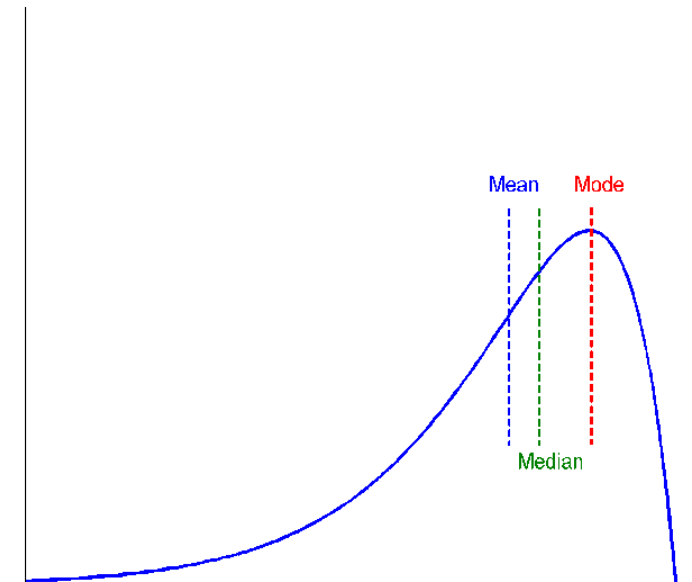
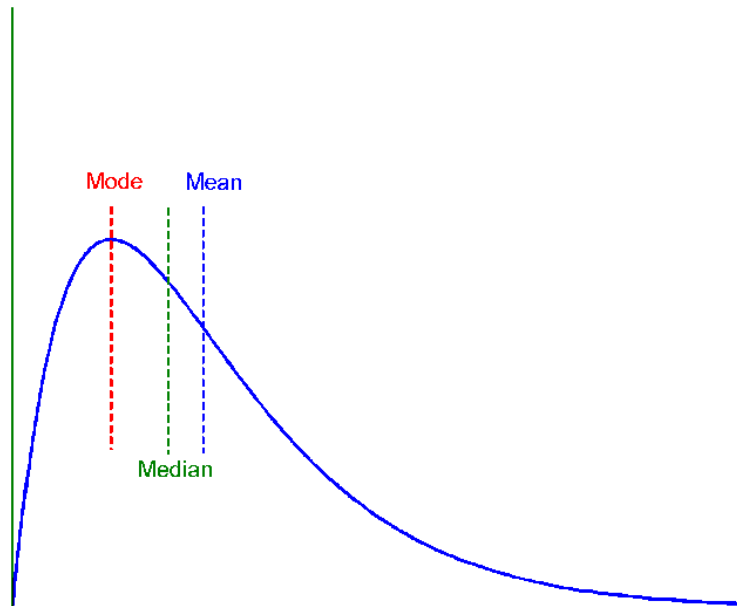
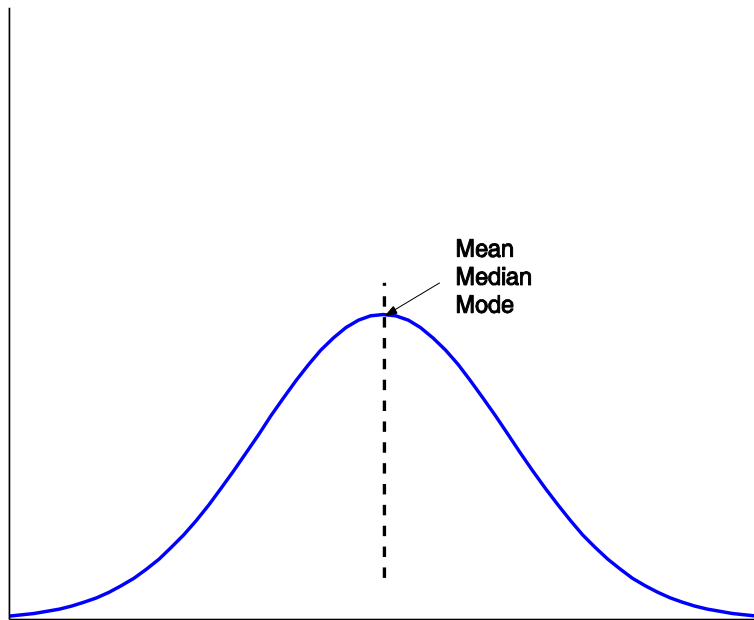


