Chapter 2

Introduction to Data Mining

Data

A. **Τύποι δεδομένων** (Data types / representation)

B. **Ποιότητα δεδομένων** (Data quality)
   - Noise
   - Outliers
   - Missing values
   - Duplicates

C. **Προεπεξεργασία** (preprocessing)
   - Διακριτοποίηση
   - Μείωση διάστασης (dimension reduction)
   - Επιλογή χαρακτηριστικών (feature selection)

D. **Συσχετίσεις μεταξύ δεδομένων**
   - Ομοιότητα / αποστάσεις μεταξύ δεδομένων

What is Data?

- Collection of data objects and their attributes

- Τα δεδομένα / πληροφορία υπό διάφορες οπτικές γωνίες
  - Εγγραφή (record).
  - Σημείο ενός χώρου (data point)
  - Διάνυσμα χαρακτηριστικών (feature vector)
  - Πρότυπο (pattern)
  - Γεγονός ή συμβάν (event)
  - Περίπτωση (circumstance)
  - Στιγμιότυπο
  - Δείγμα (sample)
  - Παρατήρηση (observation)

What is Data?

- Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- Τα δεδομένα αποτελούνται από:
  - μεταβλητές ή
  - χαρακτηριστικά ή
  - συνιστώσες ή
  - διαστάσεις

Attributes / features / variables / dimensions

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attributes

- Attribute values are numbers or symbols assigned to an attribute
- There are different types of attributes

symbolic

- Nominal (ονομαστικά)
  - Examples: ID numbers, eye color, zip codes
- Ordinal (σειριακό αριθμητικό που περιέχει αρκετή πληροφορία για να διαταχθεί)
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}

numeric

- Interval (διαστήματα τιμών)
  - Examples: calendar dates, temperatures in Celsius
- Ratio (αναλογίες)
  - Examples: temperature in Kelvin, length, time, counts

Ανάλογα με τον τύπο τιμής που έχουν

- Διακριτά χαρακτηριστικά (*Discrete attributes*)
  - Πεπερασμένο ή απείρως αριθμήσιμο πεδίο τιμών
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Συνήθως ακέραιες τιμές
  - Δυαδικά χαρακτηριστικά (true/false, 1/0)

- Συνεχή χαρακτηριστικά (*Continuous attributes*)
  - Συνεχείς πραγματικές τιμές
  - Examples: temperature, height, or weight.
  - Continuous attributes are typically represented as floating-point variables.

3 γνωρίσματα των δεδομένων που επηρεάζουν την ανάλυσή μας

- **Dimensionality (διάσταση)**
  - Πλήθος χαρακτηριστικών (μεταβλητών ανάλυσης)
  - Curse of Dimensionality («κατάρα της διάστασης»)
- **Sparsity (αραιότητα)**
  - Only presence counts
  - Μηδενικές τιμές χαρακτηριστικών σε ορισμένα στυγμιότυπα
- **Resolution (ευκρίνεια)**
  - Patterns depend on the scale
  - Η συμπεριφορά εξαρτάται από την κλίμακα των δεδομένων

Μορφές αναπαράστασης δεδομένων

- **Record data** (εγγραφές μιας βάσης δεδομένων)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- **Data Matrix :** μορφή πίνακα όπου οι στήλες είναι μεταβλητές

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

- **Feature vectors** (Διανυσματική μορφή)

$$x = \left(x_1, x_2, \ldots, x_d \right)$$

διάνυσμα *d* χαρακτηριστικών

Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

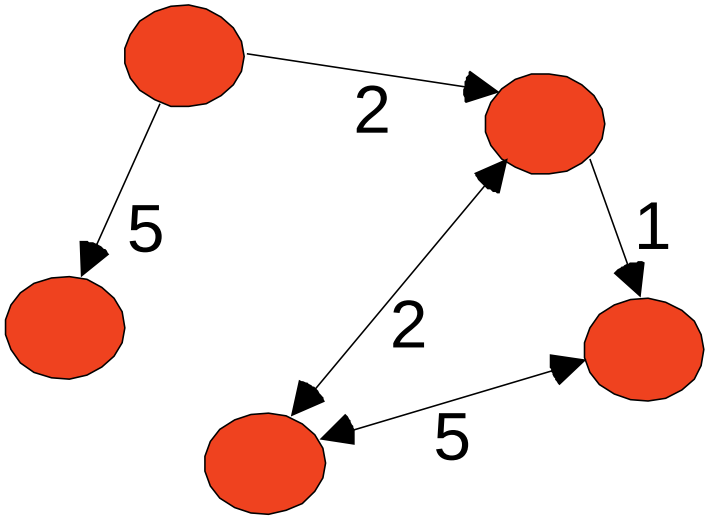
Document Data

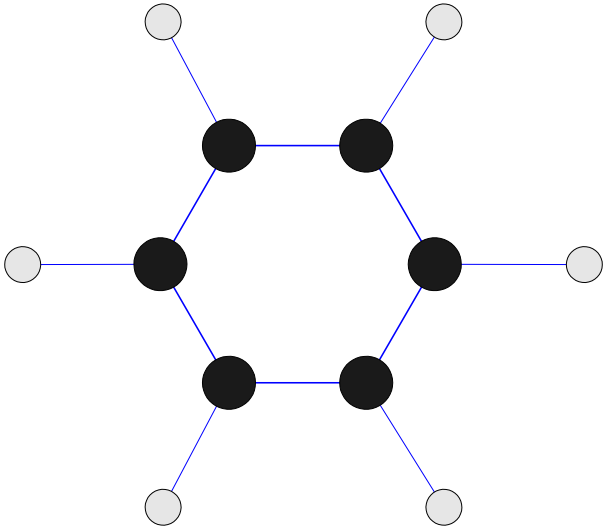
- διανυσματική αναπαράσταση κειμένων
  - Ορίζουμε ένα λεξικό από λέξεις (όροι)
  - Κάθε κείμενο αναπαρίσταται ως διάνυσμα συχνοτήτων των όρων
  - Εναλλακτική μορφή tf-idf

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Μορφές αναπαράστασης δεδομένων (συν.)

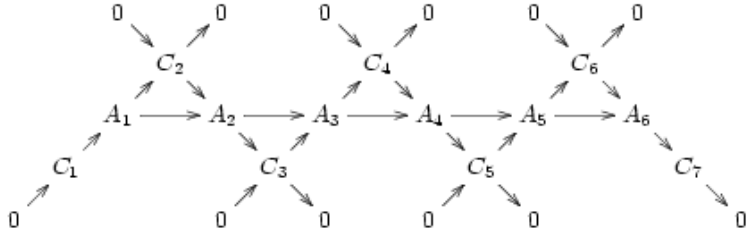
- **Image data:** διανυσματική αναπαράσταση εικόνων
  - Τιμές φωτεινότητας των pixels
  - Ιστόγραμμα της φωτεινότητας των pixels
  - Ορίζουμε χαρακτηριστικά πάνω στην εικόνα (ακμές, μεταβολές, …)
- **Graph data:** γραφική αναπαράσταση
  - Δεδομένα διαδικτύου, δίκτυα αισθητήρων, web pages links, κλπ.



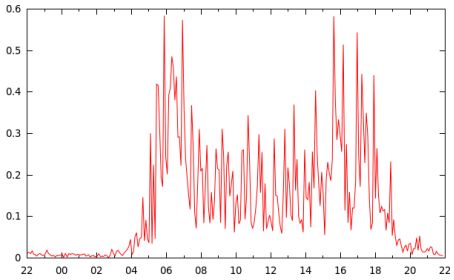


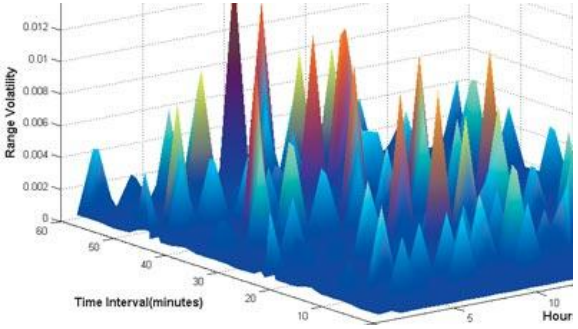
Μορφές αναπαράστασης δεδομένων (συν.)

- **Sequential Data:** ακολουθίες συμβόλων

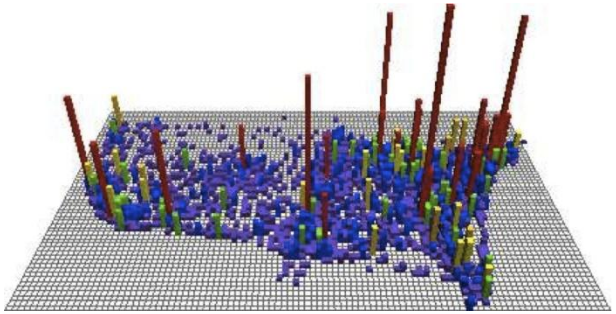


- **Time-series:** Τιμές δεδομένων με χρονική σειρά





- **Spatial data:** δεδομένα σε διάφορες θέσεις.



