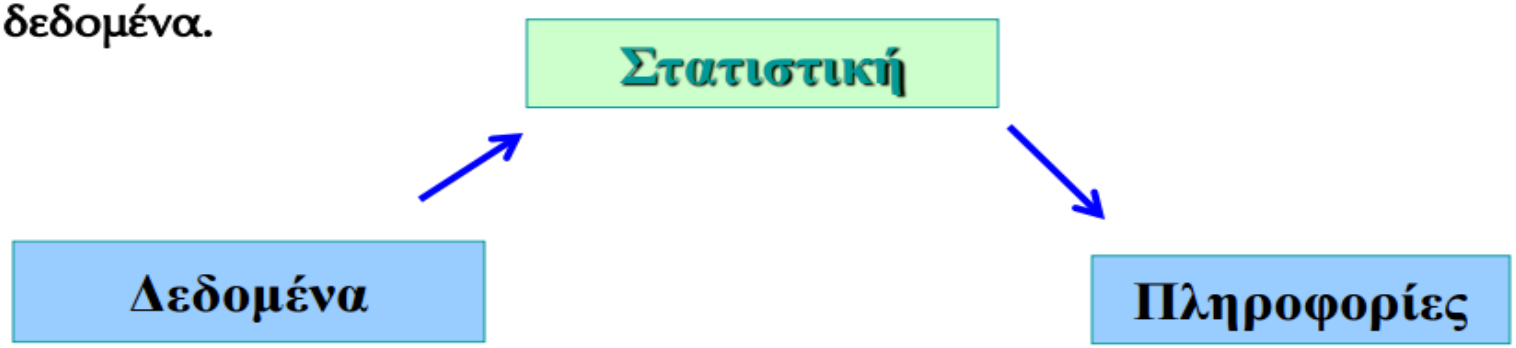
Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων

Καθηγητής
Δρ. Φιλιππάκης Μιχαήλ

Μιχαήλ Φιλιππάκης

Τι είναι Στατιστική;

Στατιστική είναι ένας τρόπος με τον οποίο αντλούμε πληροφορίες από δεδομένα.



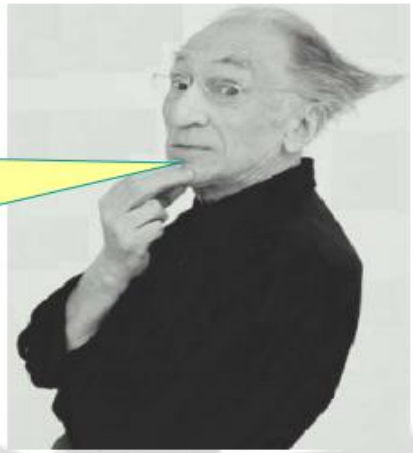
Δεδομένα: Γεγονότα, κυρίως αριθμητικά, συλλεγμένα για αναφορά ή πληροφορία.

Πληροφορία: Γνώση αναφερόμενη για κάποιο συγκεκριμένο γεγονός.

Στατιστική είναι ένα ***εργαλείο*** για να δημιουργήσουμε μία νέα αντίληψη από ένα σύνολο αριθμών.

Για παράδειγμα....

Γνωρίζω ότι οι φοιτητές φοβούνται την στατιστική γιατί θεωρούν ότι είναι δύσκολο μάθημα. Από τις περυσινές βαθμολογίες των φοιτητών μπορούμε να εξάγουμε κάποια πληροφορία?



Δεδομένα → Στατιστική → Πληροφορίες

Λίστα βαθμών από το προηγούμενο έτος

9.5
8.9
7.0
6.5
7.8
5.7
:

Νέα πληροφορία σχετικά με το μάθημα της στατιστικής.