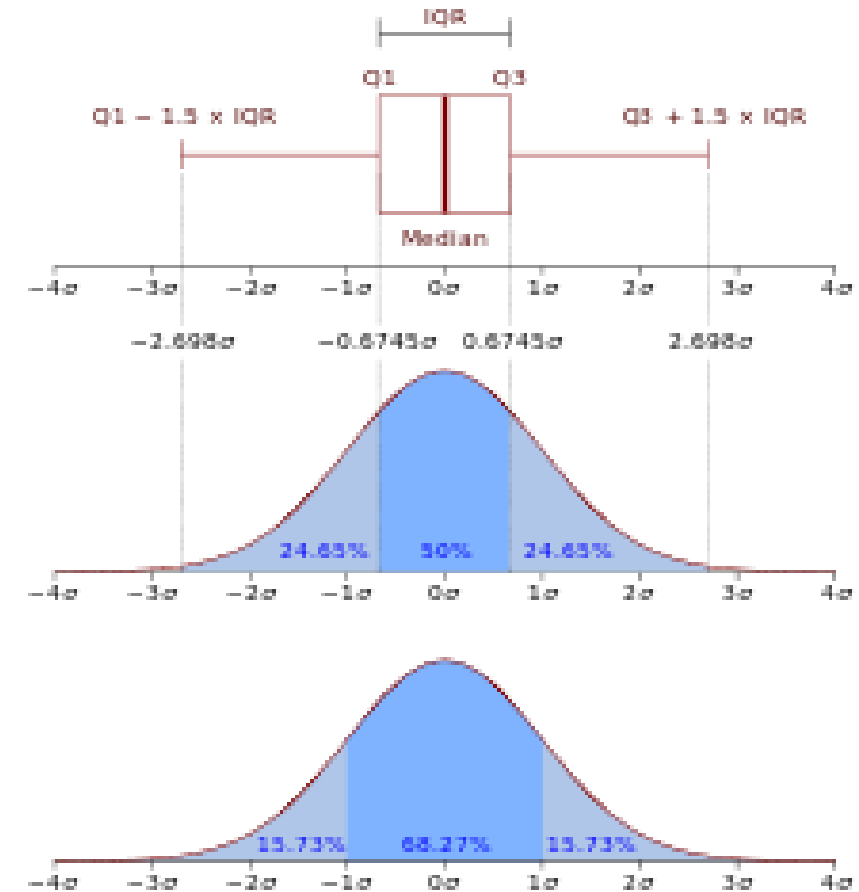
# Συμμετρία

- Συσχέτιση διαμέσου, μέσης τιμής και επικρατούσας τιμής σε συμμετρικά δεδομένα και περιπτώσεις κύρτωσης της κατανομής