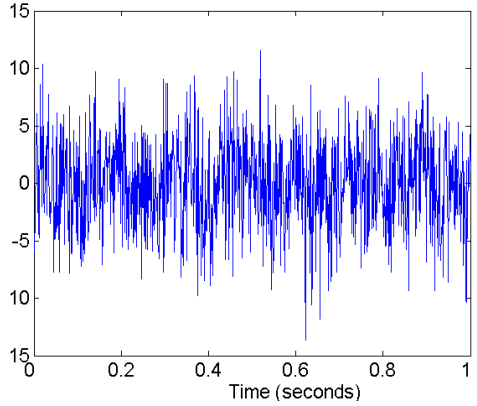
[B]. Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

➢ Data cleaning: detection & correction of data

- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

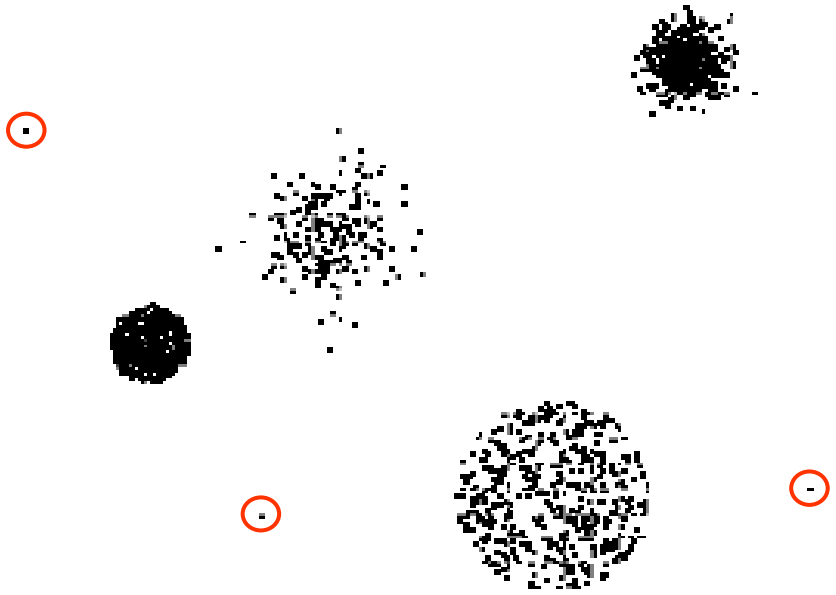
Noise



- ΄Υπαρξη θορύβου στα δεδομένα
- Need of noise extraction methods
- Προβλήματα:
  - Πολλές φορές ο θόρυβος μπορεί να παράγει ακραίες τιμές
  - Σε περιπτώσεις sequential data υπάρχουν «κρυμμένα» patterns εξαιτίας του θορύβου
  - Σε image data ο θόρυβος στο background δυσκολεύει την ανίχνευση αντικειμένων
- Ανάγκη εξαγωγής του θορύβου

Outliers (ακραίες τιμές)

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values (χαμένες τιμές)

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Στρατηγικές
  - Eliminate Data Objects (απλούστερη μεθοδολογία)
  - Estimate Missing Values
    - interpolation – prediction – mean value + noise
  - Ignore the Missing Value During Analysis
  - Ο μηχανισμός αποτελεί μέρος της μεθόδου

Duplicate Data (αντίγραφα)

- Αντίγραφα δεδομένων μπορεί να υπερεκτιμήσουν ή να πολώσουν μια μέθοδο.
- Data cleaning
- Μπορούμε να κάνουμε merge τα αντίγραφα
- Major issue when merging data from heterogeneous sources
  - Examples: Same person with multiple email addresses

[C]. Data Preprocessing

- Είναι ένα πολύ σημαντικό στάδιο το οποίο συχνά επηρεάζει την απόδοση μιας μεθόδου
- Διάφορες στρατηγικές / τεχνικές
  - Επιλογή αντιπροσωπευτικών αντικειμένων ή χαρακτηριστικών
  - Μεταβολή ή μετασχηματισμό των χαρακτηριστικών
- Στόχος: η βελτίωση της ανάλυσης (χρόνος, κόστος και ποιότητα λύσης)
  - Aggregation
  - Sampling
  - Dimensionality Reduction
  - Feature subset selection
  - Feature creation
  - Discretization and Binarization
  - Attribute Transformation

C1. Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More “stable” data
    - Aggregated data tends to have less variability

C2. Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Processing the entire set of data of interest is too expensive or time consuming.

- Cross-validation – generalization

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling χωρίς επανάθεση (without replacement)
  - As each item is selected, it is removed from the population
- Sampling με επανάθεση (with replacement)
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

C3. Discretization (διακριτοποίηση)

- Διακριτοποίηση ή συμβολική αναπαράσταση συνεχών δεδομένων.
- Χρήσιμη σε προβλήματα ταξινόμησης ή ανάλυσης συσχετίσεων
- 2 ζητήματα:
  - Πόσες κατηγορίες (διακριτές τιμές) θα έχουμε?
    - Διάταξη (sort) των τιμών και διαχωρισμό σε m διαστήματα
  - Τρόπος απεικόνισης των διακριτών τιμών?
    - Όλες οι τιμές που ανήκουν στο ίδιο διάστημα έχουν την ίδια διακριτή τιμή.
- Διαστήματα: $[(x_0, x_1), (x_1, x_2), \ldots, (x_{m-1}, x_m)]$

Διακριτοποίηση χωρίς επίβλεψη (unsupervised)

1. Διαστήματα ίσου πλάτους
2. Διαστήματα ίσης συχνότητας (ίσο πλήθος σημείων)
3. Ομαδοποίηση σε m ομάδες και διαμερισμός ανάλογα με την ομάδα που ανήκει ένα σημείο

Διακριτοποίηση με επίβλεψη (supervised)

- Διαθέσιμη η ετικέτα των δεδομένων
- Τα διαστήματα επιλέγονται με βάση την διακριτική ικανότητα και την «καθαρότητα» τους
- Ιεραρχικά
  - Κάθε τιμή ένα διαφορετικό διάστημα
  - Merging γειτονικών διαστημάτων με βάση ένα στατιστικό κριτήριο, π.χ. εντροπία ως μέτρο καθαρότητας διαστήματος

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$

k: number of labels
p_{ij} : relative frequency of label j

C4. Transformation (μετασχηματισμοί)

Είδη μετασχηματισμού ενός χαρακτηριστικού

- Μία συνάρτηση που εφαρμόζεται σε όλες τις τιμές ενός χαρακτηριστικού, π.χ. $x^k, \log x, e^x, \sqrt{x}, \frac{1}{x}, \sin x, |x|$
- Normalization (κανονικοποίηση)
  a) z-score. $\frac{x-\bar{x}}{\sigma_x} \to N(0,1) \qquad x_i \to z_i = \frac{x_i - \bar{x}}{s}$

  δηλ. ισχύει ότι: $x_i = \bar{x} + sz_i$

  b) min-max . $\frac{x-\min(x)}{\max(x)-\min(x)} \to [0,1]$

.

C5. Dimensionality reduction (μείωση διάστασης)

$$R^d \rightarrow R^m \quad , \quad m << d$$

- Αποτελεί ένα σημαντικό ζήτημα σε προβλήματα με πολυδιάστατα δεδομένα.
- Curse of dimensionality (η κατάρα της διάστασης)
  - Υπολογιστική πολυπλοκότητα
  - Μικρή διάσταση = μικρός αριθμός παραμέτρων δηλ. μικρότερος χώρος αναζήτησης
  - Μεγαλύτερη γενικευτική ικανότητα (απλούστερο μοντέλο)
  - Δεδομένα με μικρότερο θόρυβο (πιο «καθαρά»)
  - Μεγάλη διάσταση απαιτεί περισσότερα δεδομένα (μεγάλο κόστος για πολυδιάστατα)

C5. Dimension reduction (συν.)

2 οικογένειες μεθόδων μείωσης διάστασης

A. Feature Extraction (Εξαγωγή χαρακτηριστικών)

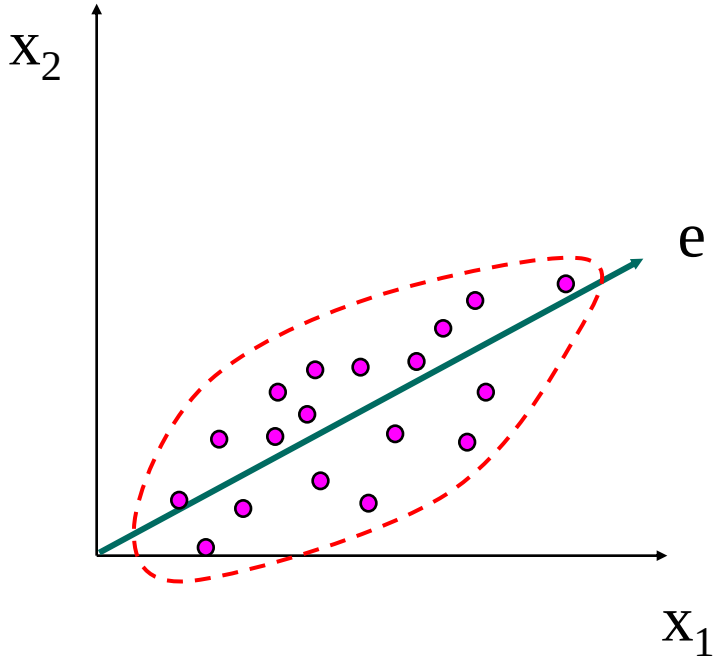
$$R^d_{(x)} \xrightarrow{\Phi} R^m_{(y)} \qquad y = \Phi x$$

B. Feature Selection (Επιλογή χαρακτηριστικών)

$$R^m \subset R^d$$

Principal Component Analysis (PCA)

- Υποθέτει ελλειπτική μορφή των δεδομένων (κανονικότητα)
- Εύρεση αξόνων της έλλειψης
- Επιλογή των πρωτευουσών συνιστωσών (μεγ. Ιδιοτιμές – variance – information)
- Πίνακας Φ τα αντίστοιχα ιδιοδιανύσματα [φ_j]



- Εφαρμογές
  - αναγνώριση εικόνων (eigenfaces)
  - Αναγνώριση ήχου (eigenvoices)

Άλλες τεχνικές μετασχηματισμού

- LDA (Linear Discriminant Analysis)
  - supervised data analysis
  - Στόχος να μεγιστοποιήσουμε την διασπορά μεταξύ των κατηγοριών και να ελαχιστοποιήσουμε την ενδοδιασπορά τους
  - Ο πίνακας που ορίζεται παρέχει τον μετασχηματισμό

- Ορίζουμε m συναρτήσεις βάσης $\varphi_1(x)$, $\varphi_2(x)$,…, $\varphi_m(x)$

  Έτσι τα αρχικά δεδομένα μετασχηματίζονται ως:

$$x \longrightarrow \phi(x) = [\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x)]$$

  παραδείγματα:
  - Fourier transform
  - Wavelets transform
  - Kernel-based transform

Feature Selection

- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
- Irrelevant features
  - contain no information that is useful for the data mining task at hand

Feature Selection (cont.)

- Techniques:
  - Exhaustive :
    - Try all possible feature subsets (για μικρό d)
  - Filter methods:
    - Features are selected before data mining algorithm is run
    - Χρησιμοποιώντας ένα κριτήριο αξιολογήσης εξετάζουμε ένα-ένα τα χαρακτηριστικά και επιλέγουμε τα m καλύτερα.
  - Wrapper methods:
    - Use the data mining algorithm as a black box to find best subset of attributes (*more expensive*)
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm

Feature Selection (cont.)

