

# ΣΤΟΙΧΕΙΑ ΣΤΑΤΙΣΤΙΚΗΣ

- **Εισαγωγή - Introduction (σελίδες 2-9)**
  - Περιγραφική Στατιστική (Descriptive Statistics)
  - Δειγματοληψία (Sampling)
  - Πιθανότητες (Probability Theory)
  - Επαγωγική Στατιστική (Inferential Statistics)
- **Περιγραφική Στατιστική – Descriptive Statistics (σελίδες 10-39)**
  - Πληθυσμός, Δείγμα (Population, Sample)
  - Μεταβλητές, Δεδομένα (Variables, Data)
  - Κατηγορίες δεδομένων (Types of Data)
  - Πίνακες/Διαγράμματα παρουσίασης Δεδομένων (Tables, Graphs)
  - Χαρακτηριστικά δεδομένων (Cross Sectional, Time Series, Bivariate variable)
- **Αριθμητικοί δείκτες – Indices (σελίδες 40-48)**
  - Μέσος (Mean)
  - Διάμεσος (Median)
  - Επικρατούσα τιμή (Mode)
- **Μέτρα Διασποράς – Measures of Dispersion (σελίδες 49-101)**
  - Μεταβλητότητα (Variability)
  - Εύρος (Range)
  - Εκατοστημόρια (Percentiles)
  - Ενδοτεταρτημοριακό Εύρος (IQR)
  - Θηκόγραμμα (Boxplot)
  - Διακύμανση (Variance)
  - Τυπική Απόκλιση (Standard Deviation)
  - Αξιοσημείωτε Εφαρμογές (Z-Score, Chebyshev's Theorem, Empirical Rule)
  - Συντελεστής Μεταβλητότητας (Coefficient of Variation - CV)

# Εισαγωγή

# Τι είναι Στατιστική;

Η επιστήμη που ασχολείται:

- με τη συλλογή,
- την ταξινόμηση,
- την παρουσίαση,
- την επεξεργασία,
- την ανάλυση και
- την ερμηνεία αριθμητικών δεδομένων,  
με στόχο την εξαγωγή συμπερασμάτων κατάλληλων για  
την **ορθή λήψη αποφάσεων**.

Παρέχει σύνολο τεχνικών και μεθόδων με σκοπό

- Το σχεδιασμό της διαδικασίας της συλλογής και της οργάνωσης των δεδομένων
- Τη συνοπτική και αποτελεσματική παρουσίαση τους
- Την ανάλυση τους και την εξαγωγή συμπερασμάτων

# Περιγραφική Στατιστική (Descriptive Statistics)

- Ασχολείται με τις μεθόδους οργάνωσης, σύνοψης και παρουσίασης των δεδομένων με τρόπο εύχρηστο και κατανοητό
- Τεχνικές:
  - **Γραφήματα:** Οπτικοποιούν τα δεδομένα διευκολύνοντας την αναγνώριση χρήσιμων πληροφορικών
  - **Αριθμητικοί Δείκτες:** αντιπροσωπεύουν ιδιότητες των δεδομένων

# Δειγματοληψία (Sampling)

Παρέχει τεχνικές για την συλλογή δεδομένων από έναν πεπερασμένο πληθυσμό (επιλογή «καλού» δείγματος)

- Επιλογή Δείγματος
- Σχεδιασμός ερωτηματολογίου
- Σφάλματα δειγματοληψίας
- Έλεγχος Ποιότητας Δείγματος
- Αμεροληψία, Αξιοπιστία, Εγκυρότητα

## Πιθανότητες (Probability Theory)

- Συνδέουν τον συνολικό πληθυσμό με το δείγμα
- Είναι απαραίτητες για την επαγωγική στατιστική
- Παίζουν σημαντικό ρόλο στη λήψη αποφάσεων

# Επαγωγική Στατιστική (Inferential Statistics)

- Παρέχει τεχνικές που επιτρέπουν την προβολή δεικτών από ένα μικρό δείγμα σε έναν ευρύτερο πληθυσμό (εκτιμητική)
  - Επαγωγή: εκτίμηση ή πρόβλεψη μιας παραμέτρου του πληθυσμού
- Απότερος στόχος: η λήψη μιας επιχειρηματικής απόφασης

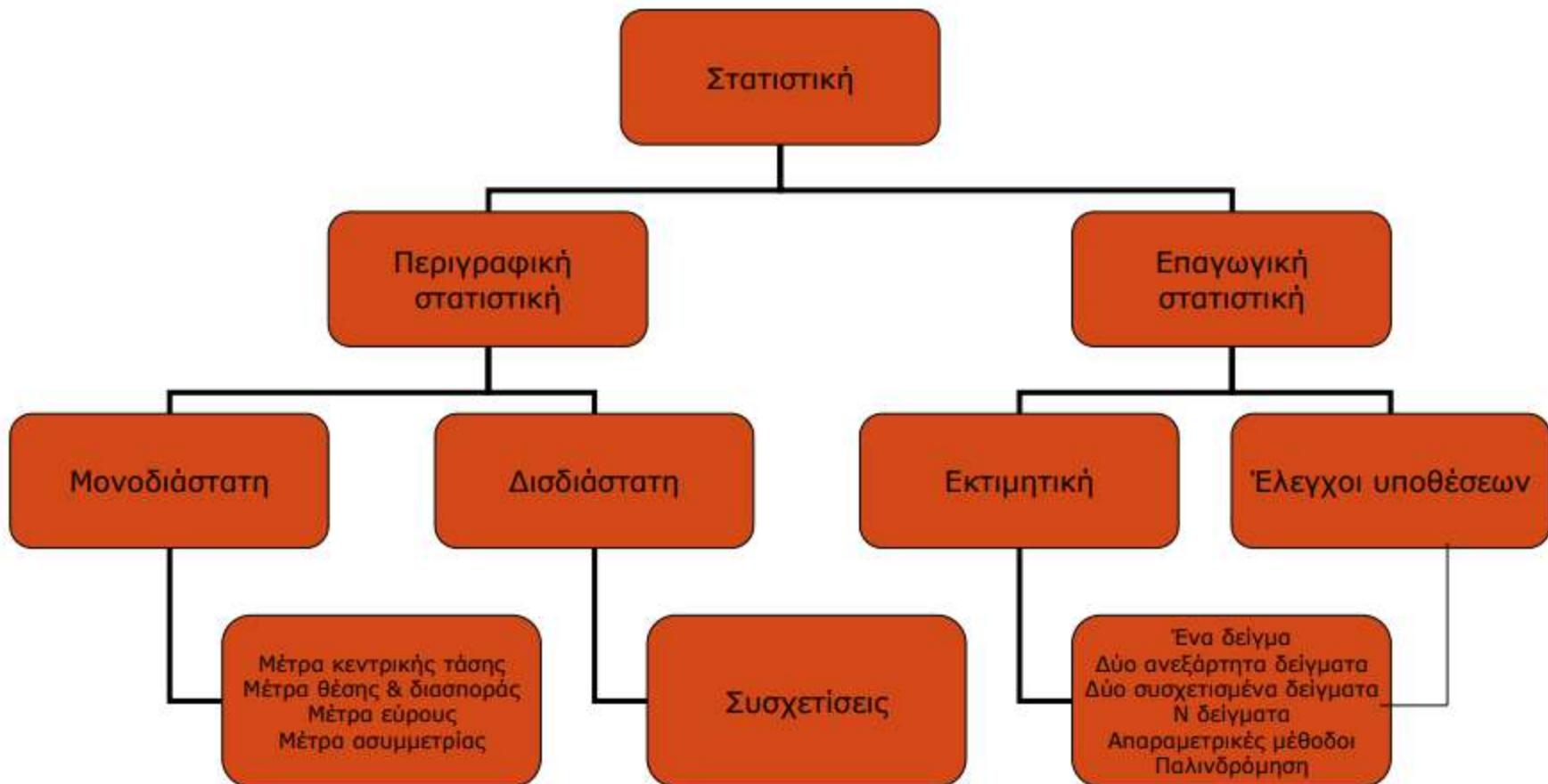
Αναφορά σε:

- Εκτιμητές, Έλεγχο Υποθέσεων, Ανάλυση Διασποράς, Παλινδρόμηση, Συσχέτιση

# Βήματα Στατιστικής Ανάλυσης

- Δειγματοληψία: επιλογή του δείγματος από τον πληθυσμό βάσει καταλληλότητας αυτού σε σχέση με το υπό εξέταση χαρακτηριστικό
- Χρήση Η/Υ και προγραμμάτων (π.χ. excel, SPSS, Minitab, Stata κτλ) για την καταχώρηση και οργάνωση των δεδομένων
- Παρουσίαση και περιγραφή δεδομένων (περιγραφική στατιστική)
- Συμπερασματολογία (επαγωγική στατιστική)

# Σχεδιάγραμμα



# Περιγραφική Στατιστική

## (Descriptive Statistics)

# Περιγραφική Στατιστική (Descriptive Statistics)

- Ασχολείται με τις μεθόδους οργάνωσης, σύνοψης και παρουσίασης των δεδομένων με τρόπο εύχρηστο και κατανοητό
- Τεχνικές:
  - **Γραφήματα:** Οπτικοποιούν τα δεδομένα
  - **Αριθμητικοί Δείκτες:** αντιπροσωπεύουν ιδιότητες των δεδομένων
- Οι τεχνικές της επιτρέπουν την εξαγωγή συμπερασμάτων και τη λήψη αποφάσεων

# Πληθυσμός, Δείγμα

- **Πληθυσμός:** το σύνολο όλων των παρατηρήσεων των οποίων κάποιο χαρακτηριστικό ή κάποια ιδιότητα θέλουμε να μελετήσουμε.
- **Στατιστικές μονάδες (άτομα):** τα στοιχεία του πληθυσμού
- **Απογραφή:** η διαδικασία καταγραφής των παρατηρήσεων ενός πληθυσμού
- **Δείγμα:** ένα υποσύνολο (τμήμα) του πληθυσμού
- **Δειγματοληψία:** διαδικασία επιλογής και επεξεργασίας δείγματος

# Μεταβλητές, Δεδομένα

- *Μεταβλητή*: ένα χαρακτηριστικό του πληθυσμού ή του δείγματος  
(Υψος, χρώμα ματιών, ομάδα αίματος, επάγγελμα, ...)
- *Τιμή μεταβλητής*: κάθε παρατηρήσιμη κατάσταση  
(1.74, καστανό, Ο+, σπουδαστής, ...)
- *Δεδομένα*: το σύνολο των τιμών μιας μεταβλητής  
(1.63, 1.84, 1.69, 1.72, ...)

# Κατηγορίες δεδομένων

- **Ποσοτικά** (Αριθμητικά): οι τιμές είναι αριθμητικές και επιδέχονται μέτρηση. Π.χ. εισόδημα, το βάρος, το ύψος, ο αριθμός των παιδιών μιας οικογένειας.
  - **Διακριτές** οι οποίες παίρνουν μόνο «μεμονωμένες» αριθμητικές τιμές, π.χ. το νούμερο των υποδημάτων, ο αριθμός των παιδιών μιας οικογένειας, ο αριθμός των ελαττωμάτων ενός προϊόντος.
  - **Συνεχείς** οι οποίες μπορούν να πάρουν οποιαδήποτε τιμή μέσα από ένα συνεχές διάστημα, όπως π.χ. το βάρος, το ύψος, η διάρκεια μιας τηλεφωνικής συνδιάλεξης.
- **Ποιοτικά:** αναφέρονται σε κάποιο ποιοτικό χαρακτηριστικό και οι τιμές τους δεν είναι αριθμητικές. Π.χ. επίπεδο εκπαίδευσης, μητρική γλώσσα, βιοτικό επίπεδο, κ.τ.λ
  - **Ονομαστικά (Nominal data):** π.χ. φύλο, χρώμα ματιών, επάγγελμα, θρησκεία, οικογενειακή κατάσταση), επιδέχονται οι οποίες επιδέχονται μόνο αυθαίρετη κατάταξη
  - **Διατάξιμα (Ordinal data):** επιδέχονται μέτρηση ανωτέρου επιπέδου που επιτρέπει την ιεράρχησή τους, όπως π.χ. χαρακτηρισμός πτυχίου (άριστα, λίαν καλώς, καλώς), σοβαρότητα μιας ασθένειας (ήπια, μέτρια, σοβαρή), της γνώμης για κάποιο μέτρο (διαφωνώ πλήρως, διαφωνώ σε κάποια σημεία, συμφωνώ, συμφωνώ πλήρως).