# Διασπορά

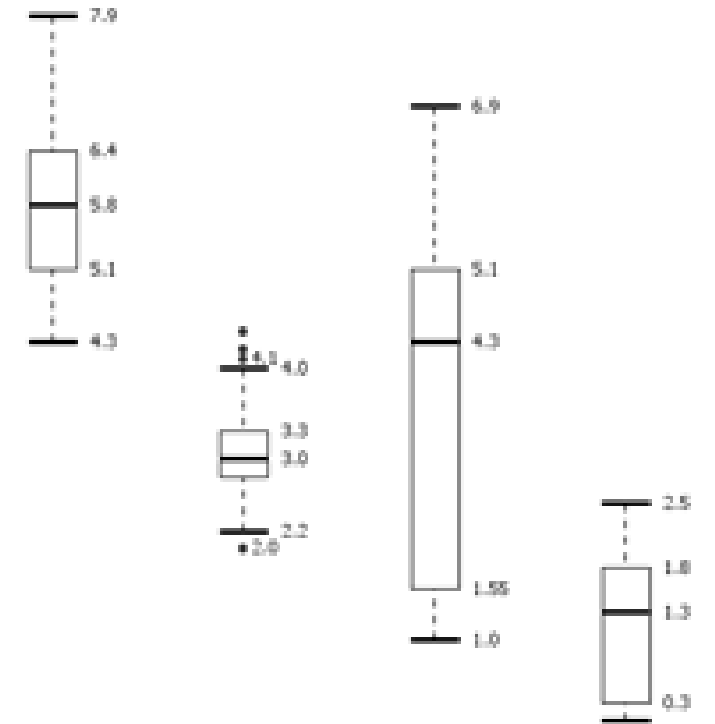
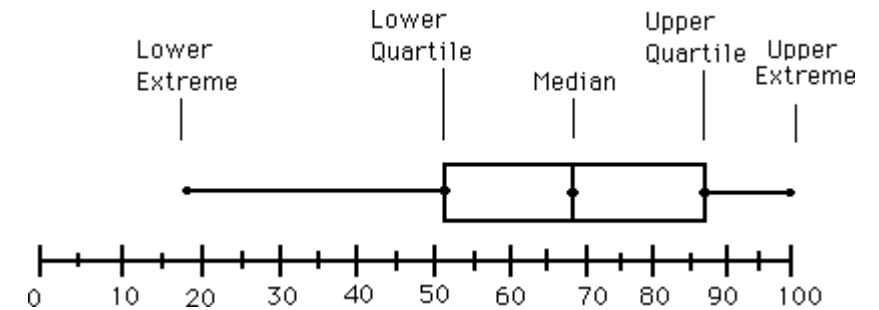
- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range:  $IQR = Q3 - Q1$
  - 5 αριθμοί: min, Q1, median, Q3, max
  - Boxplot: τα τελευταία σημεία είναι τα quartiles; Η διάμεσος μαρκάρεται; add whiskers, and plot outliers individually
  - Ανωμαλία (Outlier): usually, a value higher/lower than  $1.5 \times IQR$
- Διακύμανση και τυπική απόκλιση (δείγμα: s, πληθυσμός:  $\sigma$ )
  - Διακύμανση: (algebraic, scalable computation)
  - Τυπική απόκλιση s (or  $\sigma$ ) είναι η ρίζα της διακύμανσης  $s^2$  (or  $\sigma^2$ )



$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

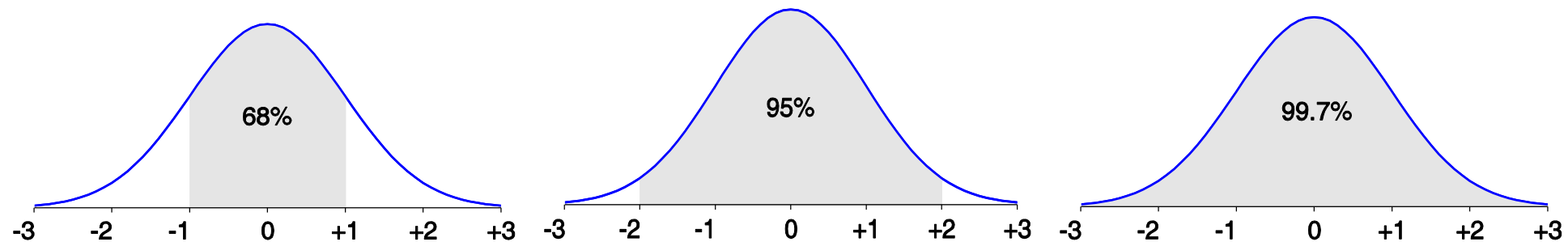
# Boxplot

- Πέντε αριθμοί καθορίζουν την κατανομή των δεδομένων
  - Ελάχιστη τιμή (Minimum), 1<sup>ο</sup> τεταρτημόριο (Q1), Διάμεσος (Median), 3<sup>ο</sup> τεταρτημόριο (Q3), Μέγιστη τιμή (Maximum)
- Boxplot
  - Τα δεδομένα αναπαριστώνται με ένα ορθογώνιο (box)
  - Οι άκρες του ορθογωνίου ορίζουν το 1<sup>ο</sup> και 3<sup>ο</sup> τεταρτημόριο, συνεπώς το ύψος του είναι το IQR
  - Η διάμεσος απεικονίζεται με μια γραμμή μέσα στο ορθογώνιο
  - Whiskers: οι δύο γραμμές εκτός του ορθογωνίου δείχνουν τη μέγιστη και την ελάχιστη τιμή
  - Outliers: είναι σημεία πέρα από ένα όριο και απεικονίζονται χωριστά