- 2 στρατηγικές μεθόδων Wrapper:
  - Σταδιακά αυξανόμενου αριθμού χαρακτηριστικών
    - Κάθε φορά προσθέτουμε το καλύτερο από τα εναπομείναντα ενσωματώνοντάς το στο υπάρχον υποσύνολο
  - Σταδιακά μειωμένου αριθμού χαρακτηριστικών
    - Ξεκινώντας με όλα τα d χαρακτηριστικά, σε κάθε βήμα εξάγουμε το χειρότερο που προκύπτει βγάζοντάς το από το σύνολο
- Διάφορα μέτρα αξιολόγησης
  - Σφάλμα ταξινόμησης - Πιθανότητα σφάλματος
  - Διακριτική ικανότητα
  - Εντροπία

[D]. Correlations among data: Similarity and distance

- **Similarity**
  - Numerical measure of how alike two data objects are

$$s\left(x_i, x_j\right)$$

  - Is higher when objects are more alike.
  - Often falls in the range [0,1]

- **Distance**
  - Measure the degree of dissimilarity between two data

$$d\left(x_i, x_j\right)$$

  - Minimum distance = maximum similarity

Απόσταση (Distance)

Συνάρτηση απόστασης $d(x_i, x_j)$

- Συνθήκες που πρέπει να πληρούνται:
  - Θετικότητα $d(x_i, x_j) \geq 0$
  - Συμμετρικότητα $d(x_i, x_j) = d(x_j, x_i)$
  - Ταυτότητα $d(x_i, x_i) = 0$
  - Τριγωνική ανισότητα $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$

Είδη αποστάσεων:

- **Ευκλείδια** (2-norm) $$d(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^2 \right)^{1/2}$$
- Γενικευμένη Ευκλείδια (p-norm ή **Minkowski**) $$d(x_i, x_j) = \left(\sum_{k=1}^{d} (|x_{ik} - x_{jk}|)^p \right)^{1/p}$$
- Σταθμισμένη Ευκλείδια $$d(x_i, x_j) = \left(\sum_{k=1}^{d} w_k (x_{ik} - x_{jk})^2 \right)^{1/2}$$

  $\{ w_k \}$: βάρη συνιστωσών
- Cityblock ή **Manhattan** $$d(x_i, x_j) = \sum_{k=1}^{d} |x_{ik} - x_{jk}|$$

- **Mahalanobis** απόσταση

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)} =$$

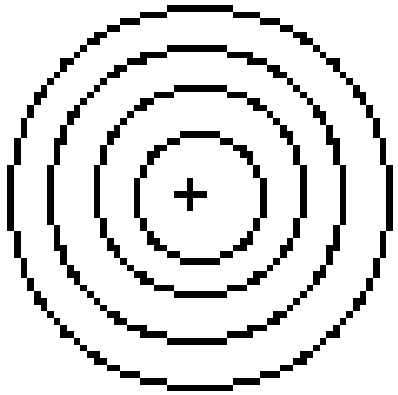
$$= \sqrt{\sum_{k=1}^{d} \sum_{m=1}^{d} (x_{ik} - x_{jk}) \Sigma_{km} (x_{ik} - x_{jm})}$$

όπου Σ είναι ο πίνακας διακυμάνσεων που καθορίζει μία έλλειψη γύρω από τα δεδομένα.

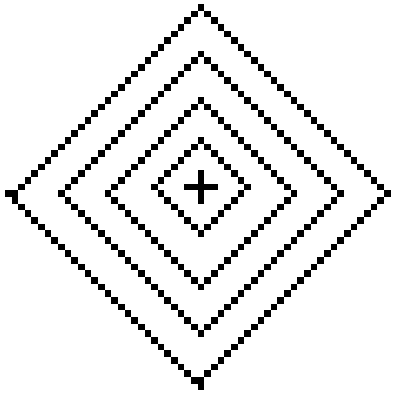
  - Σε περίπτωση όπου $\Sigma = \sigma^2 I$, σφαίρας ακτίνας σ : $d(x_i, x_j) = \frac{\sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}}{\sigma}$

  - Αν $\Sigma = I$ τότε Ευκλείδια απόσταση

- Γεωμετρικές περιοχές χρησιμοποιώντας διαφορετικές συναρτήσεις απόστασης



Euclidean



Manhattan

Mahalanobis

Άλλες αποστάσεις:

- **Hamming** (ποσό ανομοιότητας για διακριτά)

$$d(x_i, x_j) = \sum_{k=1}^{d} \delta(x_{ik}, x_{jk}) \quad \text{όπου} \quad \delta(x, y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

- Συνημιτονοειδής (**cosine**)

$$d(x_i, x_j) = 1 - \frac{x_i^T x_j}{\|x_i\| \|x_j\|} = 1 - \frac{\sum_{k=1}^{d} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{d} x_{ik}^2} \sqrt{\sum_{k=1}^{d} x_{jk}^2}}$$

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Chapter 3

Data Mining: Exploring Data

Οργάνωση και γραφική παράσταση δεδομένων

Στατιστικοί πίνακες

- Παρουσίαση των δεδομένων σε συνοπτικούς πίνακες για την ευκολία της κατανόησής τους και την εξαγωγή συμπερασμάτων.

Πίνακες συχνοτήτων

- Έστω τ.μ. X που περιγράφει τα άτομα ενός πληθυσμού και $\{X_1, X_2, \ldots X_n\}$ ένα τυχαίο δείγμα μεγέθους n. Εστω σύνολο k τιμών της μεταβλητής $\{a_1, a_2, \ldots, a_k\}$ (π.χ. κατηγορίες ή διαστήματα τιμών).

- **Συχνότητα** n_i της τιμής a_i το πλήθος των δειγμάτων με τιμή a_i, και **σχετική συχνότητα**

$$f_i = \frac{n_i}{n} \quad i = 1, \ldots, k$$

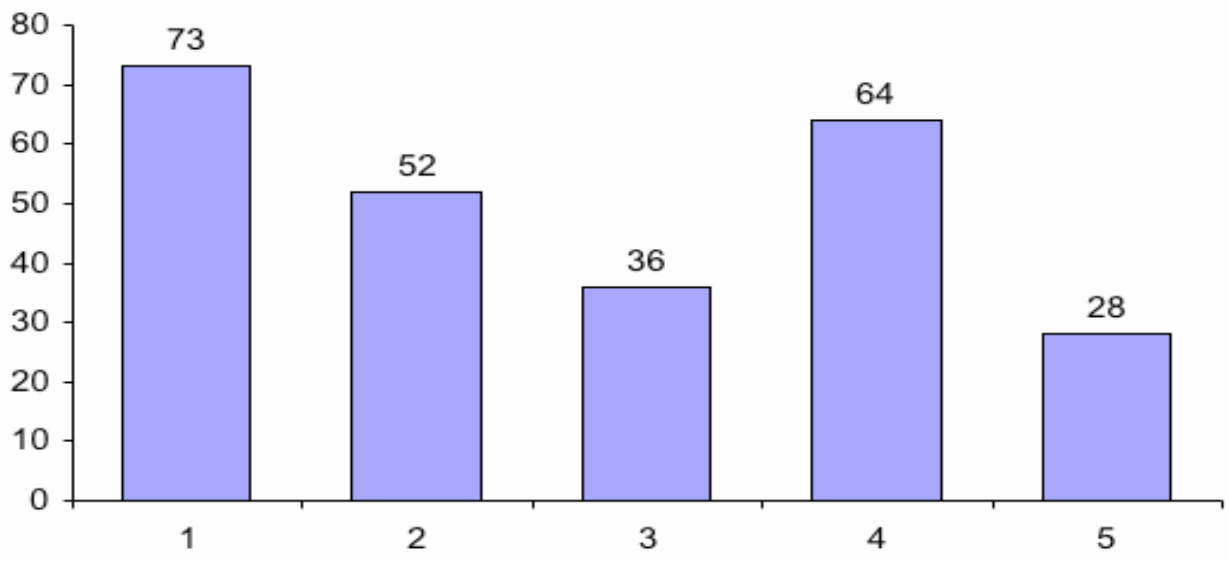
- Ένας **πίνακας συχνοτήτων** είναι ένας πίνακας που παρουσιάζει τις κατηγορίες (ή τα διαστήματα τιμών) των δεδομένων με τις αντίστοιχες συχνότητές τους.

a_i	n_i	f_i
ασπρο	6	6/11
μαύρο	3	3/11
μπλε	2	2/11

a_i	n_i	f_i
1-4	4	4/12
5-8	5	5/12
9-12	3	3/12

Γραφική παράσταση στατιστικών δεδομένων

- **Ραβδογράμματα (*bar charts*)**

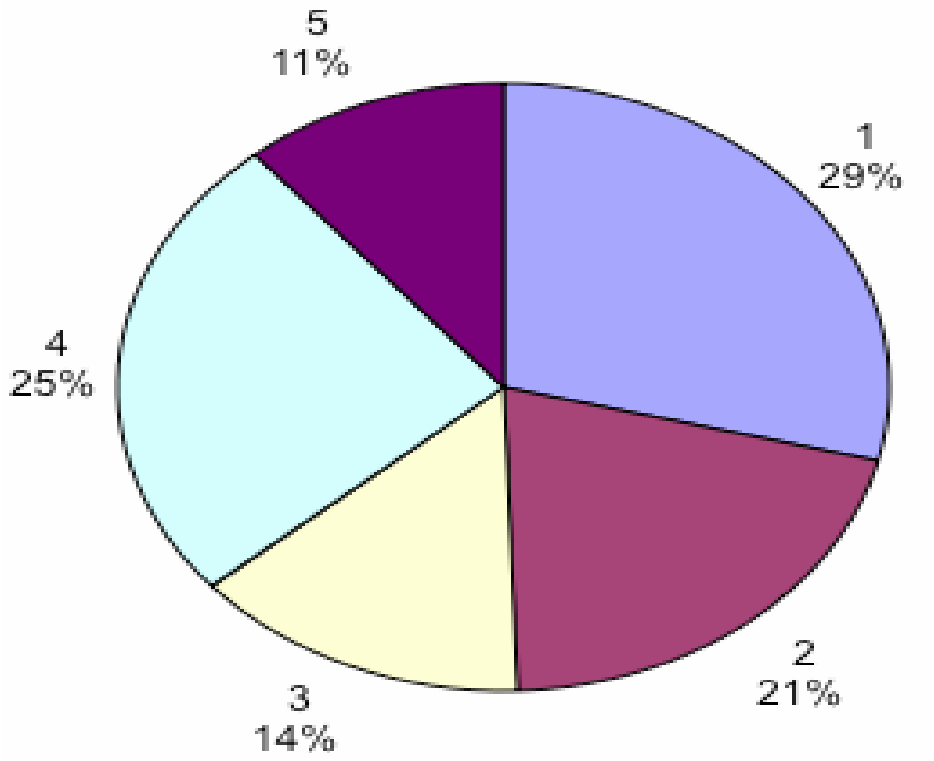


Οι κατηγορίες παρουσιάζονται στον x-άξονα ως ισομήκη διαστήματα ενώ οι αντίστοιχες συχνότητες (ή σχετικές συχνότητες) στο y-άξονα