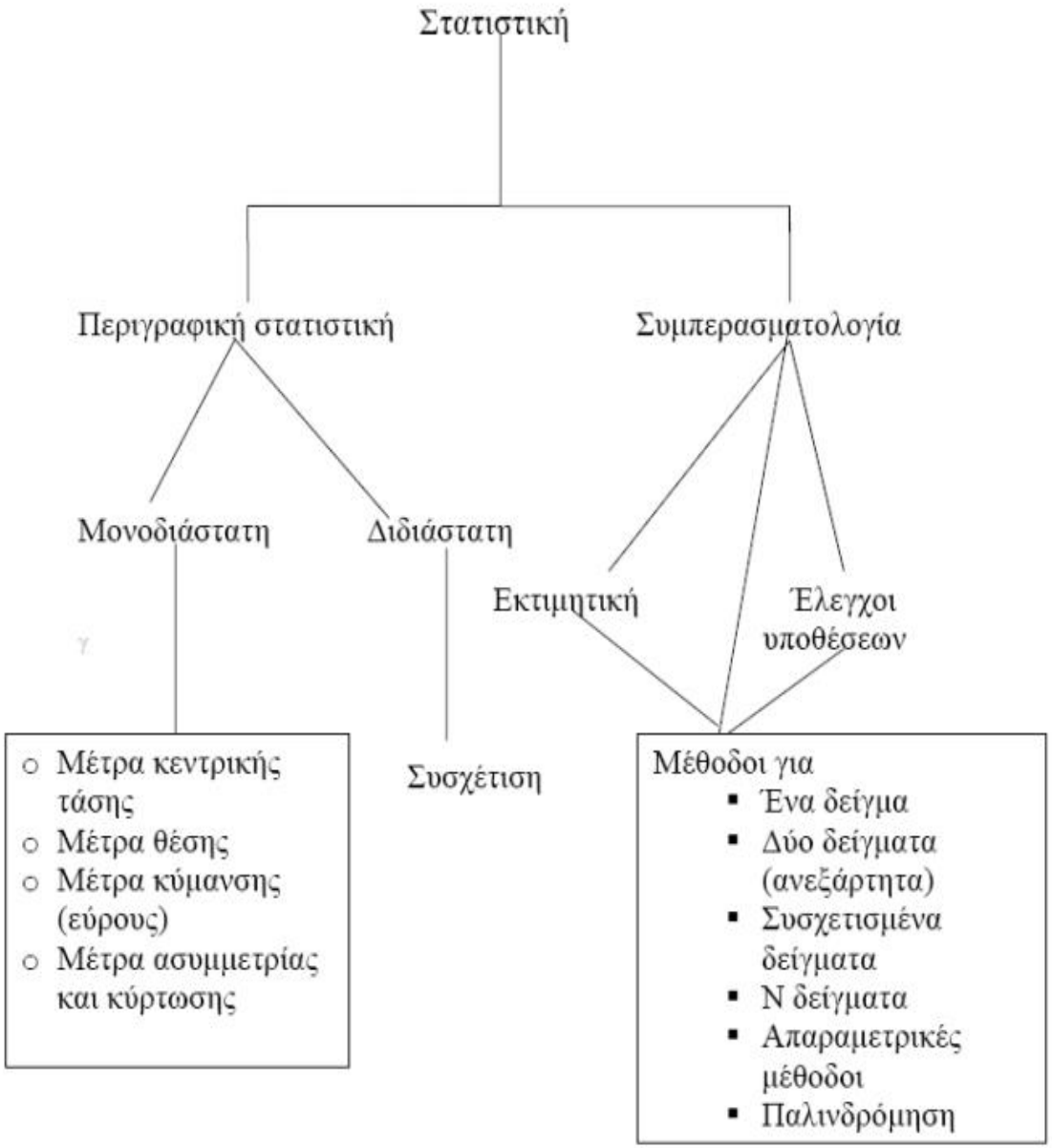
Π.χ. Μέσος όρος της τάξης,
Ποσοστό της τάξης που πήρε άριστα,
Ο βαθμός με την μεγαλύτερη συχνότητα,
Κατανομή βαθμών, κ.λ.π.

Γενικότερα....

- Η στατιστική αποτελεί αναπόσπαστο τμήμα της ερευνητικής διαδικασίας.
- Παρέχει το μέσο προκειμένου να αποφασίσουμε αν τα νούμερα που έχουμε στη διάθεση μας υποστηρίζουν ή όχι την ερευνητική ερώτηση που έχουμε κάθε φορά.

Ερωτήσεις:

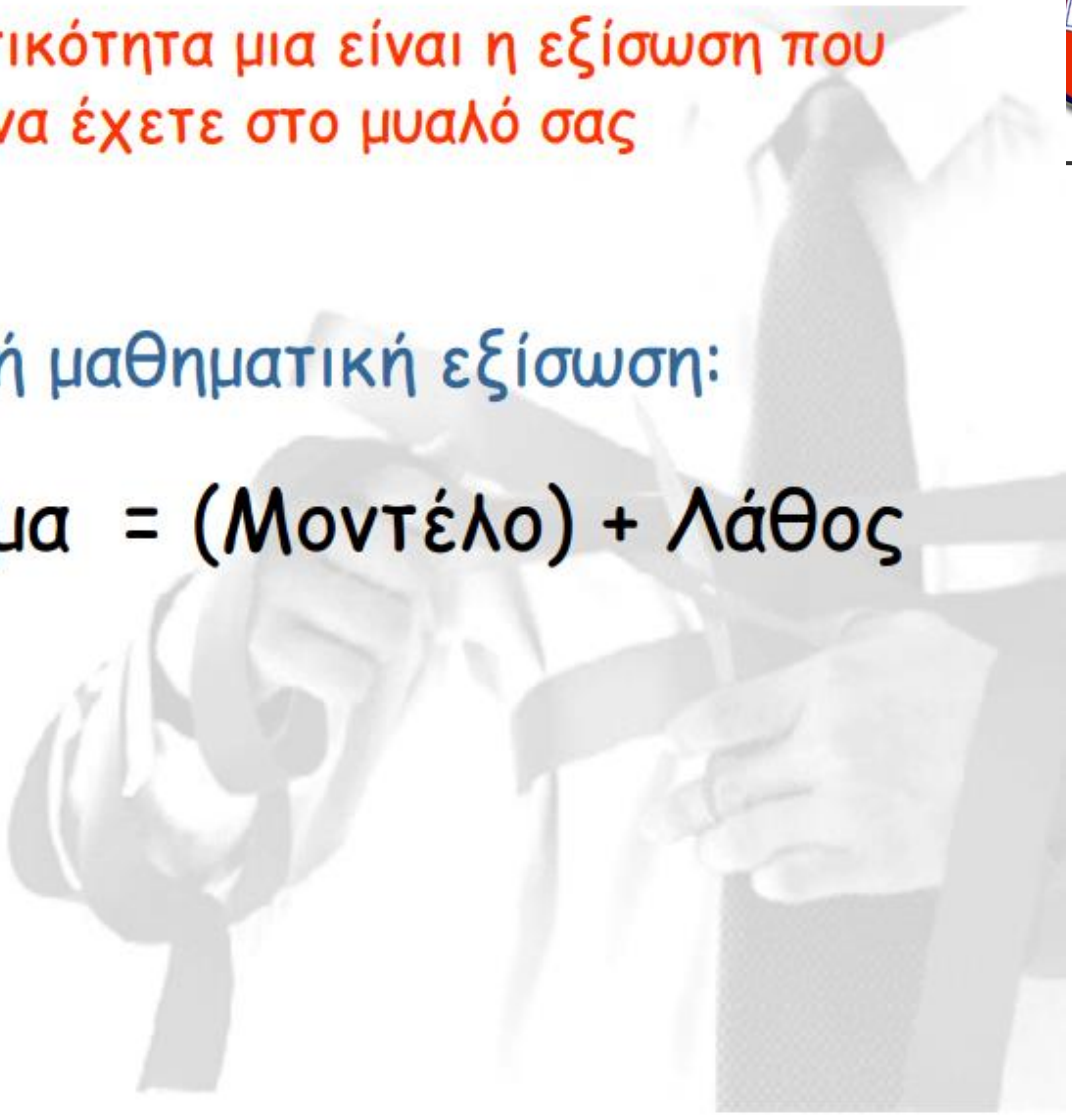
1. Τι ισχύει συνήθως?-Ποια είναι η τυπική (μέση) κατάσταση?
2. Ποια είναι η ποικιλότητα που υπάρχει?
3. Για ποιον ακριβώς μιλάμε? (Δείγμα)
4. Πόσο σίγουροι είμαστε για αυτό που λέμε?
5. Με τι θα πρέπει να συγκρίνουμε?


Στατιστική
Περιγραφική στατιστική
Συμπερασματολογία
Μονοδιάστατη
Διδιάστατη
Εκτιμητική
Έλεγχοι υποθέσεων
Μέτρα κεντρικής τάσης
Μέτρα θέσης
Μέτρα κύμανσης (εύρους)
Μέτρα ασυμμετρίας και κύρτωσης
Συσχέτιση
Μέθοδοι για
Ένα δείγμα
Δύο δείγματα (ανεξάρτητα)
Συσχετισμένα δείγματα
Ν δείγματα
Απαραμετρικές μέθοδοι
Παλινδρόμηση

Στην πραγματικότητα μια είναι η εξίσωση που θα πρέπει να έχετε στο μυαλό σας

Βασική μαθηματική εξίσωση:

Αποτέλεσμα = (Μοντέλο) + Λάθος



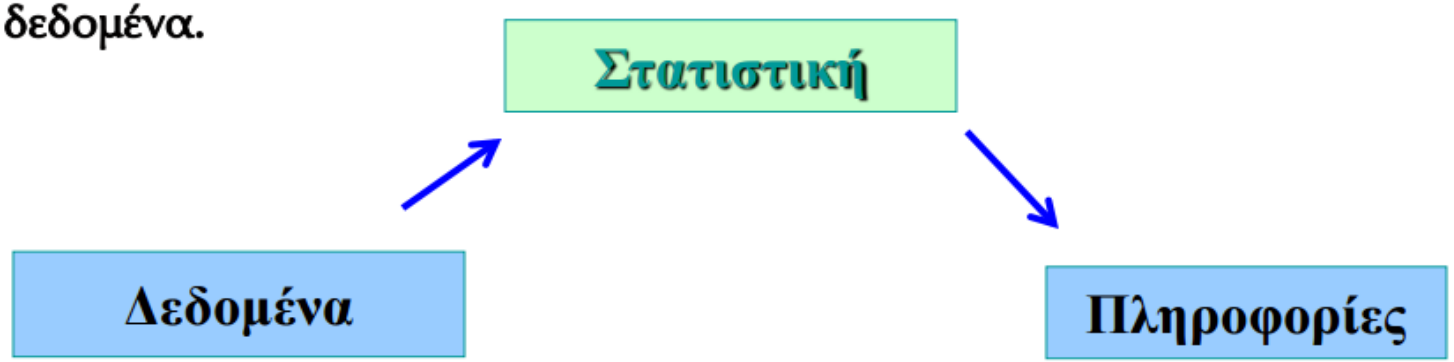
Ο μέσος όρος ως παράδειγμα ενός απλού στατιστικού μοντέλου