# Παράδειγμα

- Κανονική κατανομή
  - Από  $\mu - \sigma$  έως  $\mu + \sigma$ : περιλαμβάνονται το 68% περίπου των μετρήσεων ( $\mu$ : μέση τιμή,  $\sigma$ : τυπική απόκλιση)
  - Από  $\mu - 2\sigma$  έως  $\mu + 2\sigma$ : περιλαμβάνονται το 95% των τιμών
  - Από  $\mu - 3\sigma$  έως  $\mu + 3\sigma$ : περιλαμβάνονται το 99.7% των τιμών

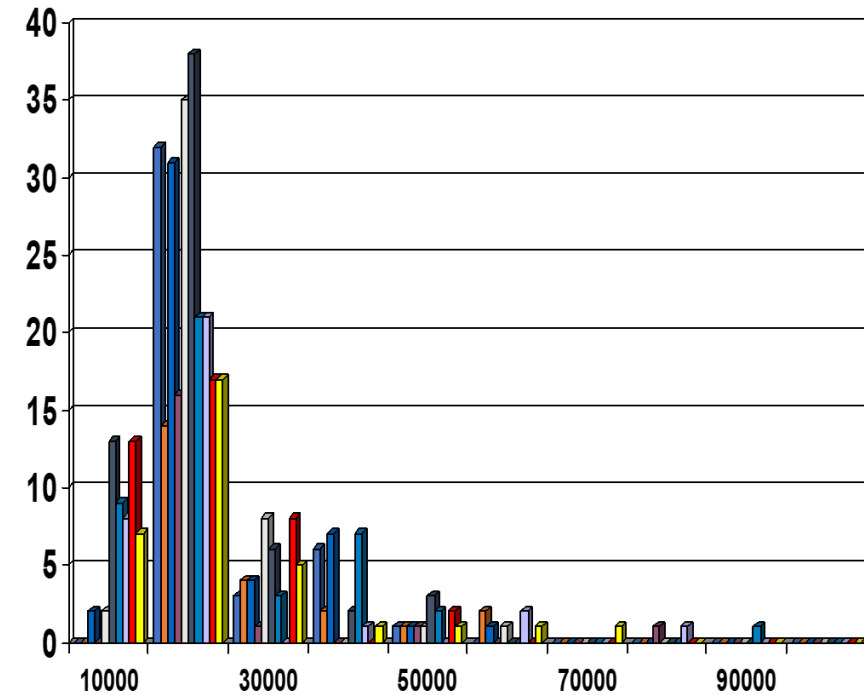


# Οπτικοποίηση

- Boxplot: γραφική αναπαράσταση των πέντε αριθμών που περιγράφηκαν
- Ιστόγραμμα (Histogram): ο x-άξονας απεικονίζει τις τιμές, ο y-άξονας απεικονίζει τις συχνότητες αυτών
- Quantile plot: κάθε τιμή  $x_i$  ταιριάζει την  $f_i$  που απεικονίζει ότι περίπου  $100 f_i \%$  των δεδομένων είναι μικρότερα ή ίσα από την  $x_i$
- Quantile-quantile (q-q) plot: απεικονίζει τα τεταρτημόρια μιας κατανομής σε σύγκριση με τα τεταρτημόρια μιας άλλης
- Scatter plot: κάθε ζεύγος τιμών είναι ένα ζεύγος συντεταγμένων που απεικονίζονται σαν σημεία στο χώρο