Είναι δυνατόν να υπάρχουν πολλαπλά ραβδογράμματα

- **Κυκλικά διαγράμματα (*pie charts*)**

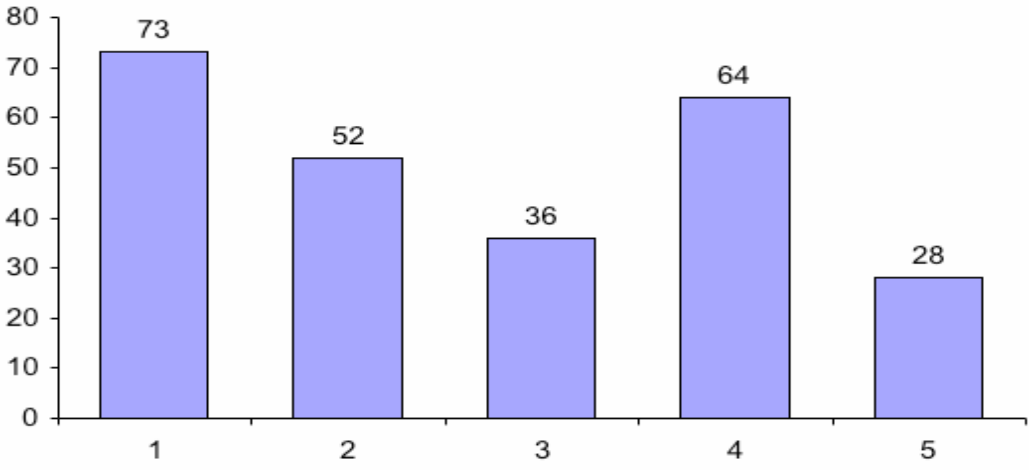
Οι κατηγορίες παρουσιάζονται σε κύκλο χωρισμένο σε κυκλικούς τομείς, τα τόξα των οποίων είναι ανάλογα προς τις αντίστοιχες συχνότητες.

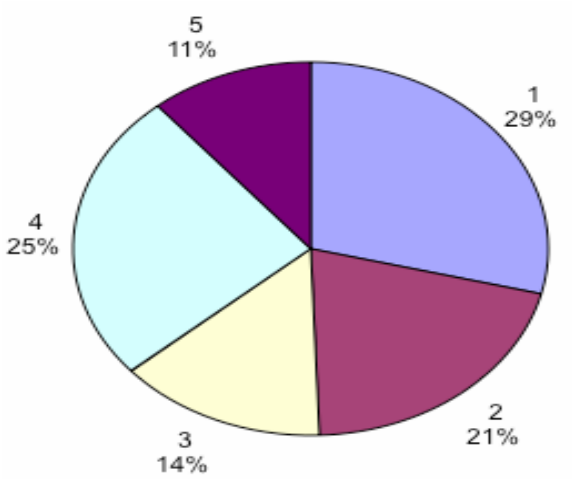


3 διαφορετικοί τρόποι παρουσίασης

a_i	n_i	F_i (%)
1	73	28.9
2	52	20.6
3	36	14.2
4	64	25.3
5	28	11.1
Σύνολο	253	100

Δείχνουν την ίδια πληροφορία (βασίζονται στα ίδια δεδομένα) απλά διαφέρει ο τρόπος παρουσίασης



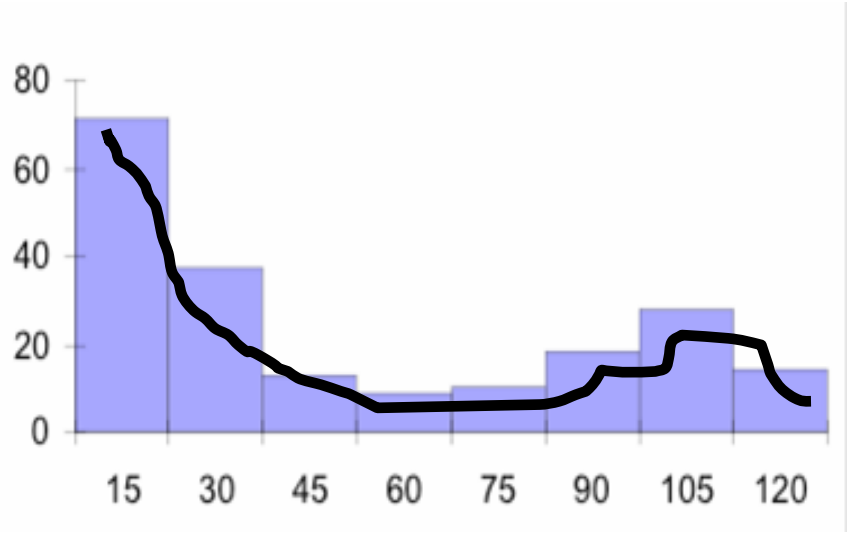


- **Ιστογράμματα (*Histograms*)**

$$P(X \in A) = \frac{k_A}{n}$$

$$P(X \in A) = \int_A f(x)dx = f(x)\int_A dx = f(x)V_A \Rightarrow f(x) = \frac{k_A(x)}{nV_A}$$

0 to 15	71
15 to 30	37
30 to 45	13
45 to 60	9
60 to 75	10
75 to 90	18
90 to 105	28
105 to 120	14
Total	200

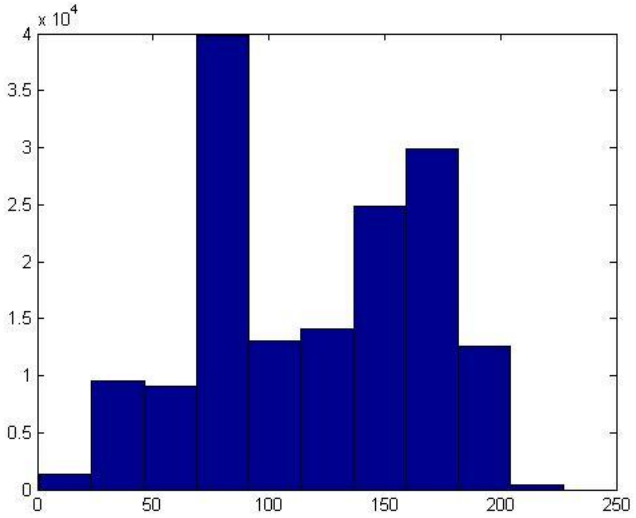


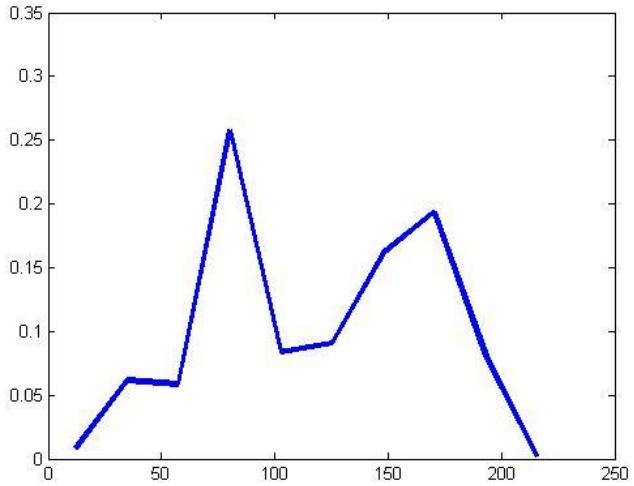
Παράγει μια κατανομή της συχνότητας των δεδομένων σε (συνήθως ίσα) διαστήματα τιμών (ή κάδους - bins). Χρησιμοποιώντας τις σχετικές συχνότητες στο ιστόγραμμα τότε κατασκευάζουμε μία κατανομή των δεδομένων