Κωδικοποίηση ποιοτικών δεδομένων μέσω αρίθμησης (δες παράδειγμα)

# Παράδειγμα 1.

Τα αποτελέσματα των εξετάσεων των φοιτητών στο μάθημα της Στατιστικής ήταν: 2, 3, 3, 4, 4, 5, 7, 5, 5, 9.

Να βρεθεί:

- i) Ποιος είναι ο πληθυσμός;
- ii) Ποια είναι τα άτομα;
- iii) Ποιες είναι οι παρατηρήσεις;
- iv) Ποια είναι η μεταβλητή και σε ποια κατηγορία ανήκει;
- v) Ποιες είναι οι τιμές των μεταβλητών;

## Απάντηση:

- i) Ο πληθυσμός είναι οι 10 φοιτητές του τμήματος.
- ii) Κάθε φοιτητής είναι ένα άτομο.
- iii) Οι παρατηρήσεις είναι: 2, 3, 3, 4, 4, 5, 7, 5, 5, 9.
- iv) Η μεταβλητή είναι «ο βαθμός στη Στατιστική» η οποία είναι ποσοτική, διακριτή και μεταβλητή διαστήματος.
- v) Οι τιμές της μεταβλητής είναι 2, 3, 4, 5, 7, 9.

## Παράδειγμα 2.

Εξετάζουμε τους κατοίκους μιας πόλης ως προς τα παρακάτω χαρακτηριστικά:

- i) Φύλο
- ii) ύψος
- iii) μορφωτικό επίπεδο
- iv) εισόδημα
- v) θρήσκευμα

Να χαρακτηρίσετε τις παραπάνω μεταβλητές.

### Απάντηση:

- i) Το «Φύλο» είναι ποιοτική ονομαστική μεταβλητή.
- ii) Το «ύψος» είναι ποσοτική συνεχής μεταβλητή διαστήματος
- iii) Το «μορφωτικό επίπεδο» είναι ποιοτικά διατάξιμη μεταβλητή.
- iv) Το «εισόδημα» είναι ποσοτική συνεχής αναλογική μεταβλητή.
- v) Το «θρήσκευμα» είναι ποιοτική ονομαστική μεταβλητή.

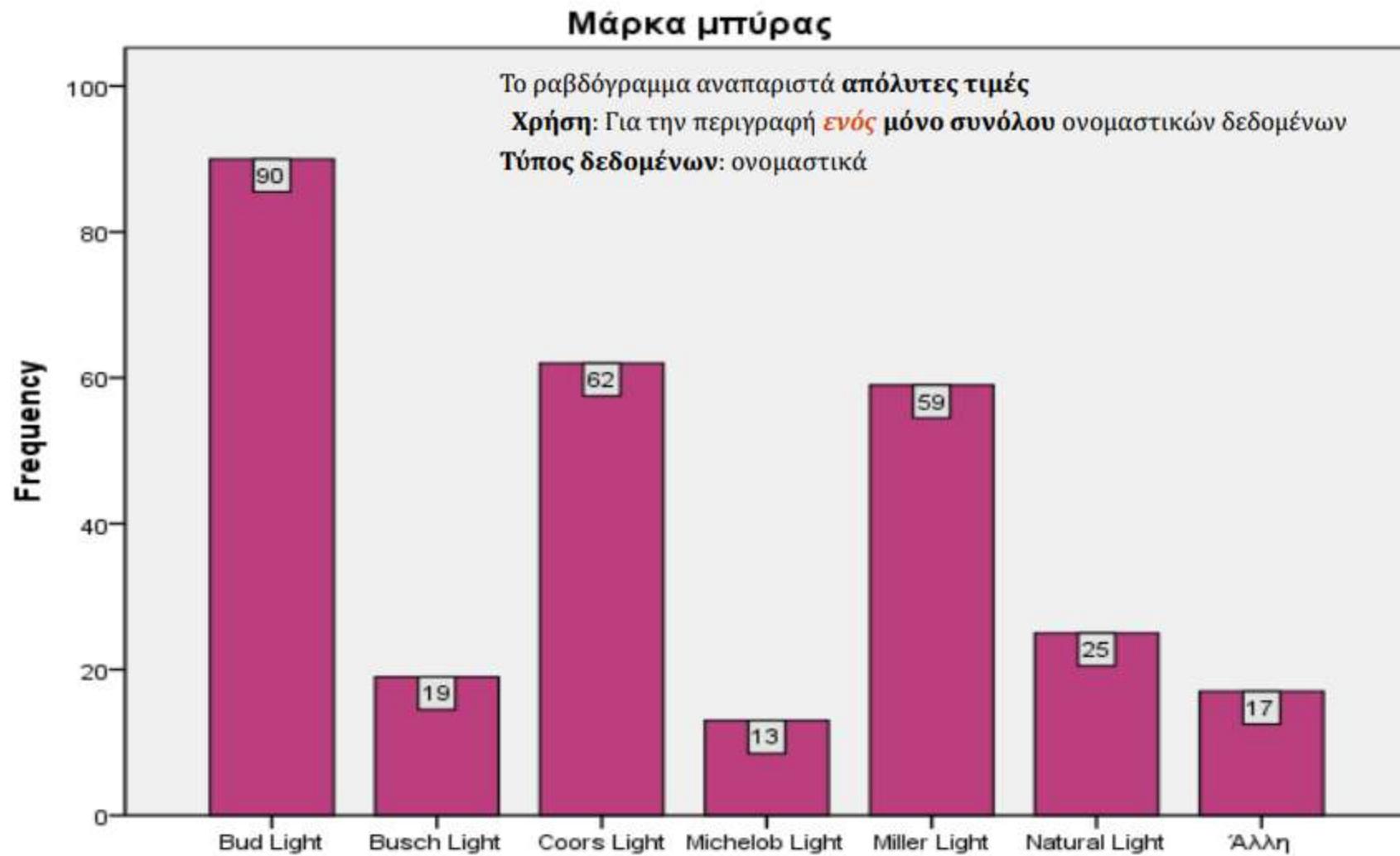
# Παρουσίαση Ονομαστικών Δεδομένων: Πίνακας Συχνοτήτων

Η οπτική εξέταση των δεδομένων δεν προσφέρει πραγματική πληροφορία!

**Διαλογή:** η καταμέτρηση των εμφανίσεων κάθε τιμής.

Αύξων Αριθμός	Μάρκα Μπύρας	Συχνότητα	Σχετική Συχνότητα
1	Bud Light	90	31,6%
2	Busch Light	19	6,7%
3	Coors Light	62	21,8%
4	Michelob Light	13	4,6%
5	Miller Light	59	20,7%
6	Natural Light	25	8,8%
7	Other	17	6,0%

# Παρουσίαση Ονομαστικών Δεδομένων: Ραβδόγραμμα



# Παρουσίαση Ονομαστικών Δεδομένων: Σύνοψη

- Ο πίνακας συχνοτήτων αναπαριστά το **απόλυτο** πλήθος των εμφανίσεων
- Ο πίνακας σχετικών συχνοτήτων αναπαριστά τα **ποσοστά** των εμφανίσεων
- Το ραβδόγραμμα αναπαριστά **απόλυτες τιμές**
- Το κυκλικό διάγραμμα αναπαριστά **ποσοστά**.

❑ **Χρήση:** Για την περιγραφή **ενός** μόνο συνόλου ονομαστικών δεδομένων

❑ **Τύπος δεδομένων:** ονομαστικά

# Παρουσίαση Ποσοτικών Συνεχών Δεδομένων: Πίνακας Δεδομένων

**Κόστος υπεραστικών κλήσεων 200 συνδρομητών**

## Συνήθεις ερωτήσεις:

- Ποια είναι η **κατανομή** των αριθμών μεταξύ του **ελάχιστου** και του **μέγιστου**;
- Ποιο είναι το **τυπικό ποσό** ενός τυχαίου λογαριασμού;
- Σε ποιο βαθμό τα ποσά είναι παρόμοια ή διαφέρουν μεταξύ τους;

## Απαντήσεις:

- Θα πρέπει να κατασκευαστεί ένας πίνακας συχνοτήτων, από το οποίο θα σχεδιαστεί ένα **ιστόγραμμα**. Συνεπώς επιβάλλεται η **ομαδοποίηση**.

# Παρουσίαση Ποσοτικών Συνεχών Δεδομένων: Ομαδοποίηση

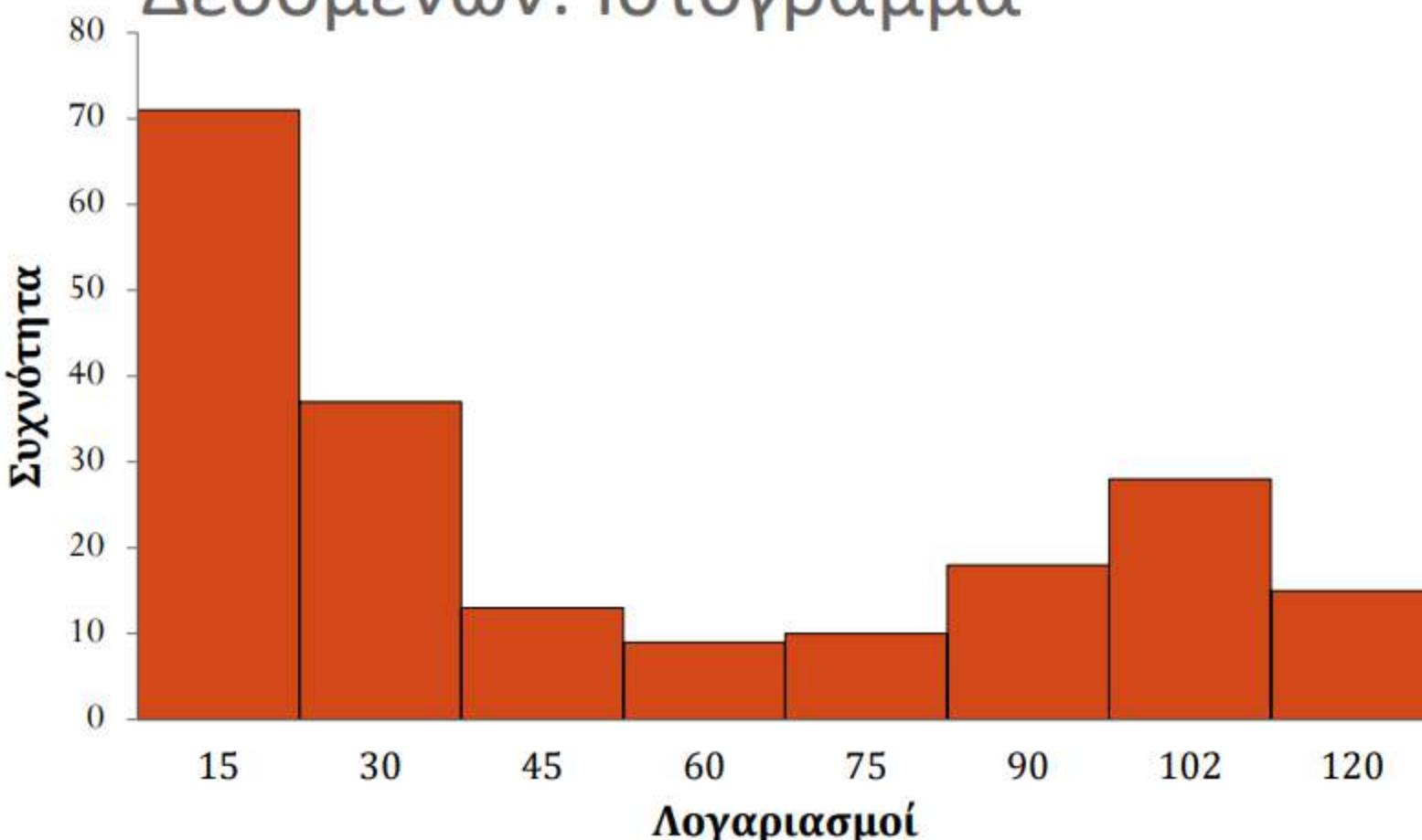
- Ομαδοποιούμε τα δεδομένα σε **κλάσεις**: Χωρίζουμε το συνολικό διάστημα των τιμών σε μια σειρά από διαδοχικές ζώνες (κλάσεις)
- Το πλήθος **k** των κλάσεων εξαρτάται από το πλήθος **n** των δεδομένων
- Τύπος Sturges:  $k = 1 + 3.32 \log_{10}(n)$
- Εύρος δεδομένων **R**: μεγαλύτερη – μικρότερη τιμή
- Πλάτος κάθε κλάσης:  $R / k$

