύο μέρη βαθμολογείται στην κλίμακα 0-100:

- Μέρος Ι: «Διαφορικές εξισώσεις» 0-100
- Μέρος ΙΙ: «Στατιστική» 0-100

Κατά την τελική εξέταση του Ιουνίου είχαμε τα εξής αποτελέσματα:

	Μέρος Ι: Διαφορικές	**Μέρος ΙΙ: Στατιστική**
$\bar{x}$	50	70
s	4	5

- Αν υποθέσουμε ότι ένας φοιτητής βαθμολογήθηκε με **55** στις «Διαφορικές» και **76** στην «Στατιστική», να εξεταστεί σε ποιο Μέρος του μαθήματος είχε καλύτερη επίδοση σε σχέση με το σύνολο των συμφοιτητών του.

Τυποποιημένες τιμές (Z-score) - **Παράδειγμα**

	Μέρος Ι: Διαφορικές	**Μέρος ΙΙ: Στατιστική**
$\bar{x}$	50	70
s	4	5
	55	**76**

- Με βάση τα δεδομένα ο φοιτητής ισχυρίζεται ότι είναι καλύτερος, σε σχέση με τους συμφοιτητές του, στην Στατιστική! Ισχύει;;;;
- Τα δύο σύνολα παρατηρήσεων παρουσιαζουν διαφορές (ως προς την τυπική απόκλιση) κατά συνέπεια δεν μπορούν να συγκριθούν ως έχουν.
  - $CV_{διαφορ} = \frac{s}{\bar{x}} = \frac{4}{50} = 0.08$ (8%)
  - $CV_{Στατιστικ} = \frac{s}{\bar{x}} = \frac{5}{70} = 0.0714$ (7.14%)
- Η σωστή απάντηση μπορεί να δοθεί μόνο μετά από **σύγκριση των τυποποιημένων τιμών** των δύο βαθμολογιών του φοιτητή (**55** & **76**).
  - $z_{διαφορικών} = \frac{55-\bar{x}}{s} = \frac{55-50}{4} = \frac{5}{4} = 1.25$
  - $z_{Στατιστική} = \frac{76-\bar{x}}{s} = \frac{76-70}{5} = \frac{6}{5} = 1.20$
- Αφού $z_{διαφορικών} > z_{Στατιστική}$, συμπεραίνεται ότι ο συγκεκριμένος φοιτητής είναι συγκριτικά **καλύτερος** στις Διαφορικές και όχι στη Στατιστική

Β. Θεώρημα Chebysheff

Δοθέντος ενός συνόλου τιμών και ενός αριθμού k>1, τότε τουλάχιστον το

$(1 - 1/k^2)$

των παρατηρήσεων βρίσκεται k τυπικές αποκλίσεις εκατέρωθεν του μέσου, δηλ. Στο διάστημα $(\bar{x} - k * s,\ \bar{x} + k * s)$

Κάνοντας εφαρμογή για k=2,3,4 προκύπτουν τα χρήσιμα συμπεράσματα:

- Τουλάχιστον το **75%** των παρατηρήσεων βρίσκεται **δύο τυπικές αποκλίσεις** εκατέρωθεν του μέσου: $(\bar{x} - 2s,\ \bar{x} + 2s)$
- Τουλάχιστον το **89%** των παρατηρήσεων βρίσκεται **τρεις τυπικές αποκλίσεις** εκατέρωθεν του μέσου: $(\bar{x} - 3s,\ \bar{x} + 3s)$
- Τουλάχιστον το **94%** των παρατηρήσεων βρίσκεται **τέσσερις τυπικές αποκλίσεις** εκατέρωθεν του μέσου: $(\bar{x} - 4s,\ \bar{x} + 4s)$

  - Εφαρμόζεται σε ιστογράμματα **οποιουδήποτε** σχήματος
  - Παρέχει μόνο **κάτω φράγματα** για τα ποσοστά (ο εμπειρικός κανόνας δίνει συγκεκριμένα ποσοστά)