Παραδείγματα με Ιστογράμματα

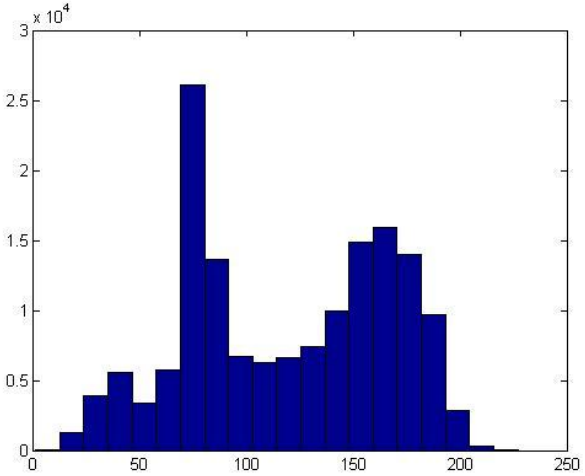
- **Εικόνα 1**

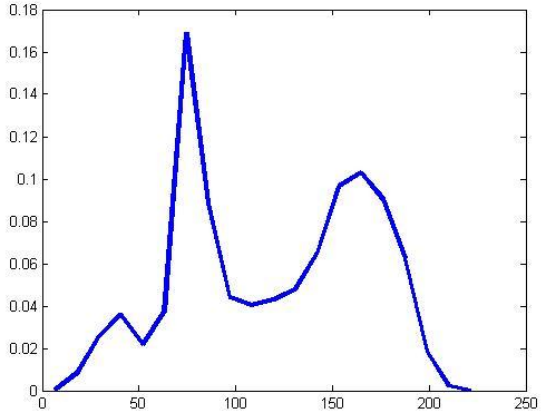
- **Ιστόγραμμα με 10 bins**



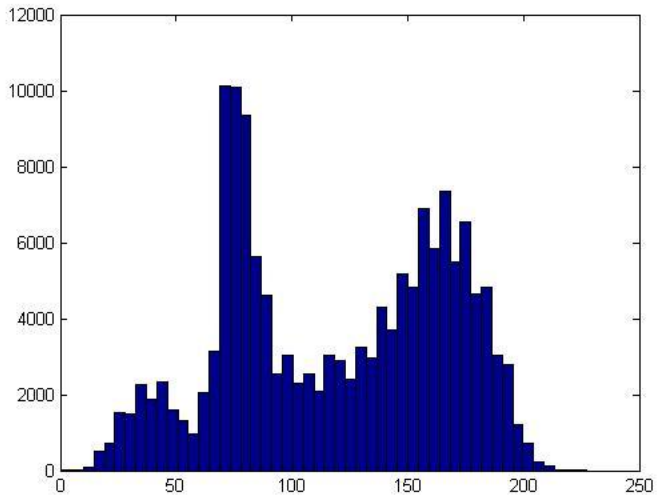


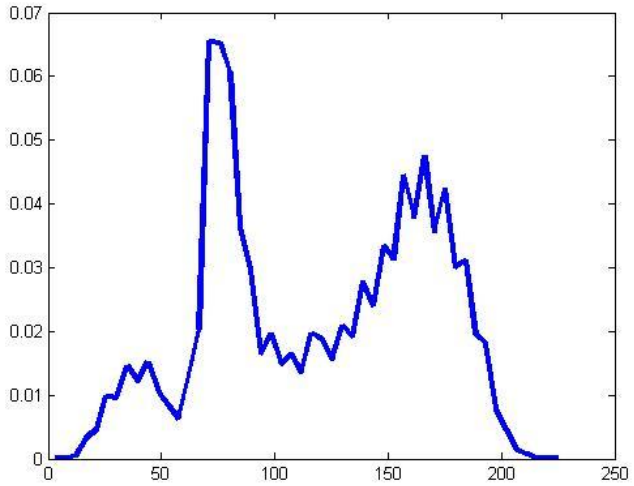
- **Ιστόγραμμα με 20 bins**



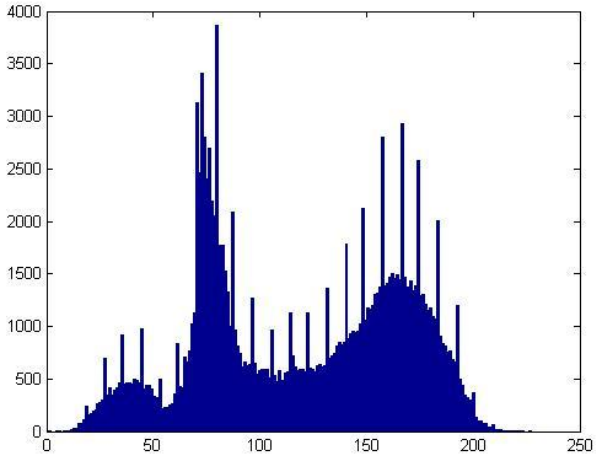


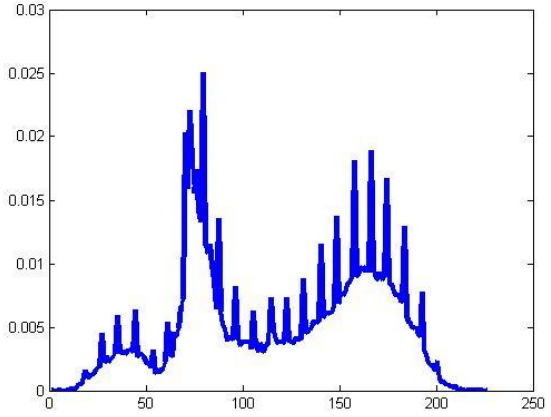
- **Ιστόγραμμα με 50 bins**





- **Ιστόγραμμα με 200 bins**

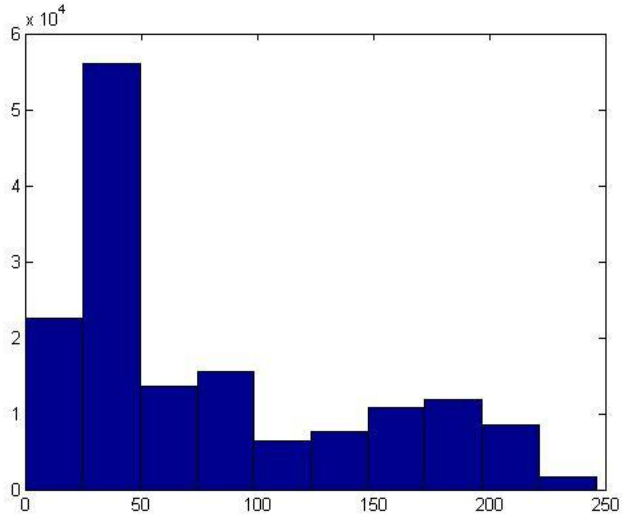


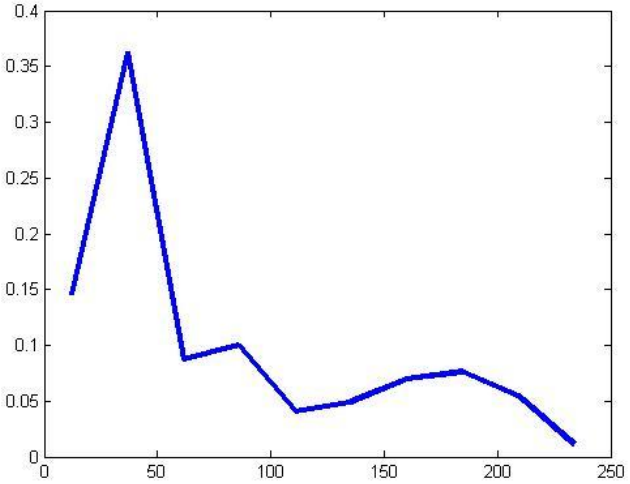


Παραδείγματα με Ιστογράμματα

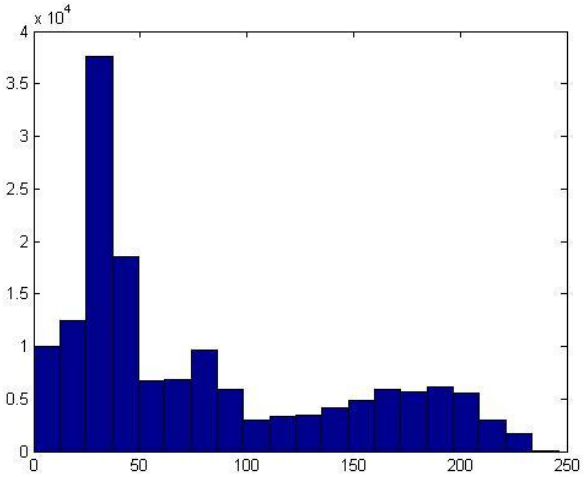
- **Εικόνα 2**

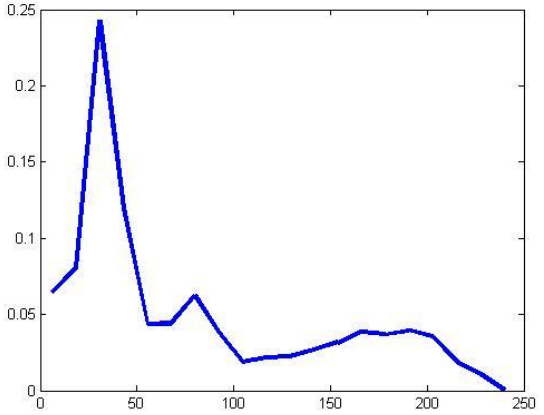
- **Ιστόγραμμα με 10 bins**



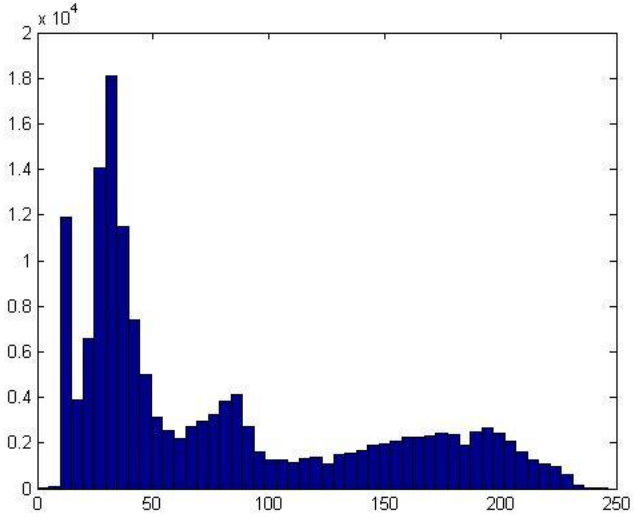


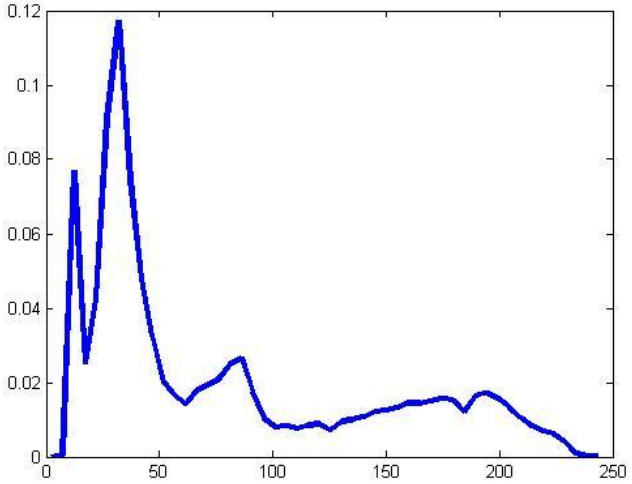
- **Ιστόγραμμα με 20 bins**



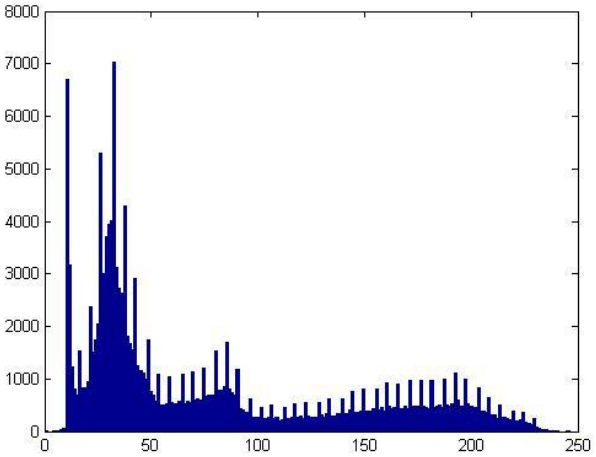


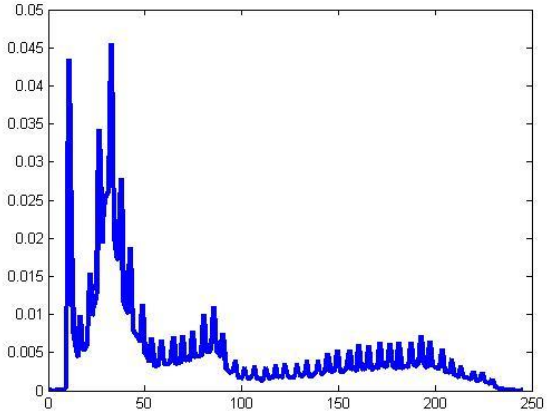
- **Ιστόγραμμα με 50 bins**





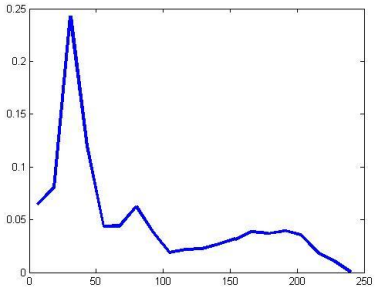
- **Ιστόγραμμα με 200 bins**



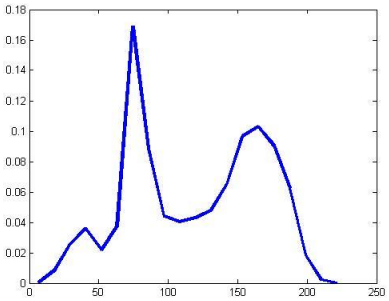


Σύγκριση εικόνων με Ιστογράμματα (20 bins)

- **Εικόνα 1**



- **Εικόνα 1**



- **cosine similarity : 8.3145e-04**
- **Euclidean distance : 0.2880**
- **Manhattan distance : 0.9868**

- **Φυλλογράμματα (*stem-leaf notes*)**

Διατηρεί τις πληροφορίες σχετικά με τις ατομικές παρατηρήσεις (χάνονται με τα ιστογράμματα)

Κάθε παρατήρηση χωρίζεται σε 2 μέρη:

- ένα **στέλεχος ή οδηγό** (*stem*) και
- ένα **φύλλο** (*leaf*)

Υπάρχουν διάφοροι τρόποι διάσπασης (συνήθως *αυθαίρετα*) ανάλογα με τον τύπο δεδομένων

Βήματα κατασκευής φυλλογραμμάτων

a) Επιλέγουμε πρώτα τα στελέχη (ή οδηγούντα ψηφία) και τα φύλλα
b) Καταγράφουμε τα στελέχη και τα φύλλα
c) Διατάσσουμε τα στελέχη κατά αύξουσα
d) Γράφουμε τα (*διαφορετικά*) φύλλα στην ίδια γραμμή των αντίστοιχων στελεχών
e) Ελέγχουμε εάν έχουν καταγραφεί όλα τα φύλλα (αριθμός τους ίσος με το συνολικό πλήθος παρατηρήσεων)