# Παρουσίαση Ποσοτικών Συνεχών Δεδομένων: Πίνακας Κλάσεων

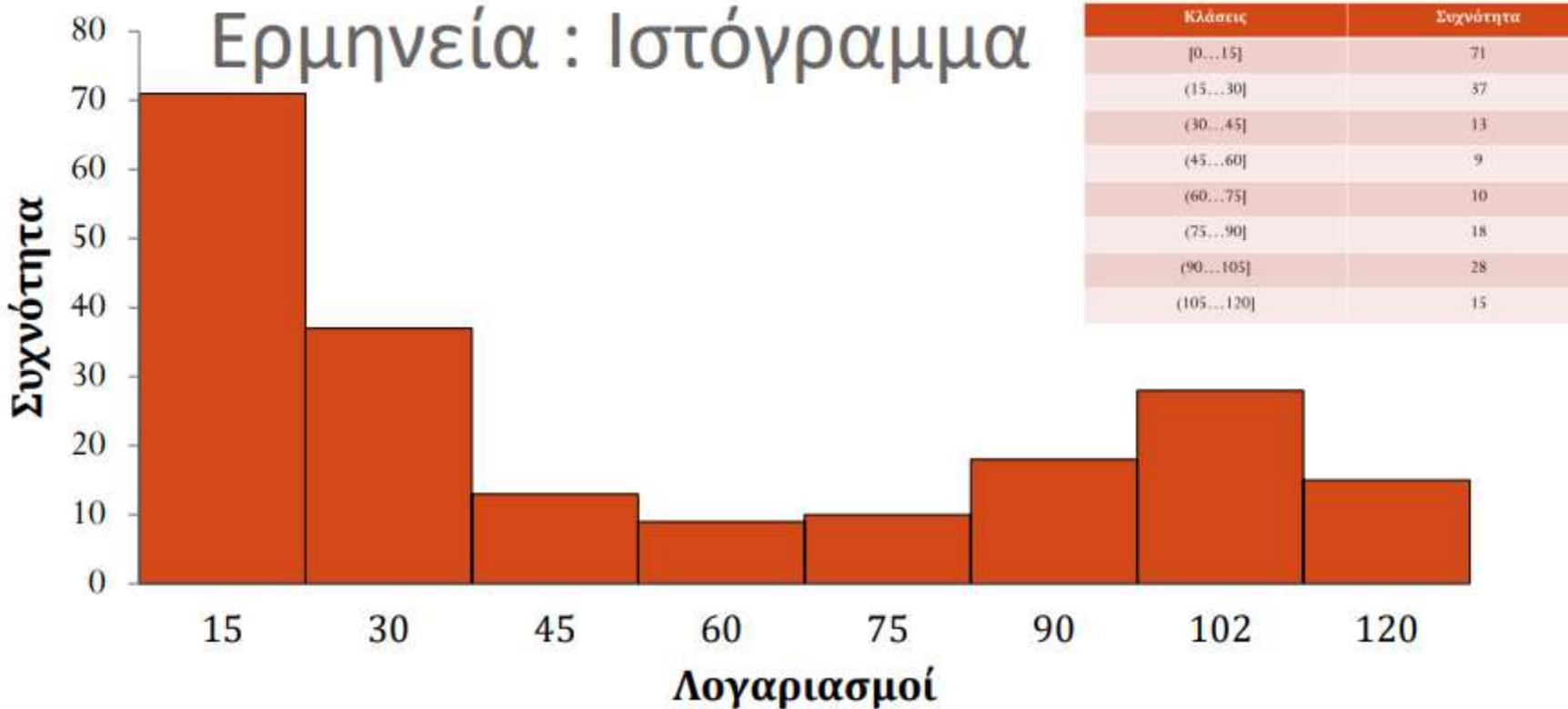
Κλάσεις	Συχνότητα
[0...15]	71
(15...30]	37
(30...45]	13
(45...60]	9
(60...75]	10
(75...90]	18
(90...105]	28
(105...120]	15

- Οκτώ (8) κλάσεις, ίδιου πλάτους (15). Σε κάθε κλάση ανήκουν τα ποσά που είναι μεγαλύτερα από το αριστερό άκρο και μικρότερα ή ίσα από το δεξί άκρο.
- Κάθε ποσό ανήκει σε μια μόνο κλάση (δεν υπάρχει επικάλυψη).
- Ο παραπάνω πίνακας ονομάζεται **κατανομή συχνοτήτων** των λογαριασμών υπεραστικών κλήσεων

# Παρουσίαση Ποσοτικών Συνεχών Δεδομένων: Ιστόγραμμα



- Το γράφημα που απεικονίζει τον πίνακα συχνοτήτων ονομάζεται **ιστόγραμμα**.
- Ο οριζόντιος άξονας αντιπροσωπεύει τις αριθμητικές τιμές των δεδομένων
- Οι ράβδοι έχουν ως βάση το διάστημα τιμών της κλάσης και ως ύψος την συχνότητα της κλάσης.



Το **ιστόγραμμα** δίνει μια σαφή εικόνα της κατανομής των λογαριασμών.

Οι μισοί περίπου (101) αφορούν μικρά ποσά [0-30]

Λίγοι (32) βρίσκονται στις ενδιάμεσες τιμές (30-75]

Αρκετά σημαντικό μέρος (61) των λογαριασμών είναι στο  
ανώτερο διάστημα τιμών (75-120]

# Πλήθος Κλάσεων μιας κατανομής συχνοτήτων

Μέγεθος δείγματος	Κλάσεις
Κάτω από 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1.000	10 - 11
1.000 - 5.000	11- 13
5.000 - 50.000	13 - 17
Πάνω από 50.000	17 - 20

Ο αριθμός των κλάσεων εξαρτάται αποκλειστικά από το πλήθος των δεδομένων  
Στον πίνακα συνδέεται ο αριθμός των κλάσεων με τον αριθμό των δεδομένων  
Εναλλακτικός τρόπος υπολογισμού από τον τύπο του Sturges:

$$k = 1 + 3.3 \log_{10}(n)$$

## Παρουσίαση Ποσοτικών Δεδομένων: Πίνακας Κλάσεων Σχετικών Συχνοτήτων

Κλάσεις	Συχνότητα	Σχετική Συχνότητα
[0...15]	71	35,5
(15...30]	37	18,5
(30...45]	13	6,5
(45...60]	9	4,5
(60...75]	10	5,0
(75...90]	18	9,0
(90...105]	28	14,0
(105...120]	15	7,0

Οι μισοί περίπου αφορούν μικρά ποσά (0-30), ποσοστό **54%**

Λίγοι βρίσκονται στις ενδιάμεσες τιμές (30-75), ποσοστό **16%**

Αρκετά σημαντικό μέρος των λογαριασμών είναι στο ανώτερο διάστημα τιμών (75-120), ποσοστό **30%**

## Αθροιστική (Σχετική) Συχνότητα

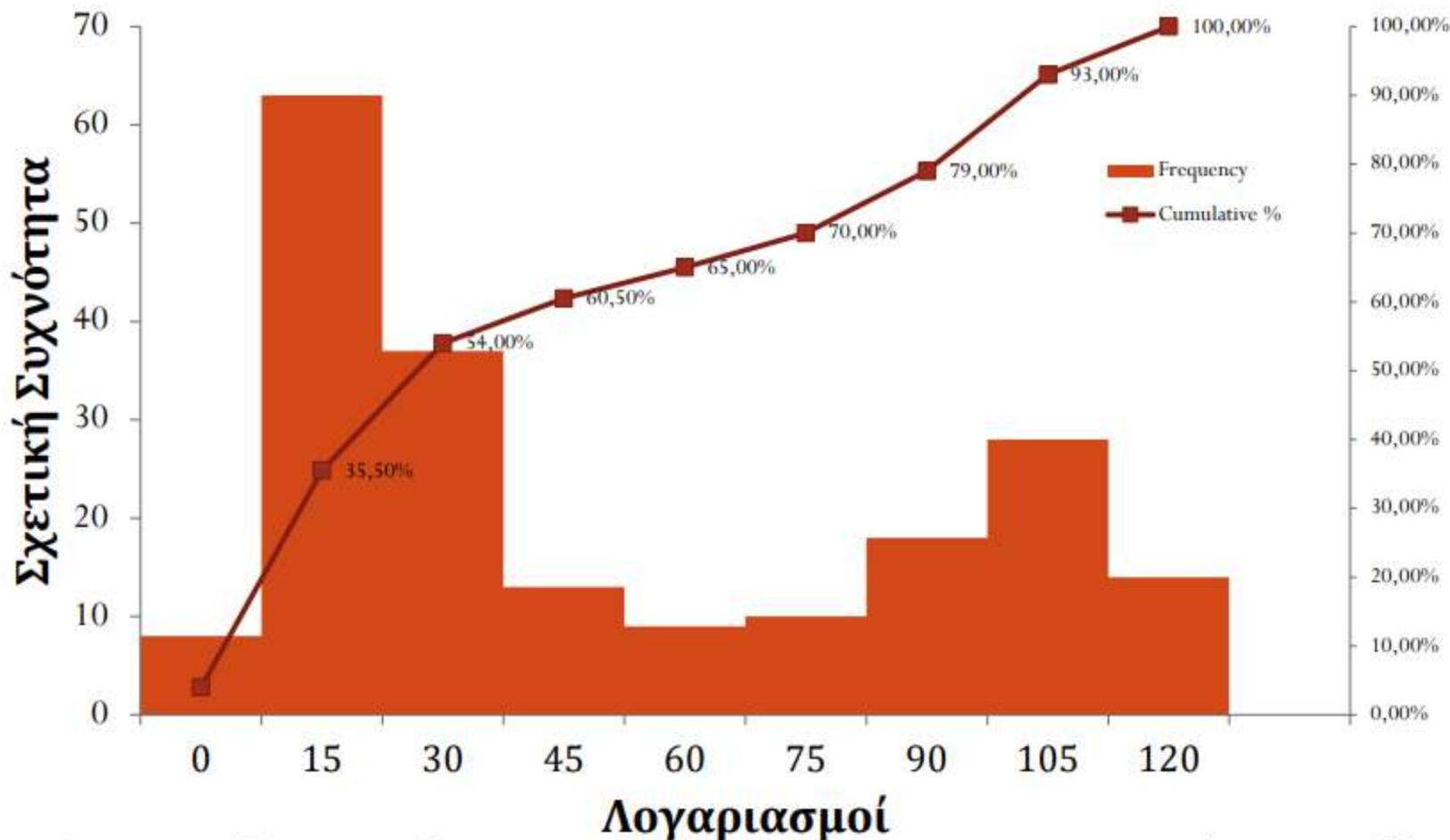
- Φανερώνει την αύξηση (του ποσοστού) των δεδομένων με την προσθήκη μιας νέας κλάσης
- Ορίζεται ως το πλήθος των παρατηρήσεων που η τιμή τους δεν ξεπερνά μια συγκεκριμένη τιμή
- Υπάρχει και η Αφαιρετική Συχνότητα:
  - Ορίζεται ως το πλήθος των παρατηρήσεων που η τιμή τους ξεπερνά μια συγκεκριμένη τιμή

Δείτε το ακόλουθο παράδειγμα (τελευταία στήλη)

# Πίνακας Κλάσεων Αθροιστικών Σχετικών Συχνοτήτων

Κλάσεις	Συχνότητα	Σχετική Συχνότητα	Αθροιστική Σχετική Συχνότητα
[0...15]	71	35,5	35,5
(15...30]	37	18,5	54,0
(30...45]	13	6,5	60,5
(45...60]	9	4,5	65,0
(60...75]	10	5,0	70,0
(75...90]	18	9,0	79,0
(90...105]	28	14,0	93,0
(105...120]	15	7,0	100

# Παρουσίαση Ποσοτικών Δεδομένων: Αθροιστικό Πολυγωνικό Διάγραμμα



- Απεικονίζει τις αθροιστικές συχνότητες με τη μορφή μιας αύξουσας τεθλασμένης γραμμής.