Ο μέσος όρος είναι στην ουσία ένα απλό μοντέλο. Αποτελεί μια σύνοψη των δεδομένων μας. Είναι μια υποθετική τιμή, η οποία μπορεί να υπολογιστεί για κάθε σύνολο δεδομένων αλλά **δεν είναι απαραίτητο** να υπάρχει στα δεδομένα μας.

Για παράδειγμα υποθέστε ότι ρωτάμε 5 φοιτητές να μας πουν τον αριθμό των «κολλητών» που έχουν. Τα δεδομένα μας μπορεί να είναι ως εξής: 1, 2, 3, 3 και 4. Στην περίπτωση αυτή ο Μ.Ο = (1+2+3+3+4)/5 = 2,6

Είναι μάλλον αδύνατο να έχουμε 2,6 φίλους. Επομένως η τιμή του Μ.Ο. είναι μια υποθετική τιμή. Υπό την έννοια αυτή ο Μ.Ο. αποτελεί ένα μοντέλο που κατασκευάσαμε προκειμένου να περιγράψουμε κατά τρόπο συνοπτικό τα δεδομένα μας.

Στατιστική είναι ένας τρόπος με τον οποίο αντλούμε πληροφορίες από δεδομένα.



Δεδομένα: Γεγονότα, κυρίως αριθμητικά, συλλεγμένα για αναφορά ή πληροφορία.

Πληροφορία: Γνώση αναφερόμενη για κάποιο συγκεκριμένο γεγονός.

Αλλά από πού τα δεδομένα έρχονται; Πως μαζεύονται; Πως εξασφαλίζεται η ορθότητα τους; Αντιπροσωπεύουν τον πληθυσμό από τον οποίο επιλέχθηκαν;

Μέτρα θέσης

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

Τα κυριότερα μέτρα θέσης:

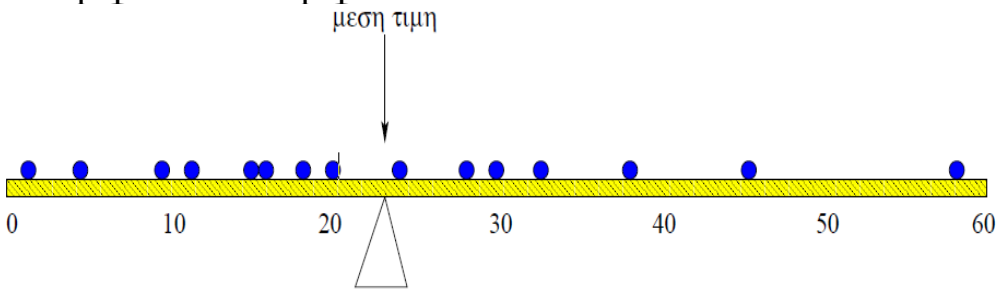
- Ο αριθμητικός μέσος *(ποσοτικά δεδομένα)*
- Η διάμεσος *(ποσοτικά ή διατακτικά)*
- Η επικρατούσα τιμή *(ποσοτικά , διατακτικά ή ονομαστικά)*

Μέτρα Θέσης (Measures of Location

- Η **Δειγματική μέση τιμή ή αριθμητικός μέσος**
- Το κέντρο ισορροπίας των δεδομένων

$$\bar{x} = \frac{x_1 + x_2 + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x} = \sum_{i=1}^{n} w_i x_i, \sum_{i=1}^{n} w_i = 1$$

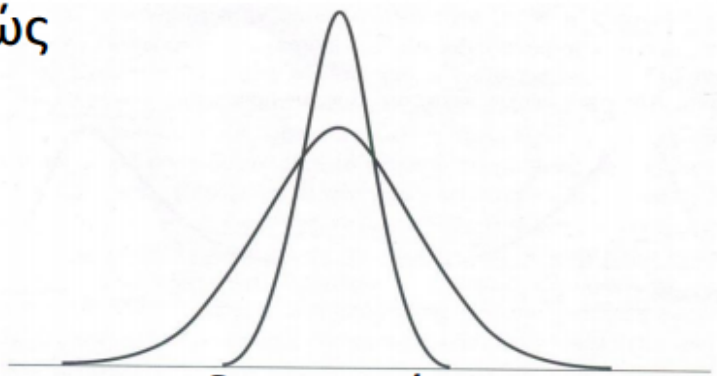


Μέτρα Θέσης (Measures of Location

- Η **επικρατούσα τιμή** ενός συνόλου δεδομένων είναι η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης.