Παράδειγμα

Σύνολο δεδομένων

$$X=\{136, 111, 120, 105, 113, 116, 99, 110\}$$

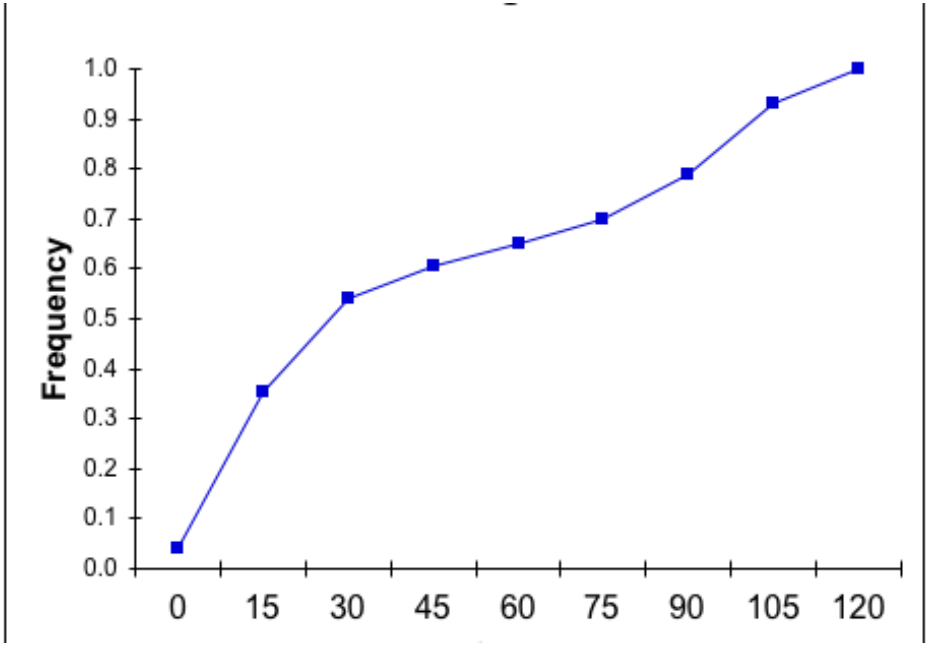
- Θεωρώντας τις **δεκάδες** ως στελέχη και τις **μονάδες** ως φύλλα τότε κατασκευάζουμε το παρακάτω φυλλόγραμμα:

Stem	*Leaf*
9	9
10	5
11	0 1 3 6
12	0
13	6

- **Καμπύλη S**

Γράφημα των αθροιστικών σχετικών συχνοτήτων

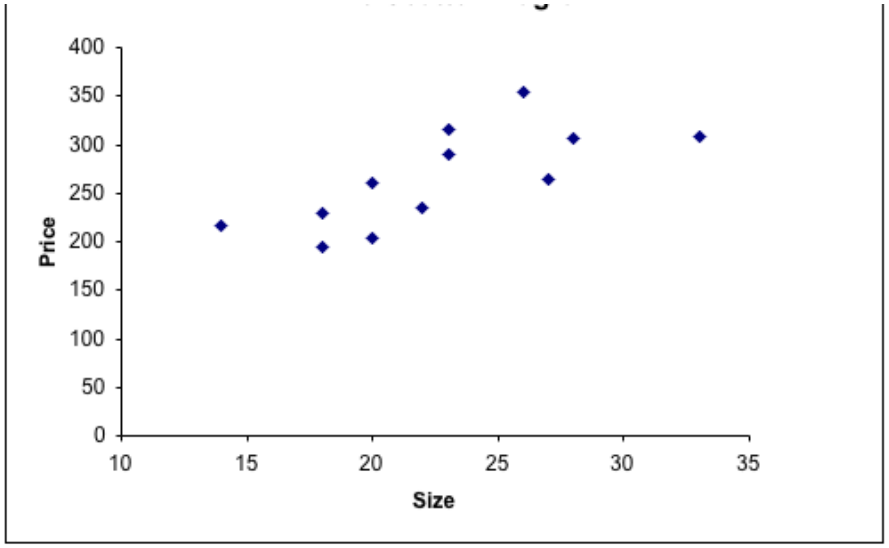
Ανάλογο με την συνάρτηση κατανομής πιθανότητας



Διάγραμμα διασποράς (*scatter plot*)

- Χρησιμοποιείται για να δείξουμε την σχέση ανάμεσα σε 2 μεταβλητές.

- Η ανεξάρτητη μεταβλητή συμβολίζεται με Χ και συνήθως βρίσκεται στον οριζόντιο άξονα, ενώ η άλλη μεταβλητή καλείται εξαρτημένη και παριστάνεται με Υ στον κάθετο άξονα.



Αριθμητικά περιγραφικά μέτρα

Μέτρα θέσης ή κεντρικής τάσης (*central tendency*)

- Χρήσιμα για την περιγραφή της θέσης της κατανομής ή του κέντρου των παρατηρήσεων. Δημοφιλέστερα: *Μέση τιμή*, *κορυφή* και *διάμεσος*.

- Ο **δειγματικός μέσος (*mean*)** $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

  δηλ. ο μέσος όρος των *n* παρατηρήσεων

  Αν a_j είναι οι κεντρικές τιμές των κλάσεων σε ομαδοποιημένα δεδομένα τότε ο μέσος υπολογίζεται ως:

  (σταθμισμένος μέσος): $\bar{x} = \frac{\sum_{j=1}^{k} n_j a_j}{n} = \sum_{j=1}^{k} f_j a_j$

- Η **κορυφή (*mode*) ή επικρατούσα τιμή** είναι η επικρατέστερη τιμή του δείγματος, δηλ. αυτή με την μέγιστη συχνότητα

- Η **διάμεσος (*median*) δ** χωρίζει το δείγμα σε 2 ίσα μέρη, ώστε ο αριθμός των παρατηρήσεων που είναι ≤ *δ* να είναι ίσος με τον αριθμό των δεδομένων που είναι ≥ *δ*. Έτσι αν διατάσουμε τις *n* παρατηρήσεις και συμβολίσουμε ως

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n-1)} \le x_{(n)}$$

τότε

$$\delta = \begin{cases} x_{(r)} & \alpha\nu \quad n = 2r-1 \\ \dfrac{x_{(r)} + x_{(r+1)}}{2} & \alpha\nu \quad n = 2r \end{cases}$$