# Χαρακτηριστικά δεδομένων

- Ένα σημαντικό χαρακτηριστικό των δεδομένων, εκτός από τον τύπο (ονομαστικά, διατακτικά ή ποσοτικά) είναι και ο **χρόνος συλλογής**.
- Τα δεδομένα που συλλέγονται κατά προσέγγιση ταυτόχρονα ονομάζονται **διαστρωματικά δεδομένα** (cross-sectional data)
- Ενώ τα δεδομένα που έχουν καταγραφεί σε μια ακολουθία χρονικών σημείων ονομάζονται **χρονολογικές σειρές** (time-series)

# Χρονολογικές Σειρές

- Τα δεδομένα που έχουν καταγραφεί σε μια ακολουθία χρονικών σημείων
- **Παράδειγμα:** η τιμή του πετρελαίου, η τιμή μιας μετοχής, η τιμή μιας κατοικίας, ...
- **Γραμμικό Διάγραμμα (line chart):** παριστάνει την εξέλιξη της τιμής της μεταβλητής σε σχέση με το χρόνο
  - Ο οριζόντιος άξονας αντιπροσωπεύει τον χρόνο
  - Ο κατακόρυφος τις τιμές της μεταβλητής

# Παρουσίαση Χρονολογικών Σειρών

Μέση τιμή Βενζίνης (σεντς ανά γαλόνι) από το 1978 έως το 2006



- $2006 - 1978 = 28$  έτη
- $28$  έτη  $\times 12 = 336$  μήνες

# Διμετάβλητες Τεχνικές (bivariate)

- Παρουσιάζουν τη σχέση που έχουν μεταξύ τους **δύο** μεταβλητές
  - Μονομετάβλητες (univariate): αφορούν μία μόνο μεταβλητή
- **Όνομαστικές Μεταβλητές**
  - Η περιγραφή γίνεται μέσω του **πίνακα διπλής εισόδου**
  - Η οπτικοποίηση γίνεται μέσω μιας παραλλαγής του **ραβδογράμματος**
- **Ποσοτικές Μεταβλητές**
  - Η περιγραφή γίνεται μέσω του **πίνακα δεδομένων**
  - Η οπτικοποίηση γίνεται μέσω **διαγράμματος διασποράς**

# Παρουσίαση Δύο Ονομαστικών Μεταβλητών: Πίνακας Δεδομένων

Παράδειγμα: στατιστική έρευνα για τις εφημερίδες που αγοράζουν συγκεκριμένοι επαγγελματίες

Αναγνώστης	Επάγγελμα	Εφημερίδα
1	2	2
2	1	4
3	2	1
4	3	2
5	1	3
6	3	3
7	2	1
8	1	3

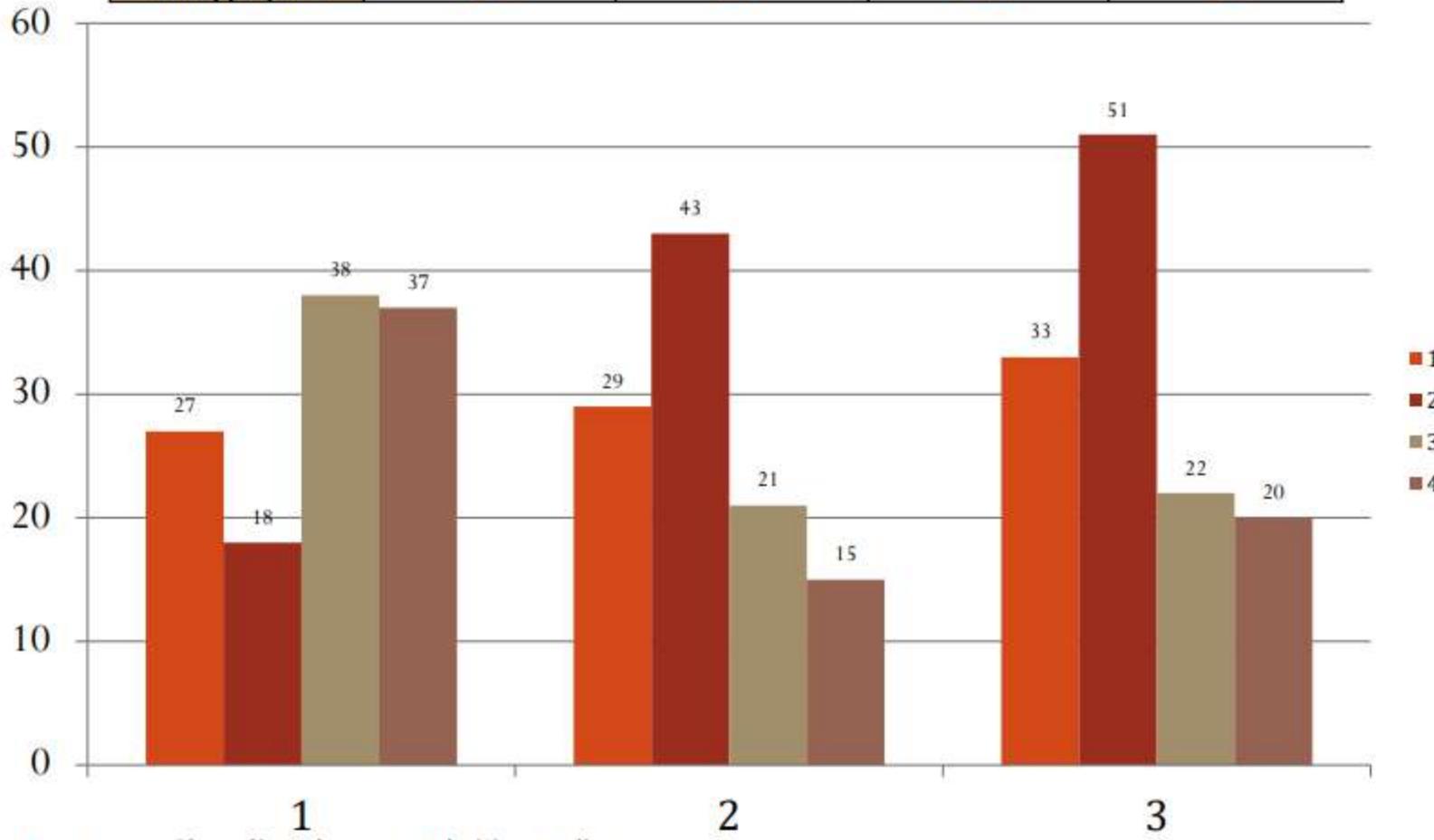
## Παρουσίαση Δύο Ονομαστικών Μεταβλητών: Πίνακας Διπλής Εισόδου (cross-tab)

Επάγγελμα	Εφημερίδα				Σύνολο
	Εφημερίδα 1	Εφημερίδα 2	Εφημερίδα 3	Εφημερίδα 4	
Επάγγελμα 1	27	18	38	37	120
Επάγγελμα 2	29	43	21	15	108
Επάγγελμα 3	33	51	22	20	126
Σύνολο	89	112	81	72	354

- Πόσες μεταβλητές χρειάζομαι για να αναπαραστήσω τον παραπάνω πίνακα σε μια βάση δεδομένων?
- Πάντα τρεις (3)! Ποιες?

# Ράβδογραμμα Δύο Ονομαστικών Μεταβλητών

Επάγγελμα	Εφημερίδα 1	Εφημερίδα 2	Εφημερίδα 3	Εφημερίδα 4
Επάγγελμα 1	27	18	38	37
Επάγγελμα 2	29	43	21	15
Επάγγελμα 3	33	51	22	20



- Στον οριζόντιο άξονα έχουμε τα τρία (3) επαγγέλματα
- Για κάθε επαγγελμα εμφανίζουμε μια συστάδα ράβδων (cluster bar) που το ύψος τους απεικονίζει την συχνότητα
- Κάθε συστάδα αποτελείται από τέσσερις (4) ράβδους - μία για κάθε εφημερίδα

# Παρουσίαση Δύο Ποσοτικών Μεταβλητών:

## Πίνακας Δεδομένων

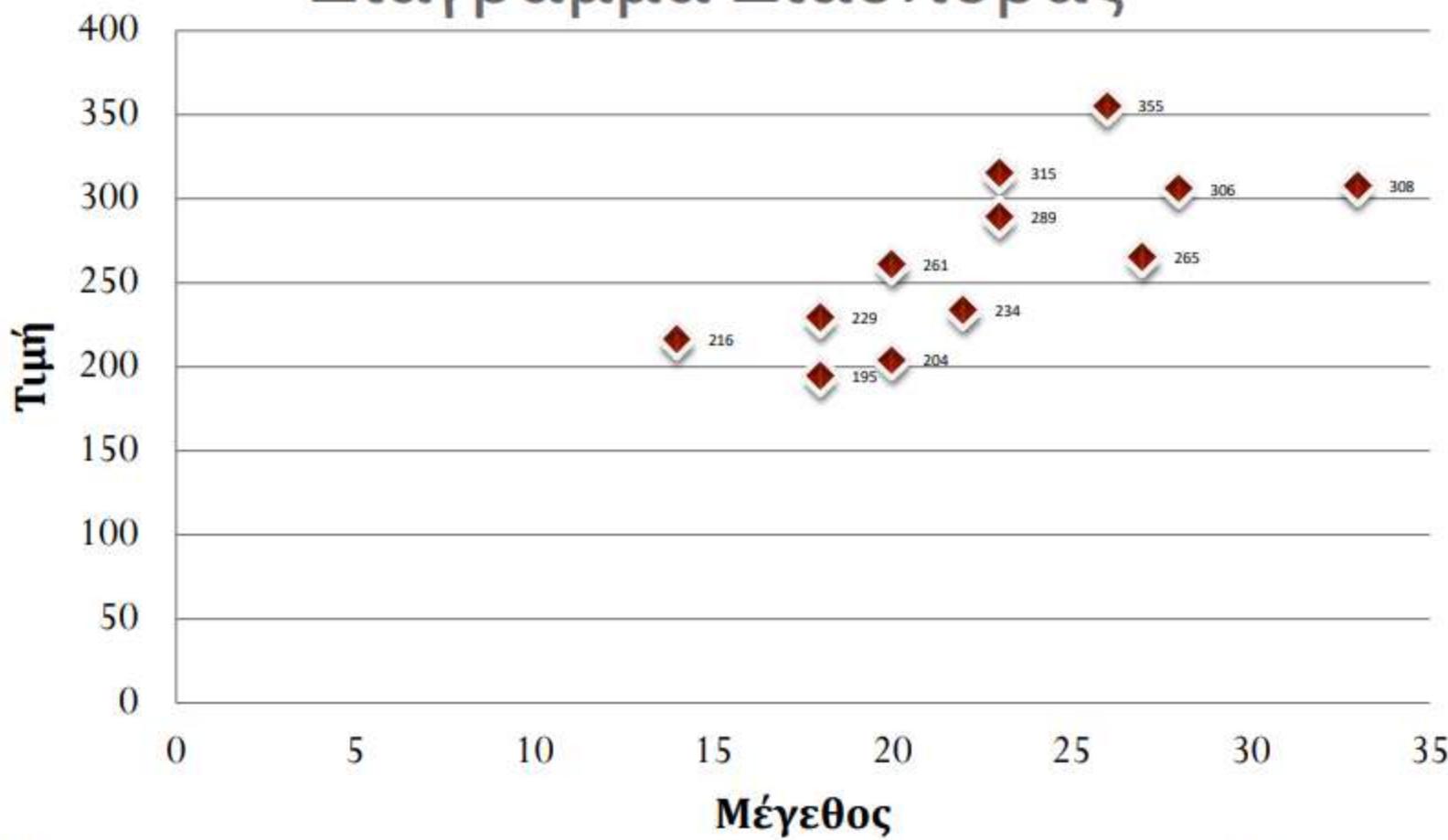
**Παράδειγμα:** τιμή πώλησης ακινήτου αναφορικά με το μέγεθος

- Η τιμή εξαρτάται από το μέγεθος
- Η σχέση των δύο μεταβλητών φαίνεται στο διάγραμμα διασποράς (scatter plot)
  - Γραμμικότητα (ισχυρή, μέτρια, ασθενής)
  - Κατεύθυνση (θετική, αρνητική)

Μέγεθος	Τιμή
23	315
18	229
26	355
20	261
22	234
14	216
33	308
28	306
23	289
20	204
27	265
18	195

# Παρουσίαση Δύο Ποσοτικών Μεταβλητών:

## Διάγραμμα Διασποράς



- Η ανεξάρτητη μεταβλητή,  $X$ , είναι το μέγεθος, ενώ η εξαρτημένη μεταβλητή,  $Y$ , είναι η τιμή
- Τα σημεία κινούνται σε μια ανοδική γραμμή
- Αυτό σημαίνει ότι η τιμή αυξάνεται με το μέγεθος
- Υπάρχουν και άλλες μεταβλητές που καθορίζουν την τιμή και θα χρειαστεί περαιτέρω ανάλυση για να καθοριστούν οι μεταβλητές αυτές.