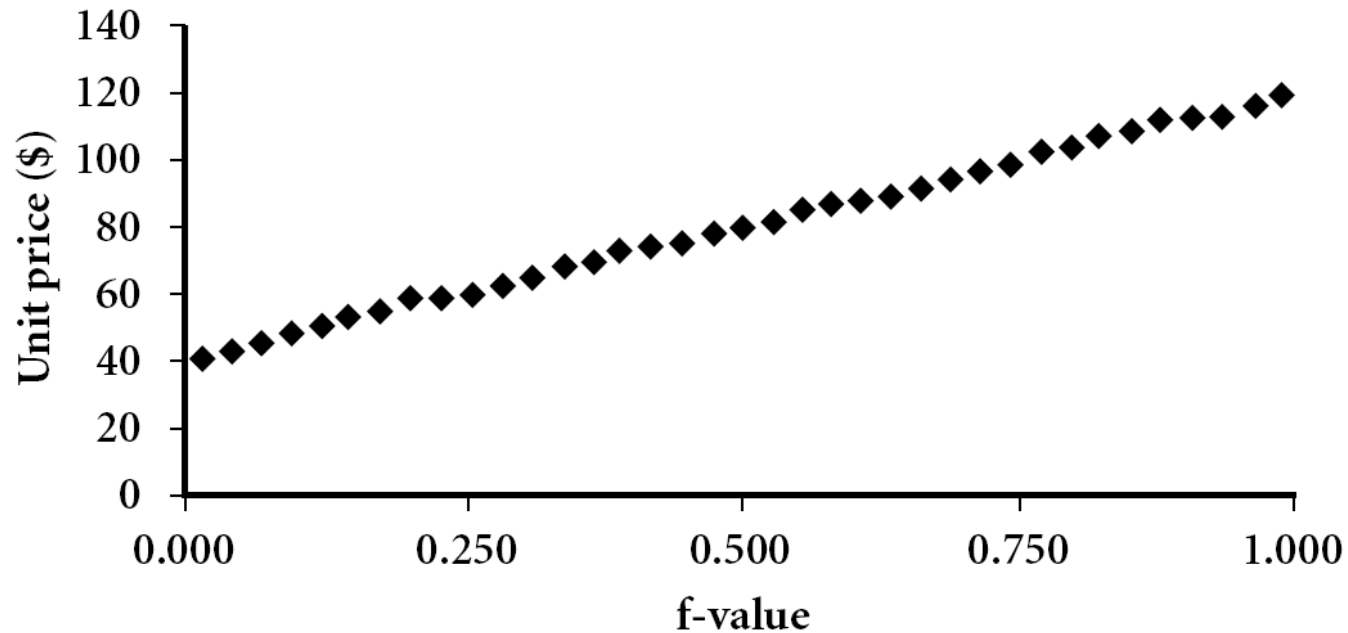
# Ιστόγραμμα

- Ιστόγραμμα (Histogram): Γραφική αναπαράσταση συχνοτήτων των τιμών που απεικονίζονται σαν μπάρες
- Δείχνει την αναλογία των τιμών που περιλαμβάνονται σε συγκεκριμένες κατηγορίες
- Διαφέρει από ένα διάγραμμα μπαρών ως προς το ότι είναι η περιοχή της μπάρας που απεικονίζει την τιμή και όχι το ύψος
- Αυτή η διαφορά είναι σημαντική όταν οι κατηγορίες δεν έχουν ομοιόμορφο πλάτος
- Οι κατηγορίες συνήθως απεικονίζονται ως μη επικαλυπτόμενα διαστήματα κάποιων μεταβλητών
- Οι κατηγορίες/μπάρες πρέπει να είναι γειτονικές