Β. Θεώρημα Chebysheff

Άσκηση:

Με βάση τα δεδομένα του προηγ. Παραδείγματος να υπολογιστεί το **ποσοστό** των φοιτητών που έχουν:

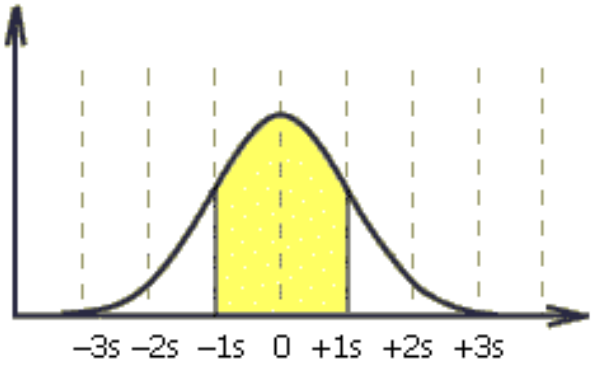
1) Επίδοση στη Στατιστική μεταξύ 60 και 80 (Απ. 75%)
2) Επίδοση στις Διαφορικές μεταξύ 38 και 62 (Απ. 88.9%)

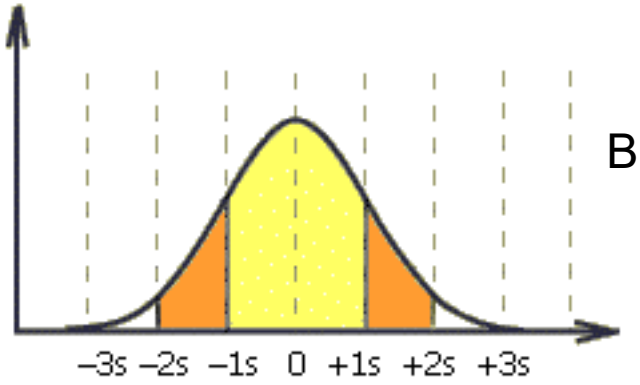
	Μέρος Ι: Διαφορικές	Μέρος ΙΙ: Στατιστική
$\bar{x}$	50	70
s	4	5

Γ. Εμπειρικός Κανόνας

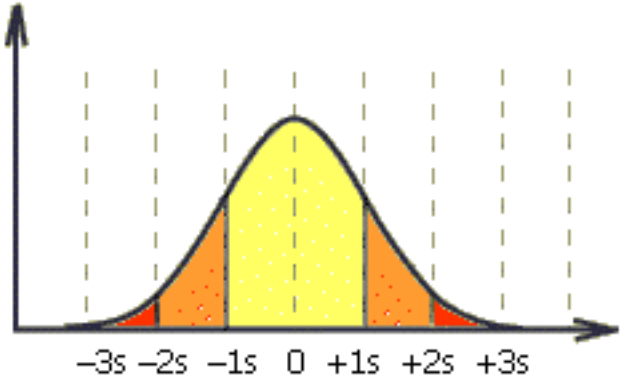
(αν το ιστόγραμμα έχει σχήμα καμπάνας, μόνο για συμμετρικές κατανομές)

Α) Περίπου το **68%** των παρατηρήσεων βρίσκονται σε απόσταση **μιας** τυπικής απόκλισης από τον αριθμητικό μέσο





Β) Περίπου το **95%** των παρατηρήσεων βρίσκονται σε απόσταση **δύο** τυπικών αποκλίσεων από τον αριθμητικό μέσο



Γ) Περίπου το **99,7%** των παρατηρήσεων βρίσκονται σε απόσταση **τριών** τυπικών αποκλίσεων από τον αριθμητικό μέσο

Εμπειρικός Κανόνας: παράδειγμα

Π.χ. Από την ανάλυση των αποδόσεων μιας επιχείρησης προκύπτει ότι το ιστόγραμμα έχει **σχήμα καμπάνας**, ο αριθμητικός μέσος είναι **10%** και η τυπική απόκλιση **8%**.