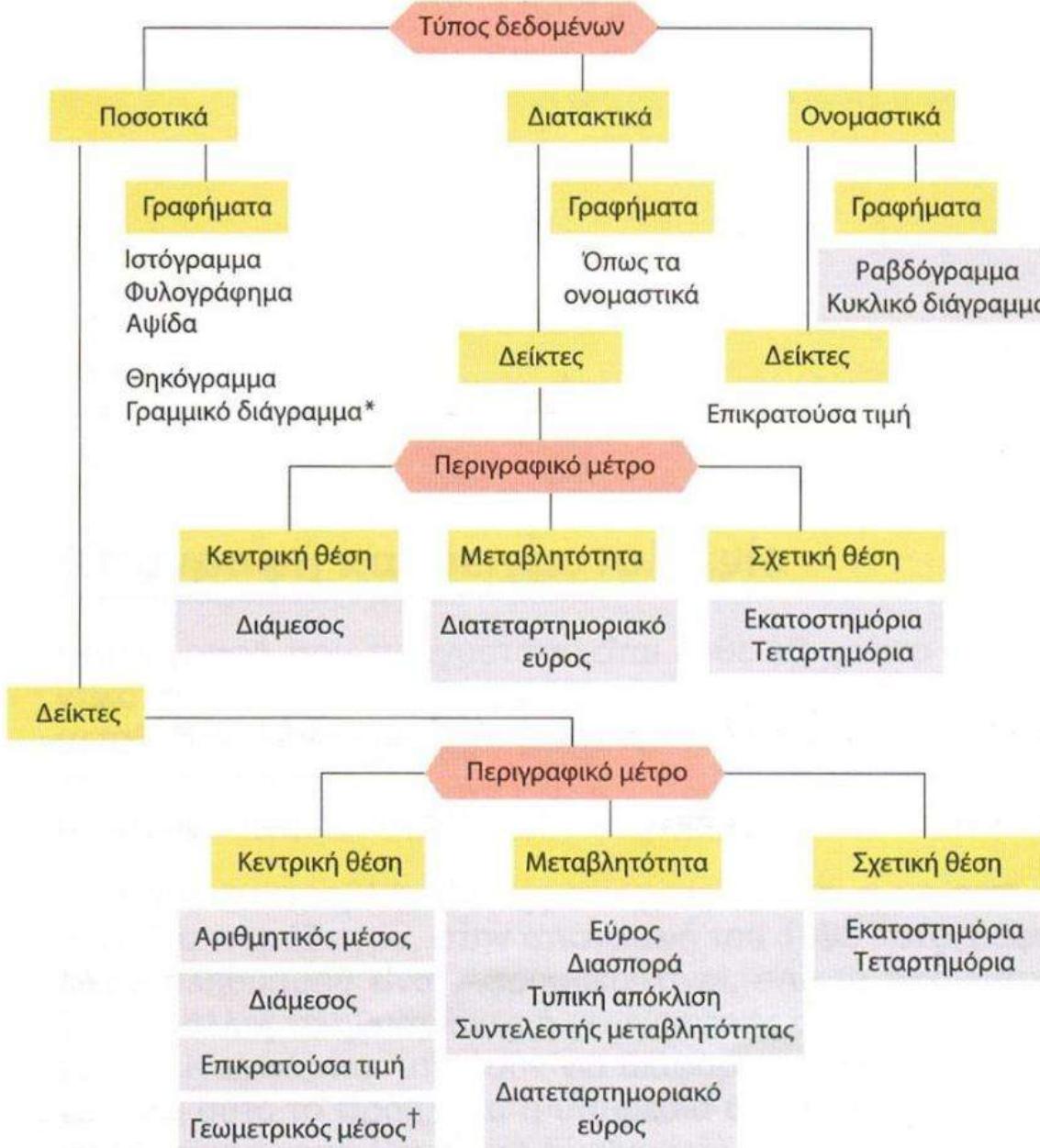
# Αριθμητικοί Δείκτες

# Περιγραφικά Μέτρα (Δείκτες)

Αντιπροσωπεύουν ένα χαρακτηριστικό του πληθυσμού ή του δείγματος

- **Δείκτες Κεντρικής Θέσης**
  - Μέσος, Διάμεσος, Επικρατούσα τιμή
- **Δείκτες Μεταβλητότητας**
  - Εύρος, Διασπορά, Τυπική Απόκλιση, Συντ. Μετ/τας
- **Δείκτες Σχετικής Θέσης**
  - Εκατοστημόριο, Διατεταρτημοριακό Εύρος
- **Δείκτες Γραμμικής Συσχέτισης**
  - Συνδιασπορά, Συντελεστής Συσχέτισης

# Περιγραφή ενός συνόλου δεδομένων



\*Χρονολογικές σειρές  
†Ρυθμοί αύξησης

# Δείκτες Κεντρικής Θέσης

---

Συνήθως, οι τιμές μιας μεταβλητής  $X$  τείνουν να συγκεντρωθούν γύρω από κάποια κεντρική της τιμή, η οποία μπορεί να χρησιμοποιηθεί:

- Ως ένα μέτρο της κεντρικής θέσης (central location)
- ή κεντρικής τάσεως (central tendency) της κατανομής της  $X$

## Αριθμητικός Μέσος (mean)

- Πληθυσμός:  $\mu = \frac{\sum_{i=1}^N x_i}{N}$ , όπου N το μέγεθος του πληθυσμού
- Δείγμα:  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ , όπου n το μέγεθος του δείγματος

**Παράδειγμα.** Χρήση διαδικτύου (σε ώρες) **10** εφήβων των τελευταίο μήνα: 0, 7, 12, 5, 33, 14, 8, 0, 9, 22

$$\bar{x} = \frac{0 + 7 + 12 + 5 + 33 + 14 + 8 + 0 + 9 + 22}{10} = 11$$

**Ερμηνεία:** 11 ώρες χρήσης κατά μέσο όρο

Τύπος δεδομένων: **ποσοτικά**

**Πρακτική ερμηνεία:** μπορούμε να θεωρήσουμε τη δειγματική μέση τιμή ως ένα «σημείο ισορροπίας». Αν κάθε παρατήρηση αντιστοιχεί σε 1<sup>gr</sup> μάζας, ένα υπομόχλιο που θα τοποθετούνταν στο σημείο  $\bar{x}$  θα ισορροπούσε το σύστημα μάζών.



## Διάμεσος (median)

- Η μεσαία τιμή, όταν ταξινομηθούν κατά αύξουσα ή φθίνουσα σειρά
  - Αν το πλήθος είναι **άρτιο**, είναι το ημιάθροισμα (μέσος όρος) των δύο μεσαίων τιμών: **n/2** και **n/2+1**
  - Αν το πλήθος είναι **περιττός**, είναι η τιμή που κατέχει τη θέση : **(n+1)/2**

**Παράδειγμα.** Χρήση διαδικτύου (σε ώρες) **10** εφήβων τον τελευταίο μήνα: 0, 7, 12, 5, 33, 14, 8, 0, 9, 22

- Διατάσουμε τις τιμές: 0, 0, 5, 7, **8, 9**, 12, 14, 22, 33
- Η διάμεσος είναι ίση με  $(8+9)/2 = 8.5$

**Ερμηνεία:** Οι μισές τιμές είναι κάτω από 8.5 και οι μισές τιμές είναι πάνω από 8.5

Τύπος δεδομένων που εφαρμόζεται: **ποσοτικά** ή **διατακτικά**

## Επικρατούσα Τιμή (mode)

- Η τιμή με την μεγαλύτερη συχνότητα

**Παράδειγμα.** Χρήση διαδικτύου (σε ώρες) **10** εφήβων τον τελευταίο μήνα: **0, 7, 12, 5, 33, 14, 8, 0, 9, 22**

Όλες οι τιμές έχουν ίδια συχνότητα, εκτός από την τιμή 0, που εμφανίζεται δύο φορές.

Άρα η επικρατούσα τιμή είναι ίση με 0

**Ωστόσο:** ο αριθμητικός μέσος είναι ίσος με **11**

και η διάμεσος είναι ίση με **8.5**

Τύπος δεδομένων: **ποσοτικά, διατακτικά ή ονομαστικά**

# Επικρατούσα τιμή

**Παράδειγμα:** Η προτίμηση μπύρας

Υπολογίζουμε τη συχνότητα εμφάνισης κάθε τιμής

Μεγαλύτερη συχνότητα εμφανίζει η μπύρα «Μύθος»

**Προβλήματα:**

- Δεν είναι πάντα αντιπροσωπευτική των τιμών του δείγματος
- Μπορεί να υπάρχουν πάνω από μία επικρατούσες τιμές

Μάρκα Μπύρας	Συχνότητα
Μύθος	90
FIX	19
Amstel	62
Heineken	13
Stella	59
Alfa	25
Corona	17

[Χωρίς τίτλο]

## Σύγκριση Κεντρικών Δεικτών

- Πρώτη επιλογή ο **αριθμητικός μέσος**
  - Ωστόσο επηρεάζεται από τις **ακραίες τιμές** (outliers)
- Υπάρχουν περιπτώσεις όπου η **διάμεσος** πλεονεκτεί
  - Δεν επηρεάζεται σημαντικά από τις ακραίες τιμές
- Η επικρατούσα τιμή επιλέγεται σπάνια
- Σε **μη ποσοτικά** δεδομένα, ο αριθμητικός μέσος **δεν είναι έγκυρος**
- Σε διατακτικά δεδομένα, μπορούμε να υπολογίσουμε **διάμεσο** και **επικρατούσα τιμή** (δώστε π.χ.)
- Σε ονομαστικά δεδομένα, **δεν μπορεί** να υπολογιστεί η διάμεσος, ούτε και η επικρατούσα τιμή (δώστε π.χ.)