Ποσοστημόρια *(quantiles)*: μέτρο σχετικής θέσης

- Γενίκευση της διαμέσου (*a=50*)
- το ***a-οστό ποσοστημόριο*** είναι η τιμή από την οποία το *a%* των τιμών είναι μικρότερο και το *(100 - a)%* είναι μεγαλύτερο από την τιμή αυτή
- Αν το a είναι ακέραιο *{1, 2, …, 99}* τότε **εκατοστημόρια** (***quantiles***).
- Αν *a={10, 20, …, 90}* **δεκατημόρια**
- Αν *a={25, 50, 75}* τότε έχουμε **τεταρτημόρια (*quartiles*)**

  a=25 : Q_1 πρώτο τεταρτημόριο

  a=75 : Q_3 τρίτο τεταρτημόριο

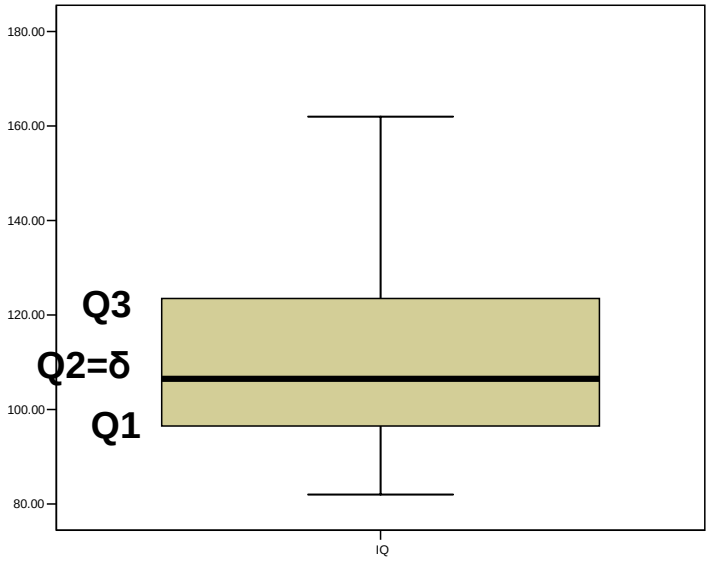
  a=50 : Q_2 δεύτερο τεταρτημόριο (δηλ. η **διάμεσος**).

Θηκογράμματα (*box plots*)

- Απλός τρόπος παρουσίασης των κυριότερων χαρακτηριστικών μιας κατανομής μέσω ενός γραφήματος

Βήματα κατασκευής

1. Αρχικά βρίσκουμε τα δύο τεταρτημόρια ***Q1***, ***Q3*** και την διάμεσο ***δ*** **(***δηλ. το* $\boldsymbol{Q_2}$**)**.
2. Κατασκευάζουμε ένα **ορθογώνιο** με την κάτω πλευρά στο *Q1*, την πάνω πλευρά στο *Q3*. Η διάμεσος παριστάνεται ως ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.



Θηκογράμματα (*συν.*)

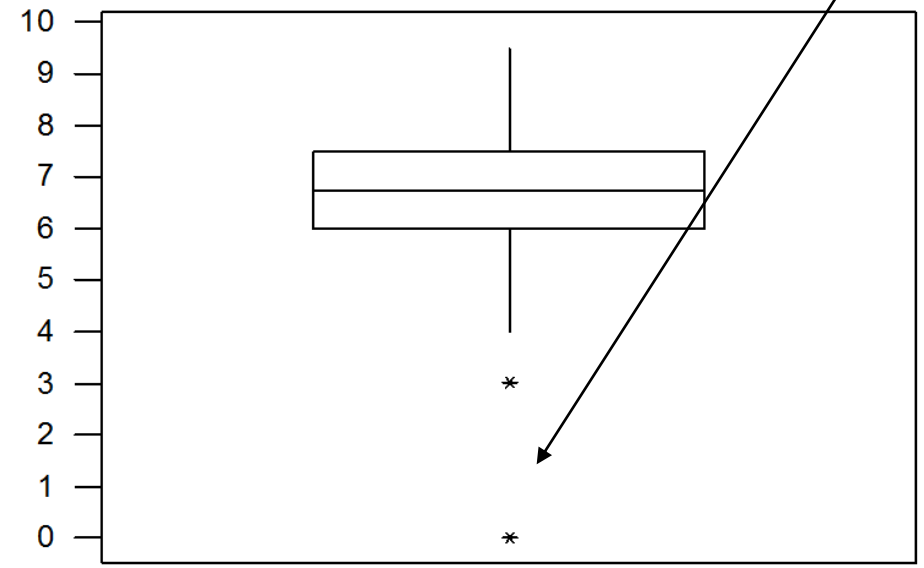
3. Φέρουμε διακεκομμένες **γραμμές** από τα μέσα των βάσεων του ορθογωνίου μέχρι τις ***οριακές τιμές (adjacent)*** που προκύπτουν:

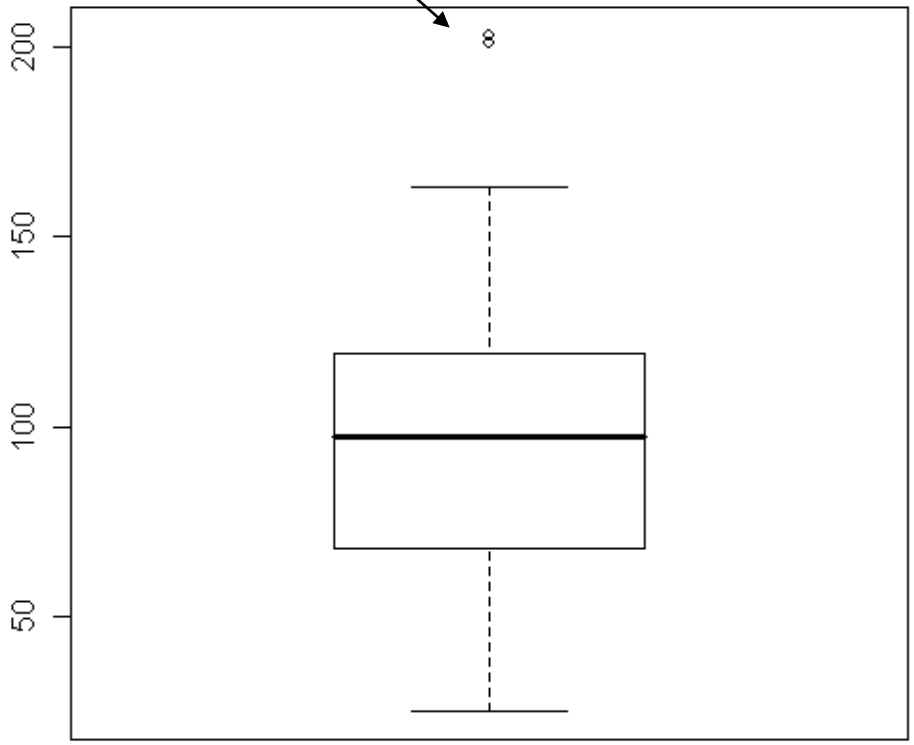
   Άνω οριακή: η μεγαλύτερη παρατήρηση που είναι ≤ του Q3+1.5(Q3-Q1)=Q3+3Q , όπου Q=(Q3-Q1)/2

   Κάτω οριακή: η μικρότερη παρατήρηση που είναι ≥ του Q1-1.5(Q3-Q1)=Q1-3Q

4. Κάθε σημείο που πέφτει έξω από το εύρος των δύο οριακών τιμών λέγονται **ακραία τιμή (*outlier*)** και παριστάνεται με κάποιο ιδιαίτερο σύμβολο (π.χ. *)

ακραίες τιμές

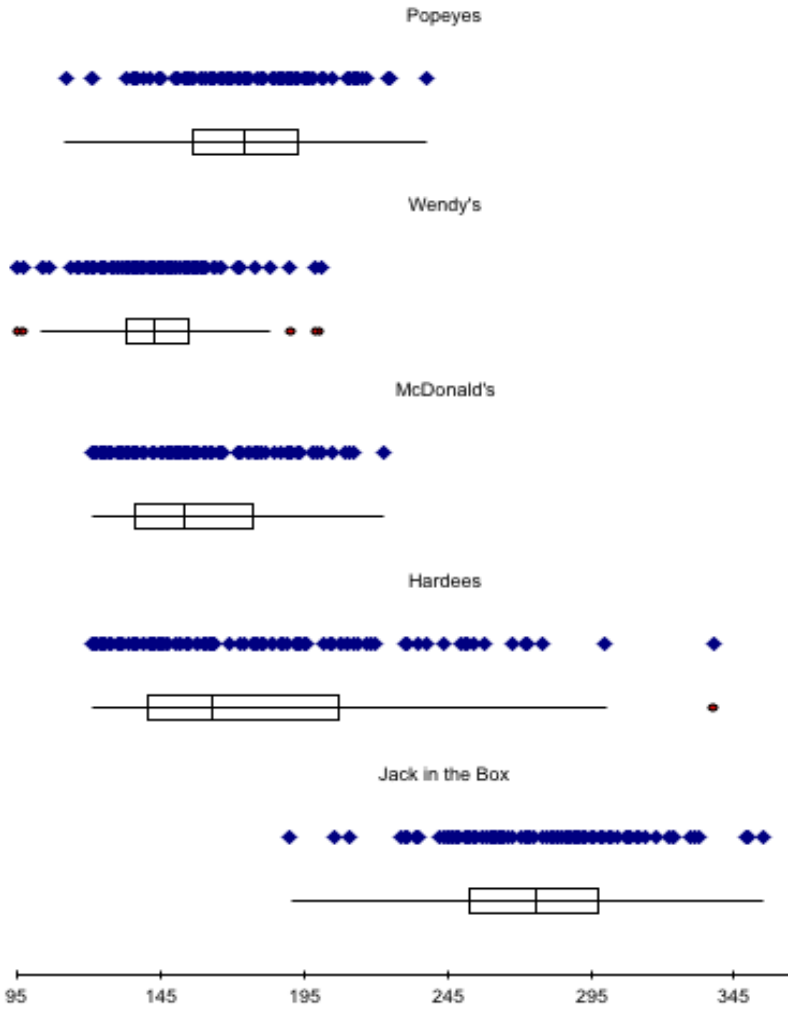




- Τα θηκογράμματα μας δίνουν το κεντρικό διάστημα με το 50% των παρατηρήσεων – μεταξύ του 1ου και 3ου τεταρτημορίου (Q1, Q3).

- Οι επεκτεινόμενες γραμμές και η θέση της διαμέσου μας δίνουν μια εικόνα της συμμετρικότητας της κατανομής.

- Δυνατότητα μελέτης των ακραίων τιμών



Μέτρα διασποράς

Εκφράζουν αποκλίσεις των τιμών μιας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης

- **Εύρος (*range*)** τιμών ή κύμανση = $max\{ x_i \} – min\{ x_i \}$

  εύκολο στον υπολογισμό, αλλά μικρής αξιοπιστίας

- **Ενδοτεταρτημοριακή απόκλιση (*interquartile range*)** = $Q_3 – Q_1$

  Μετράει το άπλωμα του 50% των μεσαίων τιμών των παρατηρήσεων.