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
- 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

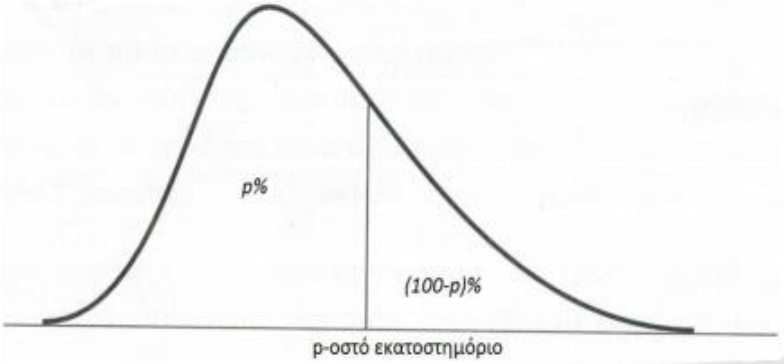
Πλεονέκτημα: Απλότητα στον υπολογισμό

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)

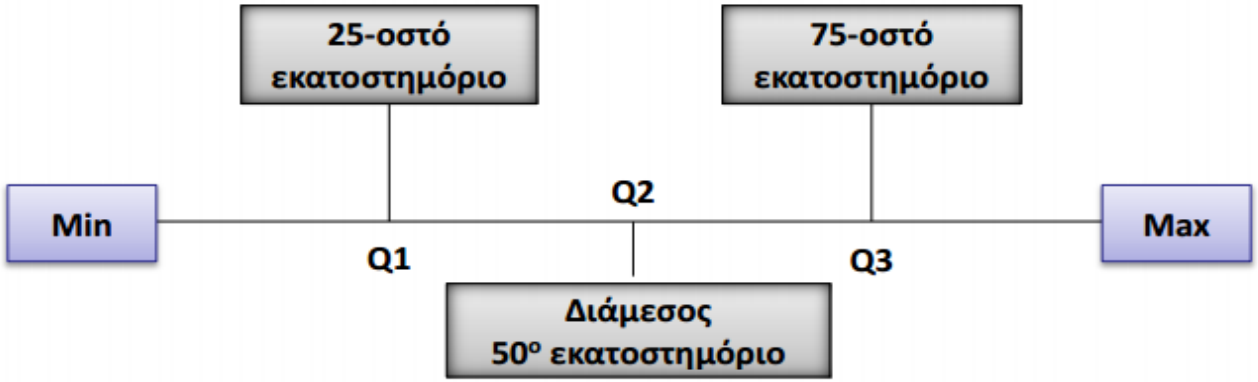
- Τα **εκατοστημόρια** *(percentiles)* αποτελούν γενίκευση της έννοιας της διαμέσου *(median)*.
- Το p-οστό εκατοστημόριο ενός συνόλου είναι εκείνη η τιμή, η οποία, όταν οι τιμές διαταχθούν σε **αύξουσα σειρά**, έχει από αριστερά της το p% των δεδομένων και από δεξιά της το υπόλοιπο (100-p)%



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

Μέτρα Διασποράς (Measures of Dispersion

- *Πρώτο Τεταρτημόριο (First Quartile)*

- Q_1 = η $\frac{n+1}{4}$ παρατήρηση.
- *Η θέση του Q_1 προκύπτει, αν χρειαστεί, από την στρογγυλοποίηση του $\frac{n+1}{4}$ στον κοντινότερο ακέραιο*

Μέτρα Διασποράς (Measures of Dispersion

- *Τρίτο Τεταρτημόριο (Third Quartile)*
- Q_3 = η $\frac{3(n+1)}{4}$ παρατήρηση στρογγυλεμένη στον κοντινότερο ακέραιο
- *Η θέση του Q_3 προκύπτει, αν χρειαστεί, από την στρογγυλοποίηση του $\frac{3(n+1)}{4}$ στον κοντινότερο ακέραιο.*

Μέτρα Διασποράς (Measures of Dispersion

Χαρακτηριστικά εκατοστιαία σημεία είναι τα *τεταρτομόρια*

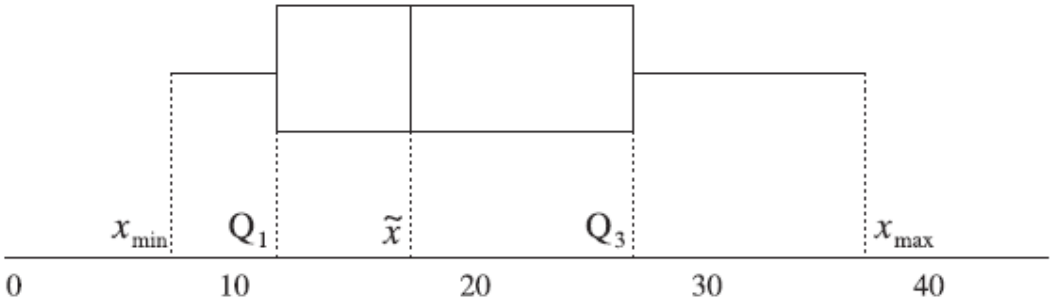
- **πρώτο ή κατώτερο τεταρτομόριο** Q_1: το 25-εκατοστιαίο σημείο
- **τρίτο ή ανώτερο τεταρτομόριο** Q_3: το 75-εκατοστιαίο σημείο

Q_1 και Q_3 ορίζονται όπως η διάμεσος αλλά περιορίζοντας το σύνολο των δεδομένων στα αντίστοιχα υποσύνολα (κατώτερο ή ανώτερο μισό).