Σύμφωνα με τον εμπειρικό κανόνα:

- Περίπου το **68%** των αποδόσεων είναι
  - **μεταξύ 2% και 18 %**
- Περίπου το **95%** των αποδόσεων είναι
  - **μεταξύ -6% και 26%**
- Περίπου το **99,7%** των αποδόσεων είναι
  - **μεταξύ -14% και 34%**

Συντελεστής Μεταβλητότητας, CV

- Εκτιμά τη σχέση της τυπικής απόκλισης με το μέγεθος των δεδομένων

Πληθυσμός: $CV = \frac{\sigma}{\mu}$

Δείγμα: $\mathbf{cv} = \frac{s}{\bar{x}}$

Παράδειγμα: $\text{cv} = \frac{s}{\bar{x}} = \frac{5{,}76}{14} = 0{,}41$

Ερμηνεία: όσο πιο μικρή είναι η τιμή του CV, τόσο πιο μικρή είναι η μεταβλητότητα των παρατηρήσεων

Συντελεστής Μεταβλητότητας, CV

Παράδειγμα: Να συγκριθεί η μεταβλητότητα των βαθμολογιων στις Διαφορικές και στην Στατιστική των 100 φοιτητών του τμήματος

	Μέρος Ι: Διαφορικές	**Μέρος ΙΙ: Στατιστική**
$\bar{x}$	50	70
s	4	5

1) Το συμπέρασμα ότι η μεταβλητότητα στην Στατιστική (λόγω του s=5) είναι μεγαλύτερη είναι λανθασμένο! Τα σύνολα έχουν διαφορετικούς μέσους.

$$CV_{διαφορ} = \frac{s}{\bar{x}} = \frac{4}{50} = 0.08 \quad (8\%)$$

$$CV_{Στατιστικ} = \frac{s}{\bar{x}} = \frac{5}{70} = 0.0714 \quad (7.14\%)$$

2) Η σχετική ανομοιογένεια (μεταβλητότητα) των επιδόσεων των φοιτητών στη Στατιστική (7.14%) είναι **μικρότερη** από εκείνη των επιδόσεων στις Διαφορικές εξισώσεις (8%)

Ομαδοποιημένα Δεδομένα: Δείκτες

Δεν γνωρίζουμε **άμεσα** τα δεδομένα αλλά την **κατανομή** τους. Μπορούμε να προσεγγίσουμε τον αριθμητικό μέσο και την διασπορά

- Αριθμητικός μέσος: $$\bar{x} \approx \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

- Διασπορά: $$s^2 \approx \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right]$$

Παράδειγμα Υπολογισμού Δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i
1	0 … 15	71
2	15 … 30	37
3	30 … 45	13
4	45 … 60	9
5	60 … 75	10
6	75 … 90	18
7	90 … 105	28
8	105 … 120	14

Παράδειγμα Υπολογισμού Δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i	Κεντρική Τιμή m_i
1	0 ... 15	71	7.5
2	15 ... 30	37	22.5
3	30 ... 45	13	37.5
4	45 ... 60	9	52.5
5	60 ... 75	10	67.5
6	75 ... 90	18	82.5
7	90 ... 105	28	97.5
8	105 ... 120	14	112.5

Παράδειγμα Υπολογισμού Δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i	Κεντρική Τιμή m_i	$f_i m_i$	$f_i m_i^2$
1	0 … 15	71	7.5	532.5	3,993.75
2	15 … 30	37	22.5	832.5	18,731.25
3	30 … 45	13	37.5	487.5	18,281.25
4	45 … 60	9	52.5	472.5	24,806.25
5	60 … 75	10	67.5	675	45,562.5
6	75 … 90	18	82.5	1485	122,512.5
7	90 … 105	28	97.5	2730	266,175
8	105 … 120	14	112.5	1575	177.187,5