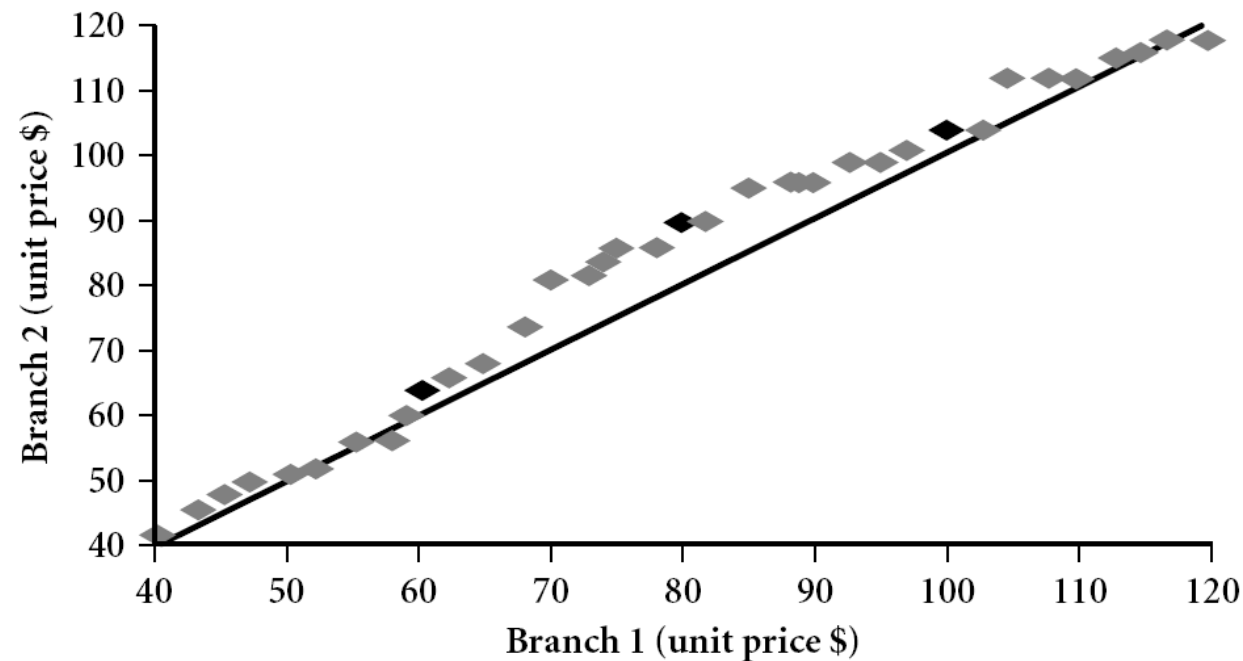
# Quantiles

- Παρουσιάζουν όλα τα δεδομένα επιτρέποντας στους χρήστες να εκτιμήσουν τη συνολική τους 'συμπεριφορά' αλλά και μη αναμενόμενες τιμές
- Για μια τιμή  $x_i$  τα δεδομένα ταξινομούνται σε αύξουσα σειρά και η 'συνδεδεμένη' τιμή  $f_i$  απεικονίζει ότι περίπου  $100 \times f_i\%$  των δεδομένων είναι μικρότερα ή ίσα της  $x_i$