ενδοτεταρτομοριακό εύρος I

$I = Q_3 - Q_1$ είναι το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά διάμεσο

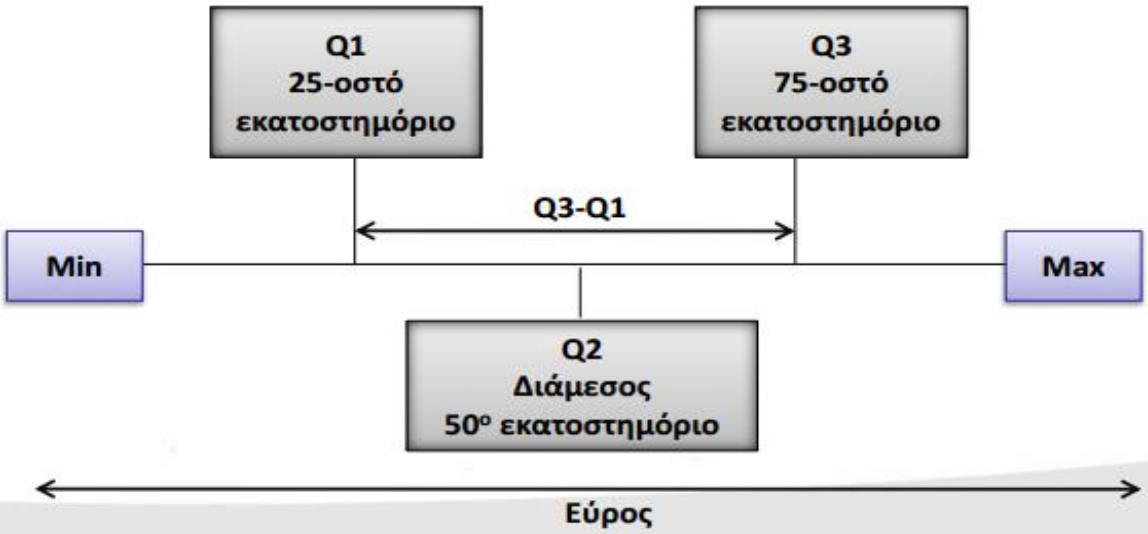
σύνοψη των 5 αριθμών - θηκόγραμμα



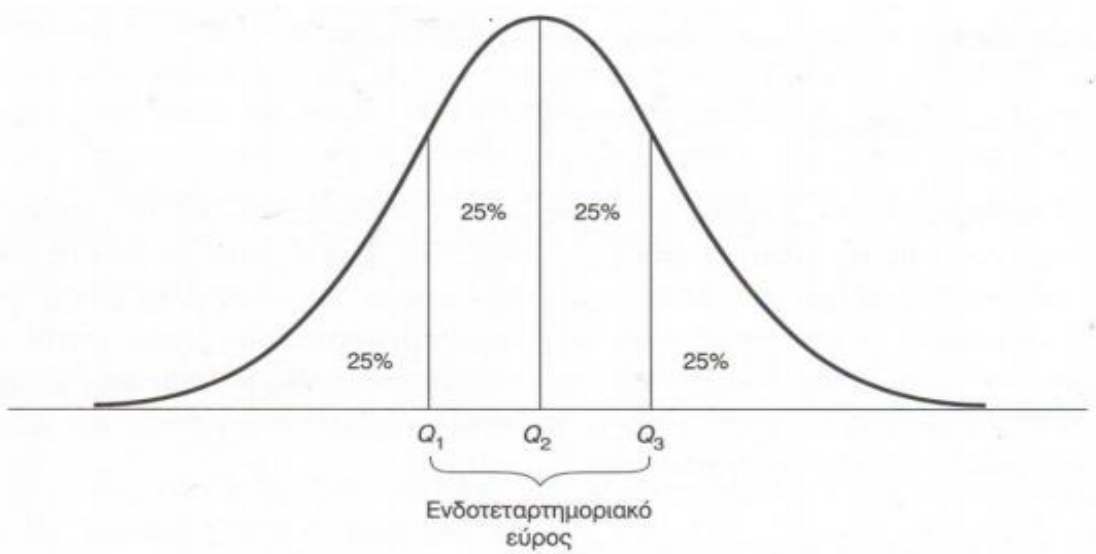
Ενδοτεταρτημοριακό εύρος

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

$$IQR = Q_3 - Q_1$$



Ενδοτεταρτημοριακό εύρος (2/2)