Παράδειγμα Υπολογισμού Δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i	Κεντρική Τιμή m_i	$f_i m_i$	$f_i m_i^2$
1	0 ... 15	71	7.5	532.5	3,993.75
2	15 ... 30	37	22.5	832.5	18,731.25
3	30 ... 45	13	37.5	487.5	18,281.25
4	45 ... 60	9	52.5	472.5	24,806.25
5	60 ... 75	10	67.5	675	45,562.5
6	75 ... 90	18	82.5	1485	122,512.5
7	90 ... 105	28	97.5	2730	266,175
8	105 ... 120	14	112.5	1575	177.187,5
Σύνολα:		**200**		**8790**	**677,250**

Παράδειγμα Υπολογισμού Δεικτών

- Προσεγγιστικές τιμές:

$$\bar{x} \approx \frac{\sum_{i=1}^{k} f_i m_i}{n} = \frac{8790}{200} = 43.95$$

$$s^2 \approx \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right] = \frac{1}{200-1}\left[677{,}250 - \frac{(8790)^2}{200}\right] = 1461.96$$

Πραγματικές τιμές: $\bar{x} = 43{,}59,\ s^2 = 1518{,}64$

- Πολύ καλή προσέγγιση του **αριθμητικού μέσου**
- Όχι καλή προσέγγιση της **διασποράς**

Άσκηση: Υπολογισμός δεικτών

Να υπολογιστεί ο **αριθμητικός μέσος**, η **διασπορά**, η **τυπική απόκλιση** και ο **συντελεστής μεταβλητότητας** των παρακάτω ομαδοποιημένων δεδομένων:

1.

Κλάση	Συχνότητα
0 ... 16	50
16 ... 32	160
32 ... 48	110
48 ... 64	80

2.

Κλάση	Συχνότητα
0 ... 200	68
200 ... 400	73
400 ... 600	101
600 ... 800	89

Άσκηση 1: Υπολογισμός δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i	Κεντρική Τιμή m_i	$f_i m_i$	m_i^2	$f_i m_i^2$
1	0 … 16	50	8	400	64	3200
2	16 … 32	160	24	3840	576	92160
3	32 … 48	110	40	4400	1600	176000
4	48 … 64	80	56	4480	3136	250880
Σύνολα:		**400**		**13120**		**522240**

- **αριθμητικός μέσος:** $\bar{x} \approx \frac{\sum_{i=1}^{k} f_i\, m_i}{n} = \frac{13120}{400} = 32.8$
- **διασπορά:** $s^2 \approx \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i\, m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i\, m_i\right)^2}{n}\right] = \frac{1}{400-1}\left[522240 - \frac{13120^2}{400}\right] = 230.34$
- **τυπική απόκλιση:** $s = \sqrt{s^2} = \sqrt{230.34} \approx 15.18$
- **συντελεστής μεταβλητότητας:** $CV = \frac{s}{\bar{x}} = \frac{15.18}{32.8} = 0.46$

Άσκηση 2: Υπολογισμός δεικτών

Κλάση	Όρια Κλάσης	Συχνότητα f_i	Κεντρική Τιμή m_i	$f_i m_i$	m_i^2	$f_i m_i^2$
1	0 ... 200	68	100	6800	10000	680000
2	200 ... 400	73	300	21900	90000	6570000
3	400 ... 600	101	500	50500	250000	25250000
4	600 ... 800	89	700	62300	490000	43610000
Σύνολα:		331		141500		76110000

- **αριθμητικός μέσος:** $\bar{x} \approx \frac{\sum_{i=1}^{k} f_i \, m_i}{n} = \frac{141500}{331} = 427.49$
- **διασπορά:** $s^2 \approx \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i \, m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i \, m_i\right)^2}{n}\right] = \frac{1}{331-1}\left[76110000 - \frac{141500^2}{331}\right] = 47332.78$
- **τυπική απόκλιση:** $s = \sqrt{s^2} = \sqrt{47332.78} \approx 217.56$
- **συντελεστής μεταβλητότητας:** $CV = \frac{s}{\bar{x}} = \frac{217.56}{427.49} = 0.50$