# Περιγραφική Στατιστική (Μέτρα διασποράς)

## Έννοιες - Κλειδιά

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

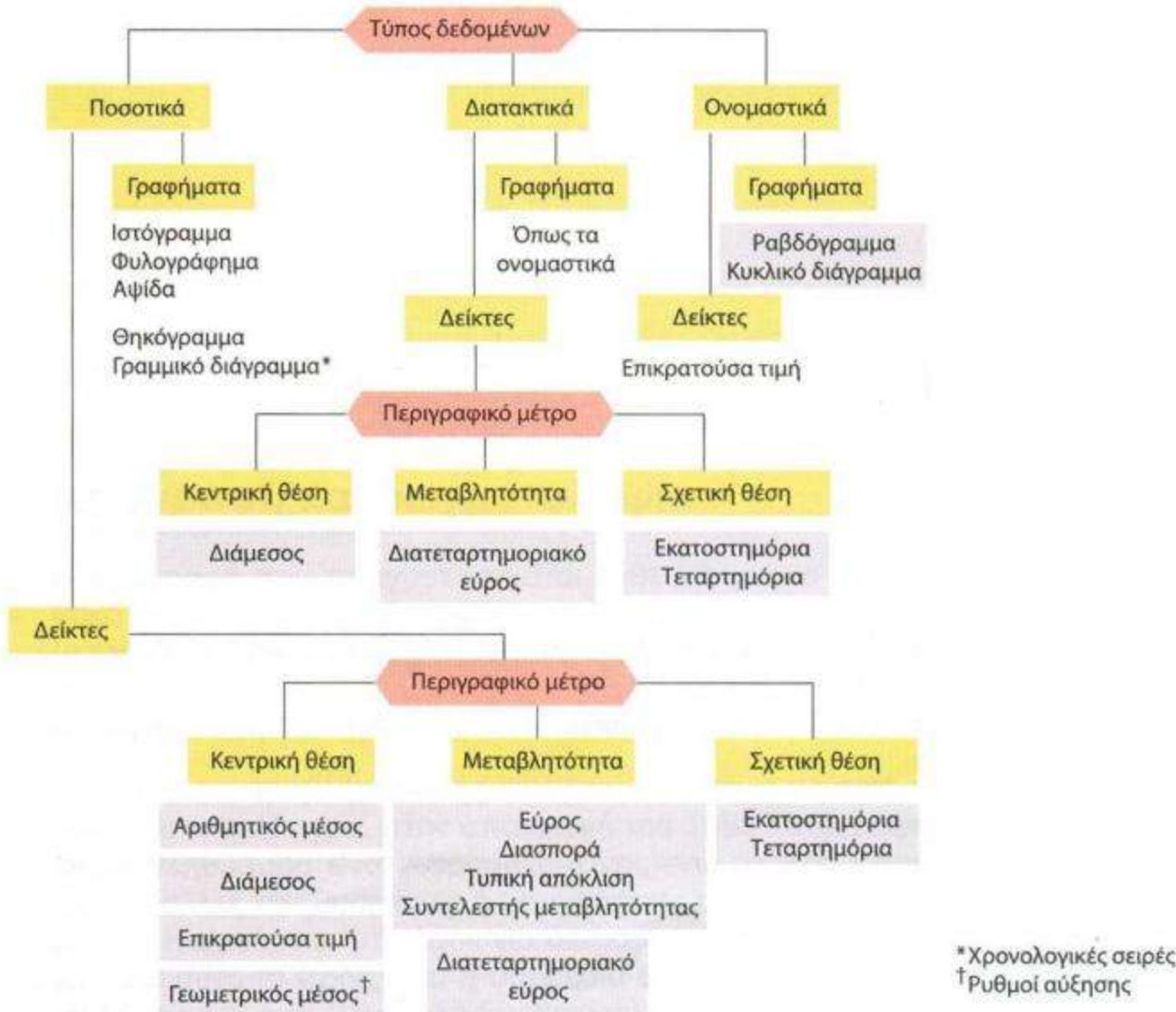
# Μέτρα θέσης (σύνοψη)

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

Τα κυριότερα μέτρα θέσης:

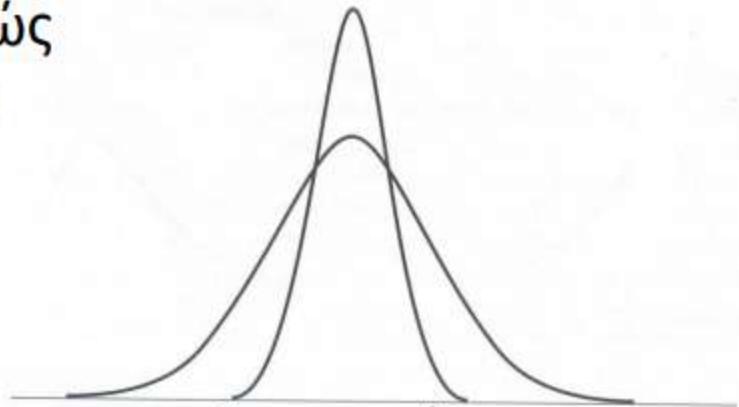
- Ο αριθμητικός μέσος (*ποσοτικά δεδομένα*)
- Η διάμεσος (*ποσοτικά ή διατακτικά*)
- Η επικρατούσα τιμή (*ποσοτικά, διατακτικά ή ονομαστικά*)

# 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

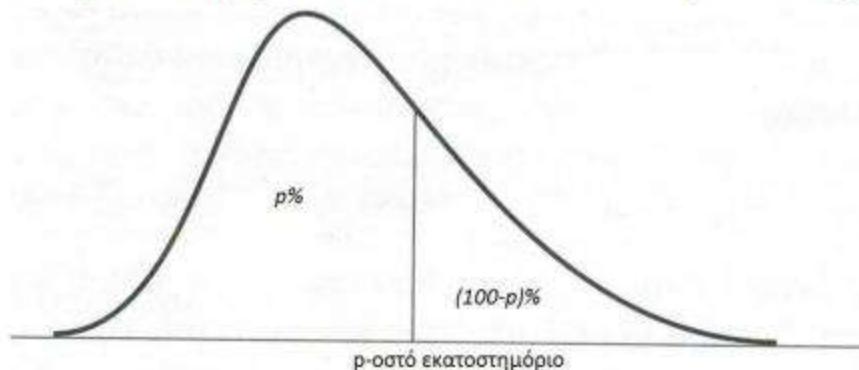
**Πλεονέκτημα:** Απλότητα στον υπολογισμό

**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  $(100-p)\%$



# 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)

- Ήρες χρήσης διαδικτύου:

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

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

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

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  $\frac{3}{4}$  (= 0.75) της απόστασης ανάμεσα στη 2<sup>η</sup> και την 3<sup>η</sup> τιμή

Άρα:  $Q_1 = 0 + \frac{3}{4} (5 - 0) = 3.75$

**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 5<sup>η</sup> και την 6<sup>η</sup> τιμή

Άρα:  $Q_2 = 8 + \frac{1}{2} (9-8) = 8.5$

[Χωρίς τίτλο]

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

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

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

## 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 **8<sup>η</sup>** και την **9<sup>η</sup>** τιμή

Άρα:  $Q_3 = 14 + \frac{1}{4} (22 - 14) = 16$

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 3<sup>η</sup> στήλη του πίνακα υπολογίζουμε τις αθροιστικές συχνότητες  $F_i$ .

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  $L$ .

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

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

# Απάντηση (2/3)

**β)** Για τον υπολογισμό του  $Q_2$  έχουμε:

- Προσδιορίζουμε την τιμή:  $p \ n = \frac{50}{100} \cdot 30 = 15$
- 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  $F_{i-1}$ .  $F_{i-1} = 7$

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	

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

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

**β)** Για τον υπολογισμό του  $Q_3$  έχουμε:

- Προσδιορίζουμε την τιμή:  $p \ n = \frac{75}{100} \cdot 30 = 22,5$

- 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  $F_{i-1}$ .  $F_{i-1} = 21$

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	

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

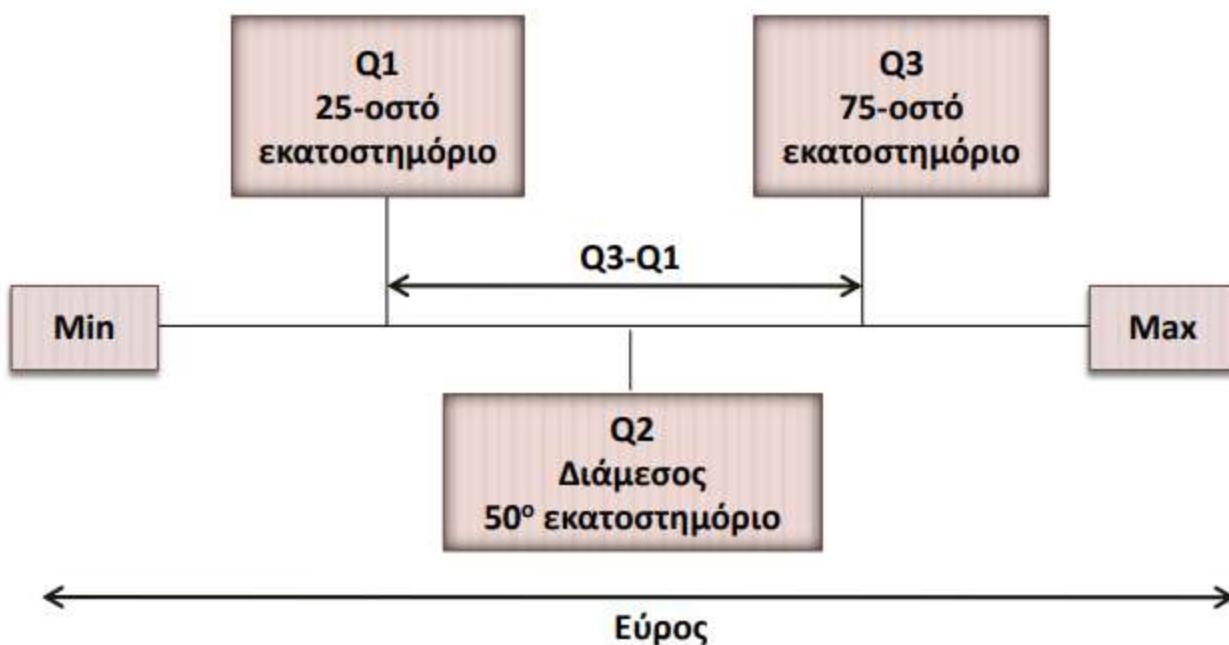
- Παρατηρούμε ότι ο επόμενος όρος, δηλ. ο 26, ανήκει στο ταξικό διάστημα 800-900, το κατώτατο όριο του οποίου το συμβολίζουμε με  $L$ , δηλ.  $L = 800$
- Πηγαίνουμε στην τάξη από την οποία προσδιορίσαμε την τιμή  $L$  και παρατηρούμε πόσες συχνότητες έχει. Αυτή είναι η τιμή του  $f_i$ , εδώ  $f_i = 5$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