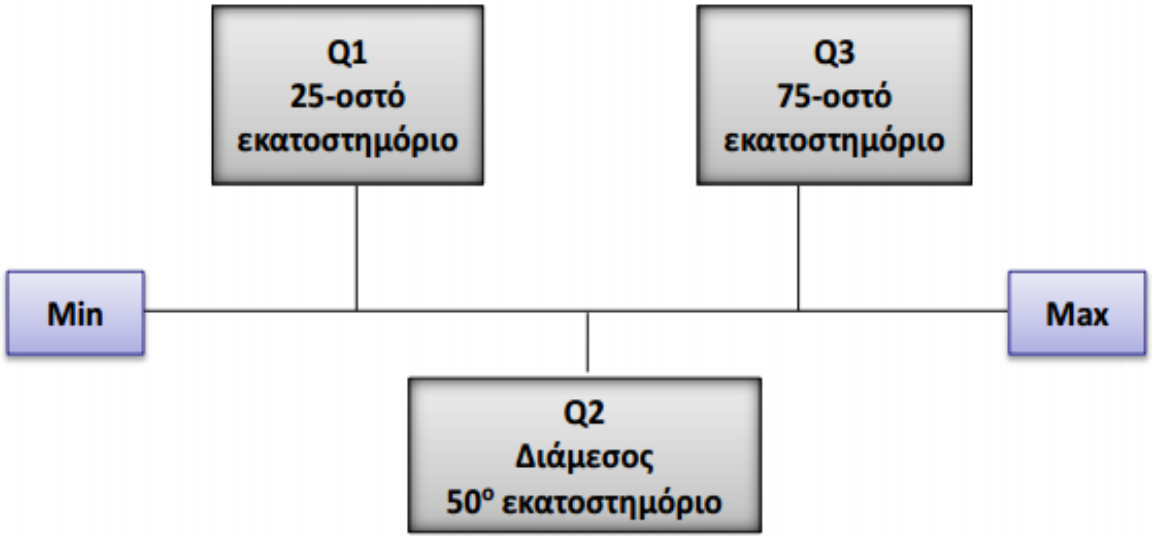


- Στο μεταξύ τους διάστημα περιέχεται το 50% των τιμών του δείγματος
- Μικρό διάστημα → μεγάλη συγκέντρωση τιμών → μικρή διασπορά τιμών
- Μεγάλη τιμή του IQR δείχνει **μεγάλη μεταβλητότητα**

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**, απομακρυσμένε 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

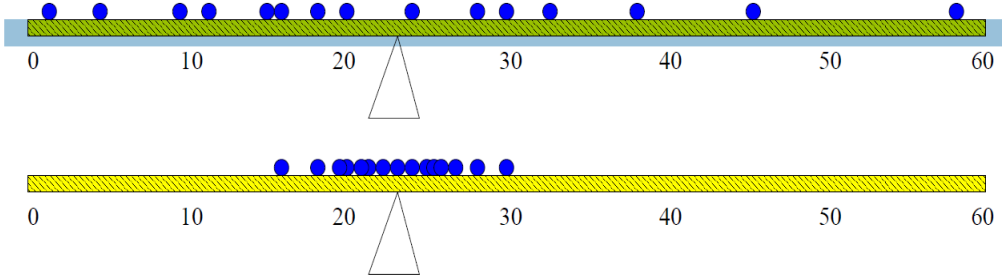
Τα θηκογράμματα (box plots) είναι γραφήματα τα οποία συνοψίζουν βασικά περιγραφικά μέτρα, όπως:

- 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.

Μέτρα Διασποράς (Measures of Dispersion

- Διακύμανση (variance)
- Μετραει την μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή



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

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

- Η διαφορά τους ασήμαντη όταν το μέγεθος του δείγματος μεγάλο

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

Μέτρα Διασποράς (Measures of Dispersion

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

ισοδύναμα

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

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.

Μέτρα Διασποράς- Σχετικής Μεταβλητότητας

- **Τυπική απόκλιση** (standard deviation)

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

- **Συντελεστής μεταβλητότητας (**coefficient of variation)

$$CV = \frac{S}{\bar{x}} 100\%$$

- **Εκφράζει την μεταβλητότητα των μετρήσεων απαλλαγμένη από την επίδραση της μέσης τιμής**

Μέτρα Διασποράς (Measures of Dispersion

- 10 δείγματα επίδοσης που προκύπτουν από simulations φαίνονται στον παρακάτω πίνακα

#run	ΔC/C
1	0.40
2	0.51
3	0.51
4	0.54
5	0.55
6	0.59
7	0.63
8	0.67
9	0.75
10	2.10
Σύνολο	7.25

Μέτρα Διασποράς (Measures of Dispersion

δειγματική μέση τιμή:

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{7.25}{10} = 0.725$$

δειγματική διάμεσος:

$$\tilde{x} = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{x_5 + x_6}{2} = \frac{0.55 + 0.59}{2} = 0.57$$

εύρος τιμών δείγματος:

$$x_{\min} = 0.40 \qquad x_{\max} = 2.10 \quad \longrightarrow \quad R = 1.70$$

δειγματική διασπορά (Πρώτα το άθροισμα τετραγώνων)

$$\sum_{i=1}^{10} x_i^2 = 0.40^2 + 0.51^2 + \cdots + 0.75^2 + 2.10^2 = 7.44$$

$$s^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) = \frac{1}{9}\left(7.44 - 10 \cdot 0.725^2\right) = 0.243$$

Μέτρα Διασποράς (Measures of Dispersion

δειγματική τυπική απόκλιση $s = \sqrt{s^2} = \sqrt{0.243} = 0.493$

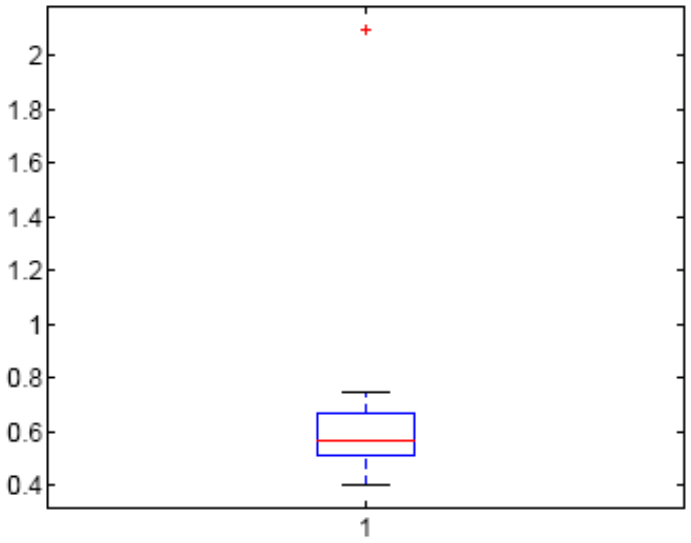
$x_{\min}$=0.40 $\tilde{x}$ =0.57 $x_{\max}$ =2.10

0.40 0.51 0.51 0.54 0.55 0.59 0.63 0.67 0.75 2.10

Q_1 =0.51 Q_3 =0.67

ενδοτεταρτομοριακό εύρος

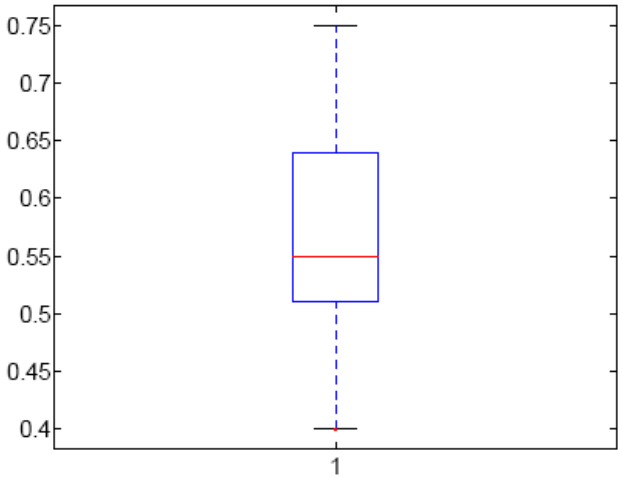
$I = Q_3 - Q_1 = 0.67 - 0.51 = 0.16$



Μέτρα Διασποράς (Measures of Dispersion

Απαλοιφή ακραίας τιμής ($\longrightarrow$ 9 παρατηρήσεις)

δειγματική μέση τιμή	$\bar{x} = \frac{5.15}{9} = 0.572$
δειγματική διάμεσος	$\tilde{x} = x_{(n+1)/2} = x_5 = 0.55$
Διασπορά	$s^2 = \frac{1}{8}\left(5.15 - 9 \cdot 0.572^2\right) = 0.010$
Τυπική απόκλιση	$s = \sqrt{0.010} = 0.10$
Ελάχιστη τιμή	$x_{\min} = 0.40$
Μέγιστη τιμή	$x_{\max} = 0.75$
Εύρος	$R = 0.75 - 0.40 = 0.35$
Πρώτο τεταρτομόριο	(διάμεσος των $\{x_1, \ldots, x_5\}$) $Q_1 = x_3 = 0.51$
Τρίτο τεταρτομόριο	(διάμεσος των $\{x_5, \ldots, x_9\}$) $Q_3 = x_7 = 0.63$
Ενδοτεταρτομοριακό εύρος	$I = 0.63 - 0.51 = 0.12$



Συντελεστής Μεταβλητότητας, CV

Εκτιμά τη σχέση της τυπικής απόκλισης με το μέγεθος των δεδομένων

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

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