  Μεγάλες τιμές αυτής της στατιστικής σημαίνει ότι το 1ο και 3ο τεταρτημόριο απέχουν σημαντικά υποδεικνύοντας υψηλό επίπεδο μεταβλητότητας.

  Έτσι όσο μικρότερο είναι το διάστημα αυτό, τόσο μεγαλύτερη θα είναι η συγκέντρωση τιμών και άρα μικρότερη διασπορά τιμών.

- **Μέση απόκλιση (*mean deviation*)** $MO = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

  Ο αριθμητικός μέσος των αποκλίσεων των τιμών από το μέσον τους

- **Δειγματική διασπορά ή διακύμανση (*variance*)** $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

  στον παρονομαστή έχουμε *n-1* (και όχι *n*) για καλύτερη εκτίμηση του

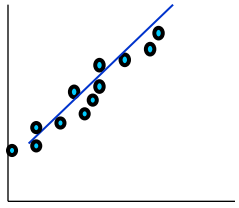
- **Δειγματική τυπική απόκλιση (*standard deviation*)** $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

Μέτρα συσχέτισης δύο μεταβλητών

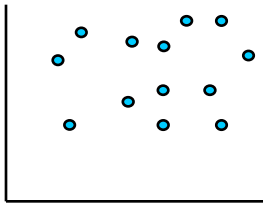
- **Συνδιακύμανση (*covariance*):** μέτρο κατευθυντικότητας
  - Αν οι 2 τ.μ. κινούνται προς την ίδια κατεύθυνση τότε συνδιακύμανση μεγάλη θετική
  - Αν κατεύθυνση αντίθετη τότε μεγάλη αρνητική, ενώ αν όχι σχέση τείνει στο μηδέν.

$$\text{Sample covariance} = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]$$

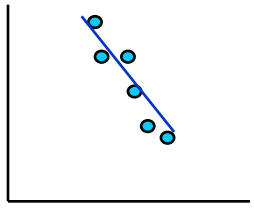
- **Συντελεστής συσχέτισης (*correlation coefficient*):** μέτρο γραμμικότητας μεταξύ των δύο μεταβλητών



r -> 1



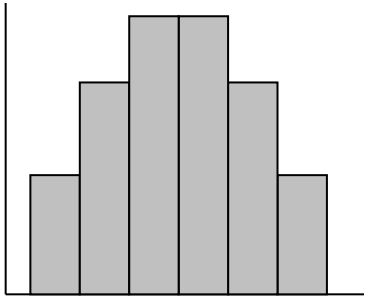
r -> 0



r -> -1

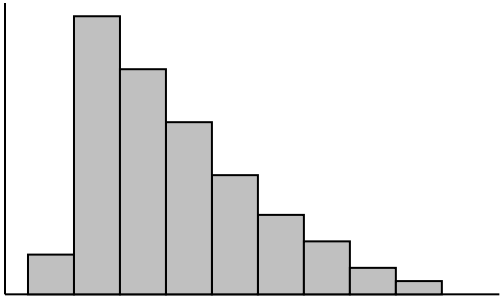
$$r = \frac{S_{xy}}{S_x S_y} \quad \in [-1, 1]$$

Μέτρα ασυμμετρίας



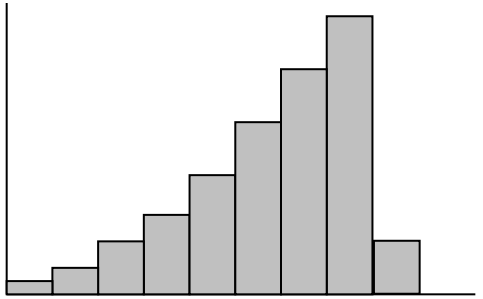
$\bar{x} = \delta = M_0$

Συμμετρική κατανομή
Η κορυφή, ο μέσος και η διάμεσος συμπίπτουν



$M_0 < \delta < \bar{x}$

Θετική συμμετρία
Οι περισσότερες παρατηρήσεις είναι δεξιά της κορυφής (M0).



$\bar{x} < \delta < M_0$

Αρνητική συμμετρία
Οι περισσότερες παρατηρήσεις είναι αριστερά της κορυφής (M0).

- **Συντελεστής ασυμμετρίας *Pearson***

$$Y_1 = \frac{\bar{x} - M_0}{s} \quad Y_2 = \frac{3(\bar{x} - \delta)}{s}$$

Αν *Y=0* => συμμετρία

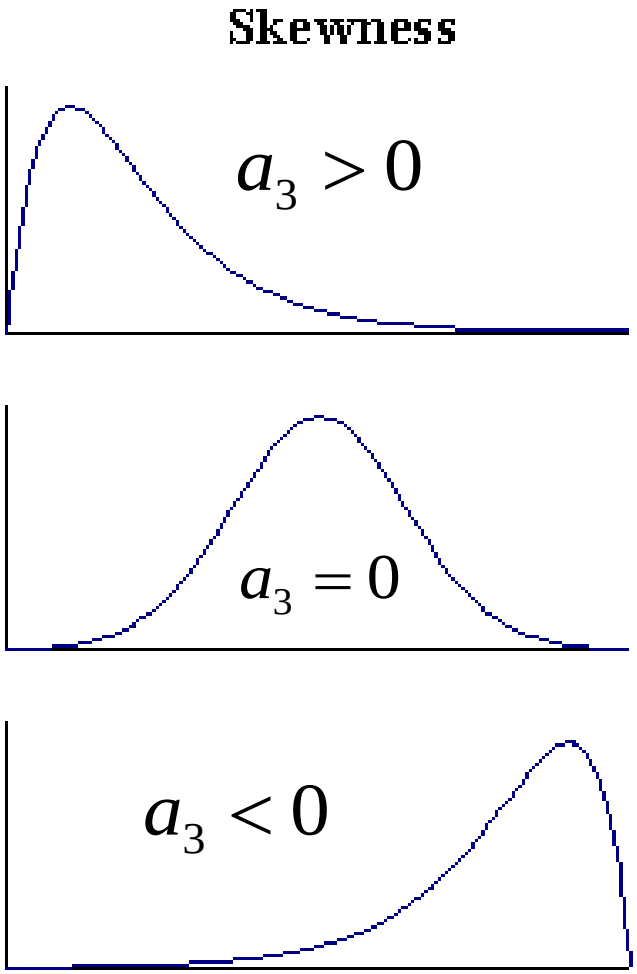
Αν *Y<0* => αρνητική συμμετρία

Αν *Y>0* => θετική συμμετρία

- **Λοξότητα (skewness):**

βαθμός **συμμετρίας** της κατανομής

$$a_3 = \frac{\sum_x (x - \bar{x})^3}{s_x^3}$$

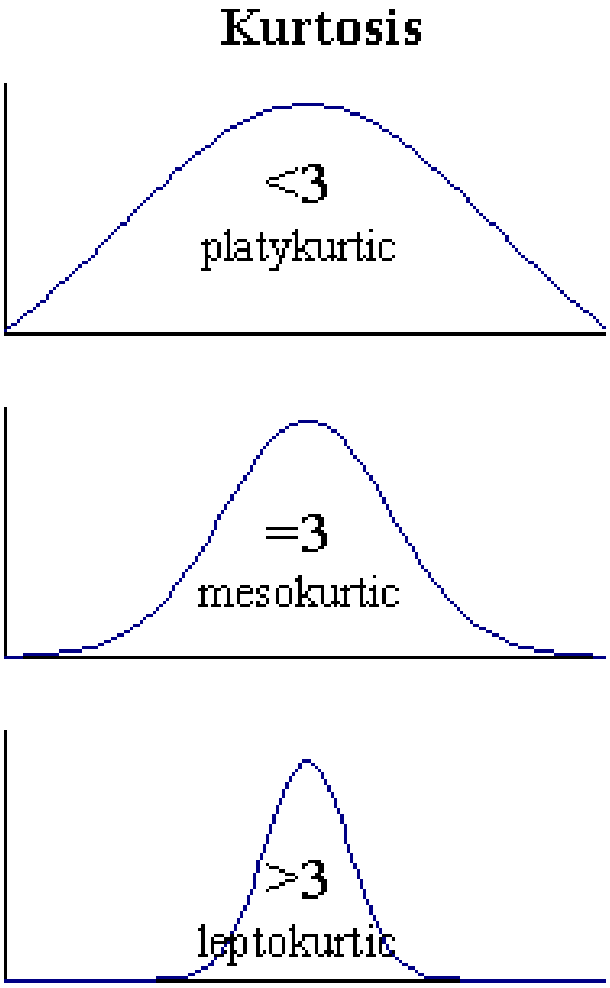


- **Κυρτότητα (kurtosis)**

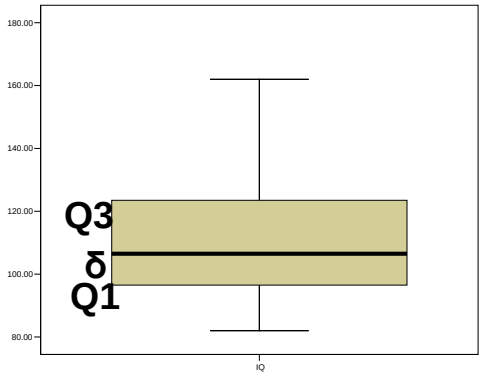
βαθμός **αιχμηρότητας** της κατανομής

$$a_4 = \frac{\sum_x (x - \bar{x})^4}{s_x^4}$$

Στη περίπτωση της **κανονικής** κατανομής ισχύει ότι: $\alpha_4 = 3$



- **Συντελεστής του *Bowley***



$$S_A = \frac{Q_1 + Q_3 - 2\delta}{Q_3 - Q_1} = \frac{(Q_3 - \delta) - (\delta - Q_1)}{Q_3 - Q_1} \quad \in [-1, +1]$$

Αν $S_A = 0$ => συμμετρία

Αν $-1 < S_A < 0$ => αρνητική συμμετρία (η διάμεσος πλησιάζει το Q3)

Αν $0 < S_A < 1$ => θετική συμμετρία (η διάμεσος πλησιάζει το Q1)