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

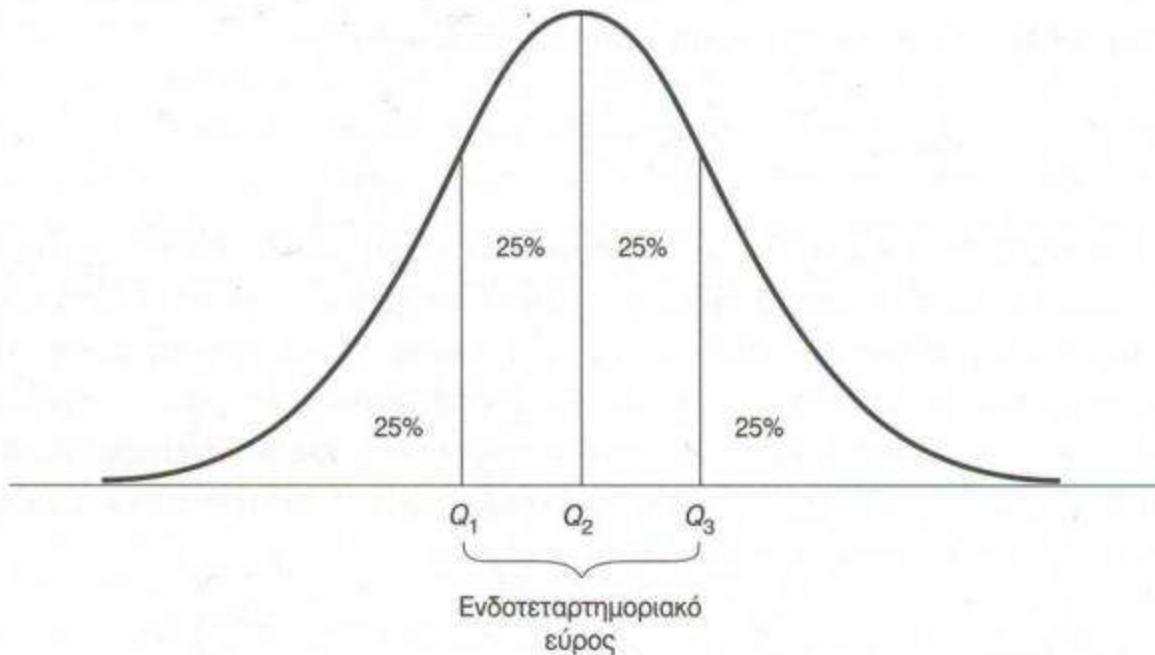
# Ενδοτεταρτημοριακό εύρος

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

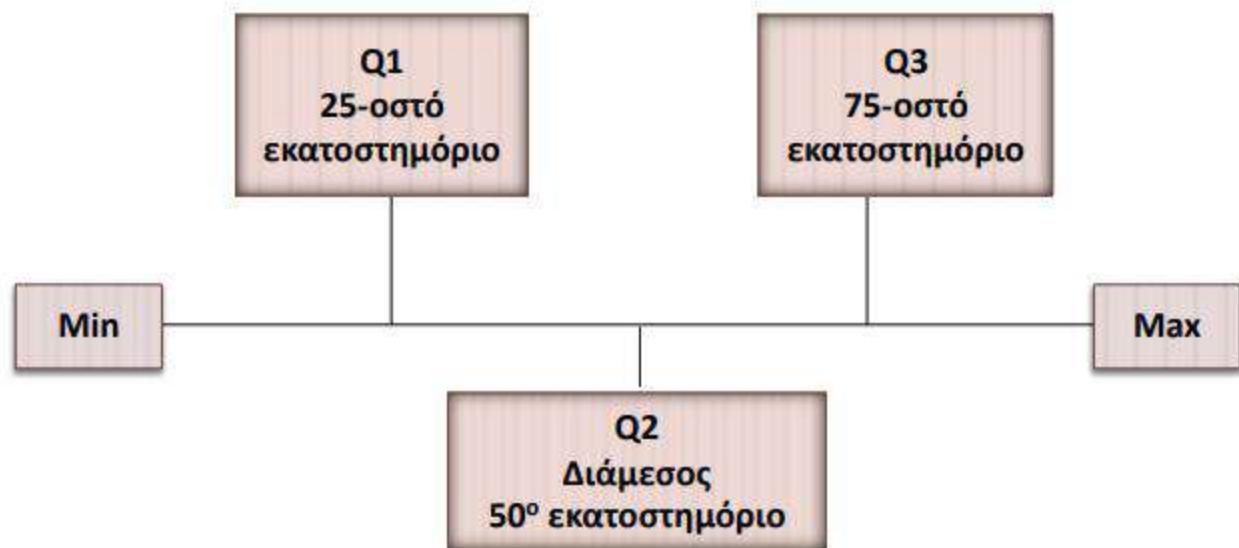
Μικρό διάστημα → μεγάλη συγκέντρωση τιμών → μικρή διασπορά τιμών

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

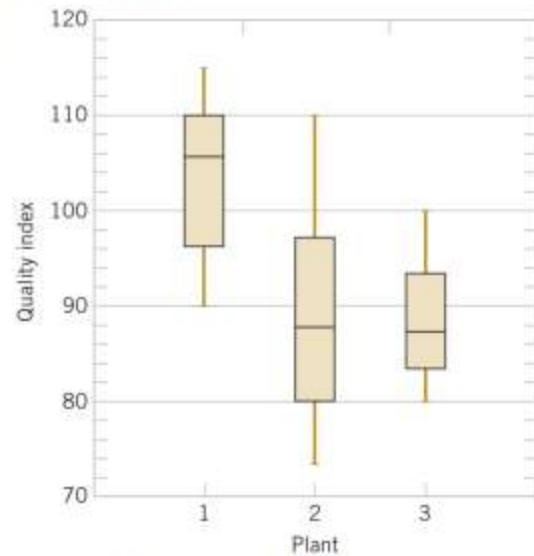
- Ίσως να οφείλονται σε λάθος καταγραφή, ή να κρύβουν χρήσιμες πληροφορίες
- Π.χ. ακραία τιμή (θετική ή αρνητική) στην απόδοση ενός πωλητή μιας επιχείρησης

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:

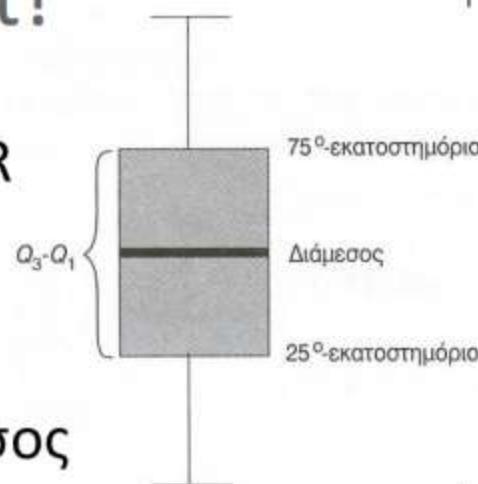
- η διάμεσος
- τα τεταρτημόρια
- το ενδοτεταρτημοριακό εύρος
- καθώς και τις ακραίες τιμές



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

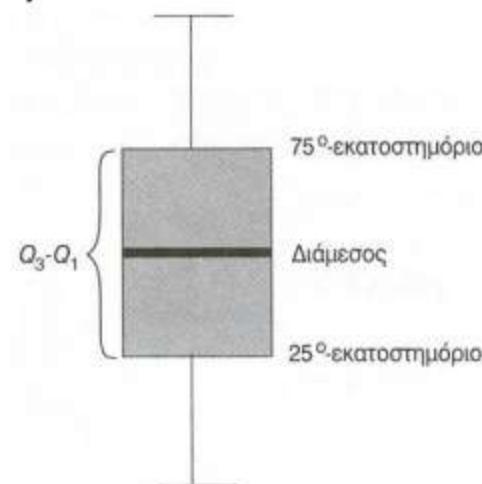
# Θηκόγραμμα – από τι αποτελείται?

- Από ένα ορθογώνιο παραλληλόγραμμο με ύψος IQR
- Κάτω οριζόντια πλευρά →  $25^{\circ}$  εκατοστημόριο
- Πάνω οριζόντια πλευρά →  $75^{\circ}$  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?

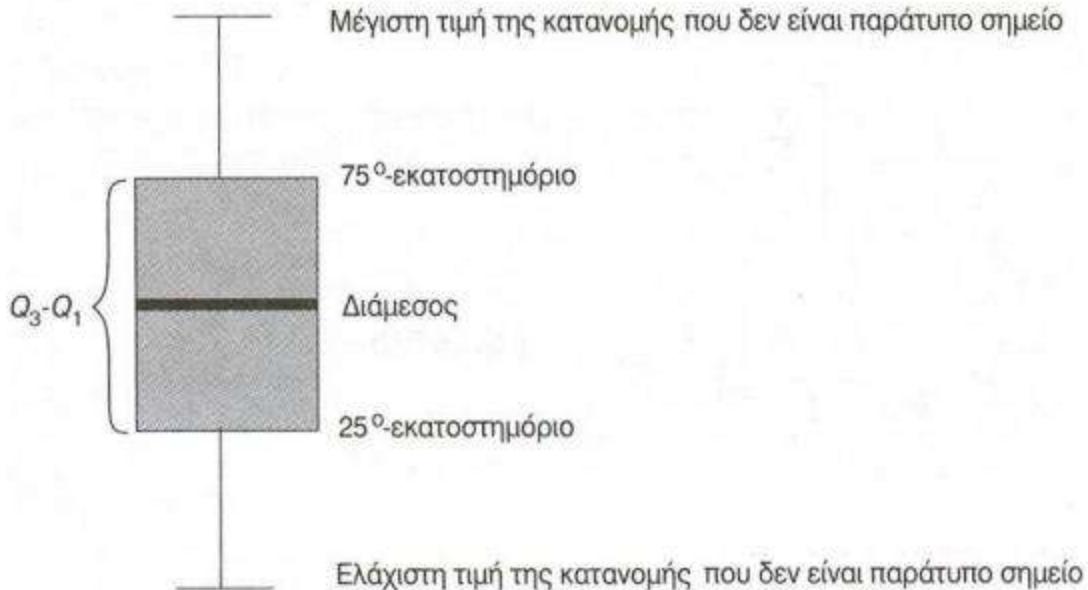
- Αν τα ακραία σημεία βρίσκονται σε απόσταση (από την άνω ή κάτω πλευρά) μικρότερη του  $3(Q_3 - Q_1)$ , δηλ. μεταξύ  $1,5(Q_3 - Q_1)$  και  $3(Q_3 - Q_1)$ , συμβολίζονται με έναν μικρό κύκλο (o)
- Διαφορετικά, συμβολίζονται με έναν αστερίσκο (\*)
- Εσωτερικός φράκτης  $\rightarrow \pm 1,5(Q_3 - Q_1)$
- Εξωτερικός φράκτης  $\rightarrow \pm 3(Q_3 - Q_1)$



# Παράδειγμα boxplot

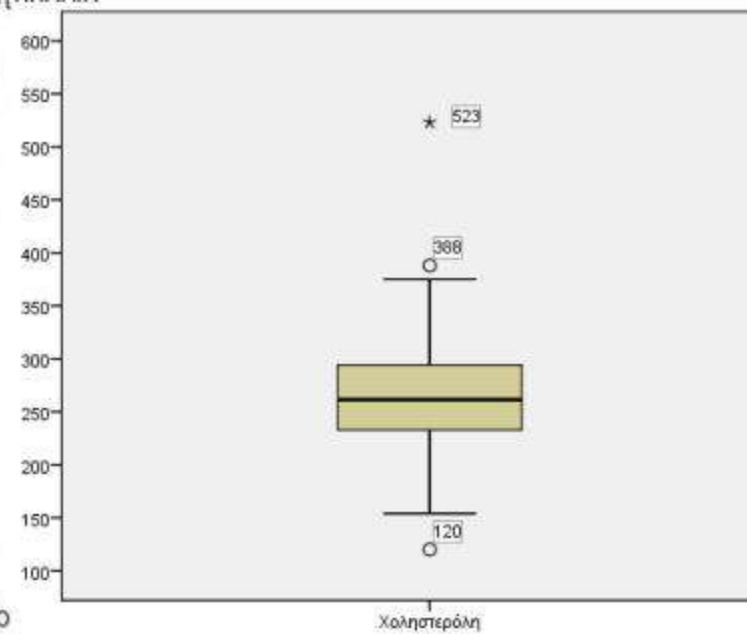
\* Τιμές μεγαλύτερες κατά  $3(Q_3 - Q_1)$  τουλάχιστον από το  $75^{\text{o}}$ -εκατοστημόριο

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



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

\* Τιμές μικρότερες κατά  $3(Q_3 - Q_1)$  τουλάχιστον από το  $25^{\text{o}}$ -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**:

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

Statistics		
workers		
	N	Valid
		25
		Missing
Mean		27,80
Median		18,00
Mode		11 <sup>a</sup>
Minimum		5
Maximum		92
Percentiles	25	11,50
	50	18,00
	75	41,00

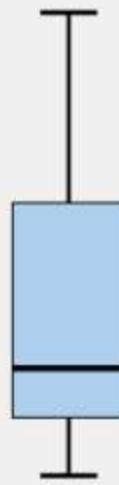
a. Multiple modes exist. The smallest value is shown

# Απάντηση

- Σχηματίζουμε το ορθογώνιο, κάτω πλευρά στο  $Q_1=11,5$ , πάνω πλευρά στο  $Q_3=41$  και αναπαριστούμε με παχιά γραμμή την διάμεσο ( $Q_2=18$ ).
- Υπολογίζουμε το:  $IRQ = Q_3 - Q_1 = 29.5$
- Συνεπώς:  $1,5 * IRQ = 44.25$  και  $3 * IRQ = 118$ .
- Εσωτερικός φράκτης  $\rightarrow \pm 1,5(Q_3 - Q_1) = \pm 44.25$  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 17<sup>η</sup> παρατήρηση με τιμή 92 είναι outlier. Το σημειώνουμε με κύκλο (ο)

- Εξωτερικός φράκτης  $\rightarrow \pm 3(Q_3 - Q_1) = \pm 88,5$  Κάτω όριο = 0, άνω όριο =  $41 + 88,5 = 129,5$ . Δεν υπάρχουν παρατηρήσεις μεγαλύτερες από 129,5 για να τις σημειώσουμε με αστερίσκο (\*)



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

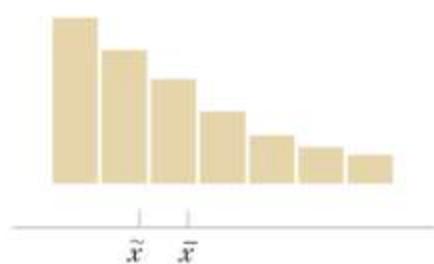
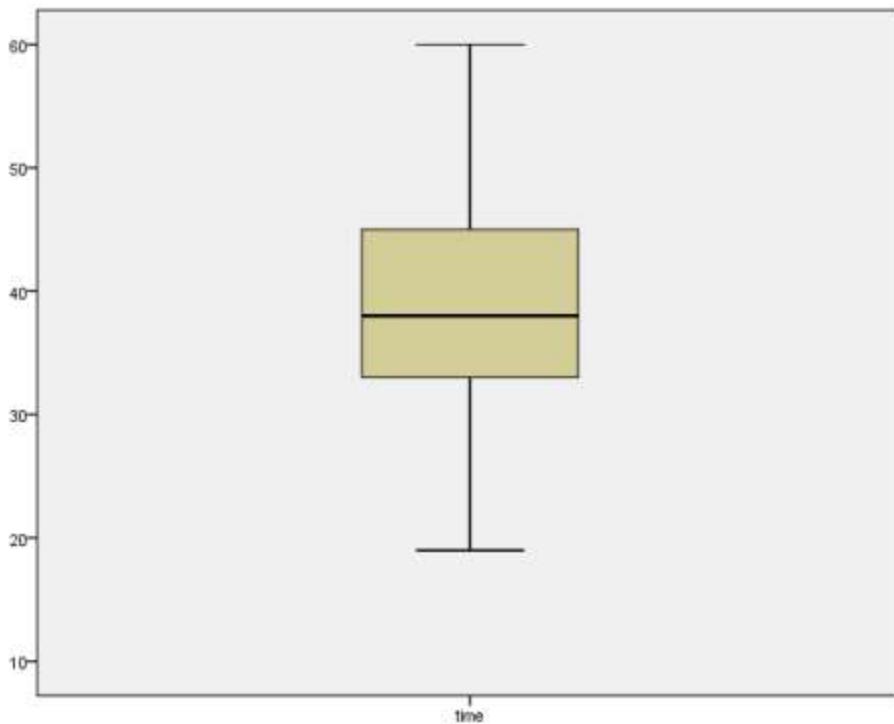
# Άσκηση

33	29	45	60	42	19	52	38	36
----	----	----	----	----	----	----	----	----

Mean 39,33  
Median 38,00  
Mode 19<sup>a</sup>  
Std. Deviation 12,247  
Variance 150,000

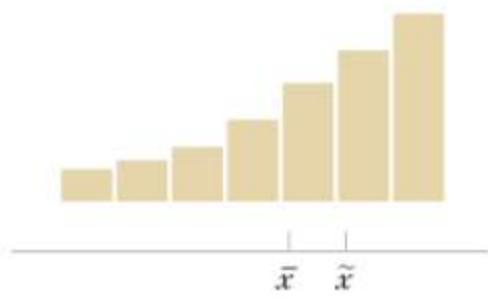
a. Multiple modes exist. The smallest value is shown

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

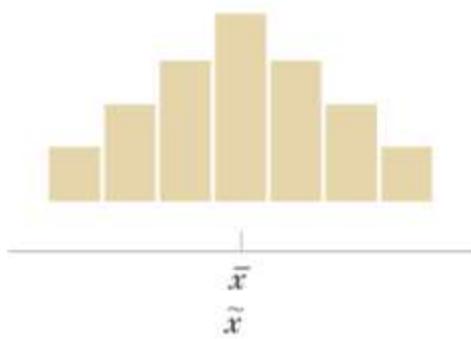


ΘΕΤΙΚΗ ή δεξιά λόξευση (right skew)

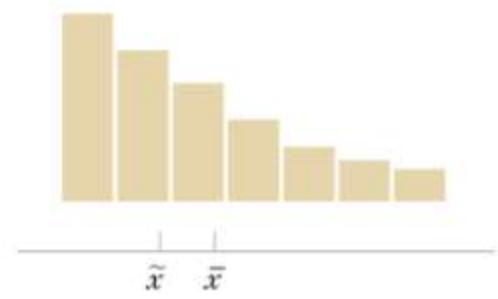
# Ασυμμετρία κατανομής δεδομένων



Αρνητική ή  
αριστερή λόξευση  
(left skew)  
(a)



Συμμετρική  
(b)



Θετική ή δεξιά  
λόξευση (right skew)  
(c)

## 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

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

- 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.

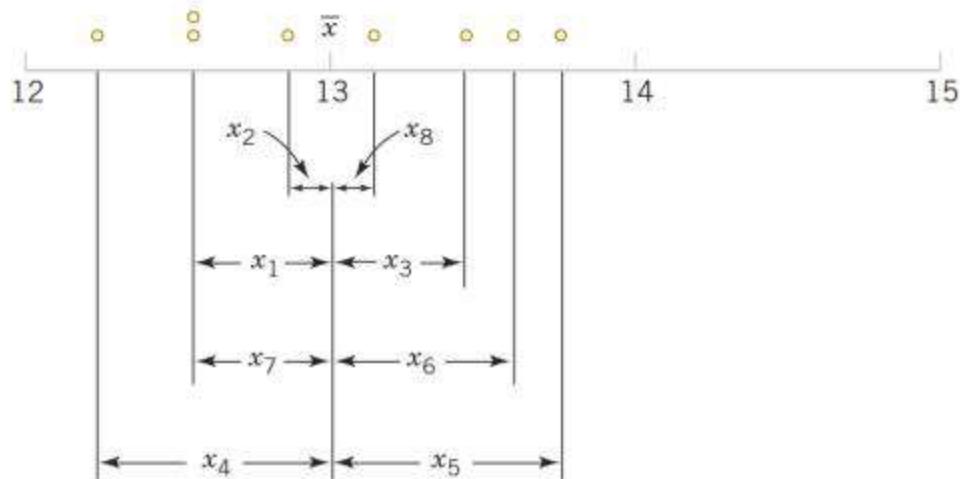
Το άθροισμα των αποκλίσεων ισούται πάντα με μηδέν! Πρέπει να χρησιμοποιήσουμε ένα μέτρο το οποίο θα μετατρέπει τις αρνητικές αποκλίσεις σε μη αρνητικές ποσότητες, υψώνουμε τις αποκλίσεις στο τετράγωνο.

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

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

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

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

## 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 (αιτήσεις)<sup>2</sup>

## 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.

Αξιοσημείωτες εφαρμογές του μέσου και της τυπικής απόκλισης

### A. 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 I: «Διαφορικές εξισώσεις» 0-100
- Μέρος II: «Στατιστική» 0-100

Κατά την τελική εξέταση του Ιουνίου είχαμε τα εξής αποτελέσματα:

	Μέρος I: Διαφορικές	Μέρος II: 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 I: Διαφορικές	Μέρος II: Στατιστική
$\bar{x}$	50	70
$s$	4	5
	<b>55</b>	<b>76</b>

- 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.
  - $CV_{\text{διαφορ.}} = \frac{s}{\bar{x}} = \frac{4}{50} = 0.08 \quad (8\%)$
  - $CV_{\text{Στατιστικ.}} = \frac{s}{\bar{x}} = \frac{5}{70} = 0.0714 \quad (7.14\%)$
- Η σωστή απάντηση μπορεί να δοθεί μόνο μετά από **σύγκριση των τυποποιημένων τιμών** των δύο βαθμολογιών του φοιτητή (**55 & 76**).
  - $z_{\text{διαφορικών}} = \frac{55 - \bar{x}}{s} = \frac{55 - 50}{4} = \frac{5}{4} = 1.25$
  - $z_{\text{Στατιστική}} = \frac{76 - \bar{x}}{s} = \frac{76 - 70}{5} = \frac{6}{5} = 1.20$
- Αφού  $z_{\text{διαφορικών}} > z_{\text{Στατιστική}}$ , συμπεραίνεται ότι ο συγκεκριμένος φοιτητής είναι **συγκριτικά καλύτερος** στις Διαφορικές και όχι στη Στατιστική

Αξιοσημείωτες εφαρμογές του μέσου και της τυπικής απόκλισης

Αξιοσημείωτες εφαρμογές του μέσου και της τυπικής απόκλισης

## B. 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 I: Διαφορικές	Μέρος II: Στατιστική
$\bar{x}$	50	70
$s$	4	5

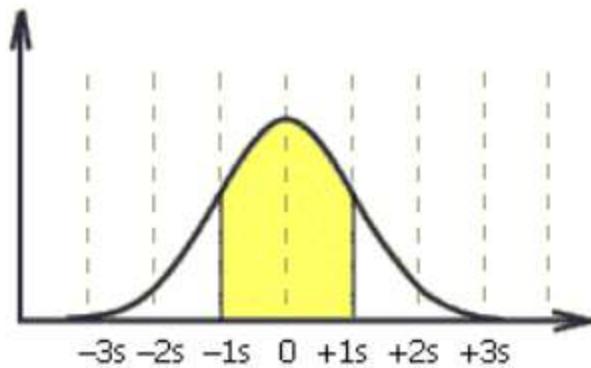
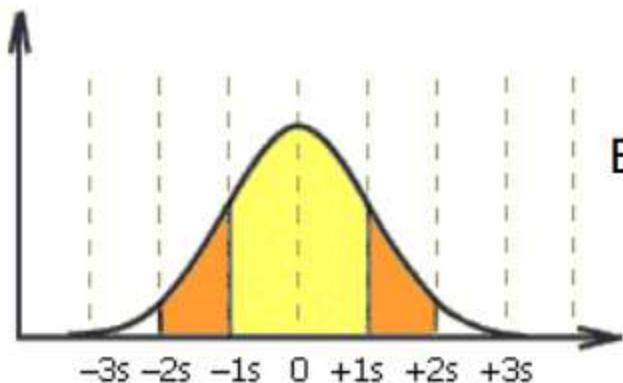
(παρατημένα μεταξύ των μέσων των δύο μεταβολών την ίδια συγκεκριμένα ποσοστά)

# Αξιοσημείωτες εφαρμογές του μέσου και της τυπικής απόκλισης

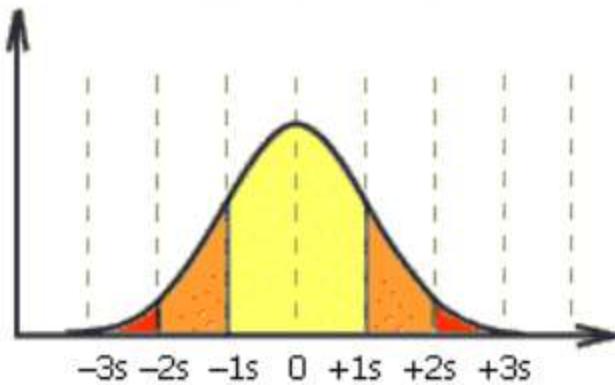
## Γ. Εμπειρικός Κανόνας

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

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



- B) 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

$$\text{Πληθυσμός: } CV = \frac{\sigma}{\mu}$$

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

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

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 βαθμολογιών στις Διαφορικές και στην Στατιστική των 100 φοιτητών του τμήματος

	Μέρος I: Διαφορικές	Μέρος II: 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.

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

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

- 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
<b>Σύνολα:</b>		<b>200</b>		<b>8790</b>	<b>677,250</b>

# 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 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:  $s^2 \approx \frac{1}{n-1} \left[ \sum_{i=1}^k f_i m_i^2 - \frac{(\sum_{i=1}^k f_i m_i)^2}{n} \right] = \frac{1}{331-1} \left[ 76110000 - \frac{141500^2}{331} \right] = 47332.78$

τυπική απόκλιση:  $s = \sqrt{s^2} = \sqrt{47332.78} \approx 217.56$

συντελεστής μεταβλητότητας:  $CV = \frac{s}{\bar{x}} = \frac{217.56}{427.49} = 0.50$