# Διάγραμμα Q-Q

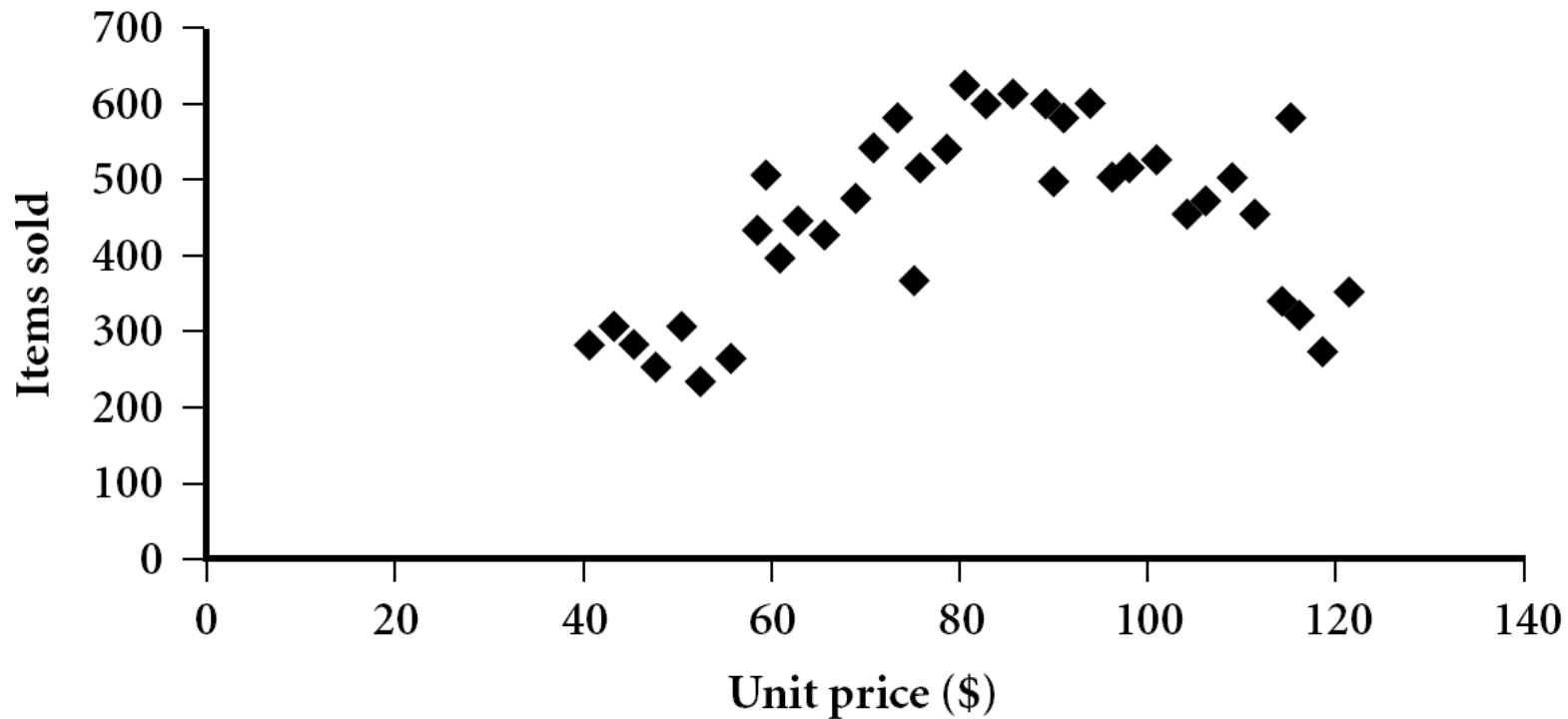
- Απεικονίζει τα τεταρτημόρια μιας κατανομής συγκρινόμενα με τα αντίστοιχα τεταρτημόρια μιας άλλης κατανομής
- Στόχος: Υπάρχει μετατόπιση όταν μεταβαίνουμε από τη μια κατανομή στην άλλη;
- Τα παράδειγμα δείχνει την τιμή κάποιων πωληθέντων προϊόντων σε δύο καταστήματα για κάθε τεταρτημόριο





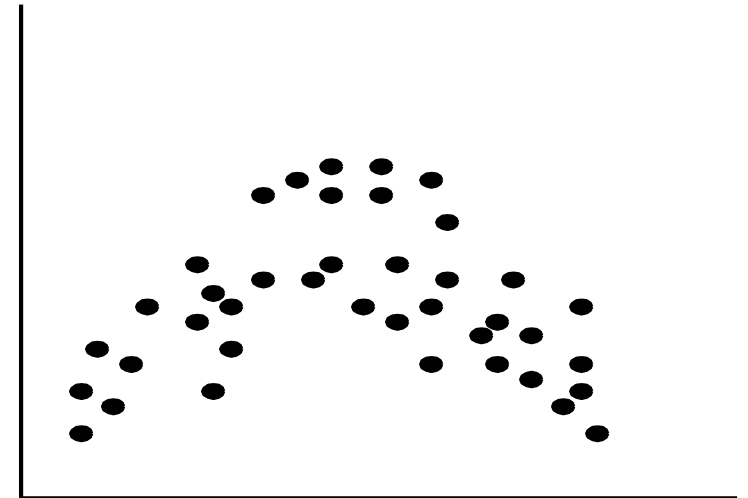
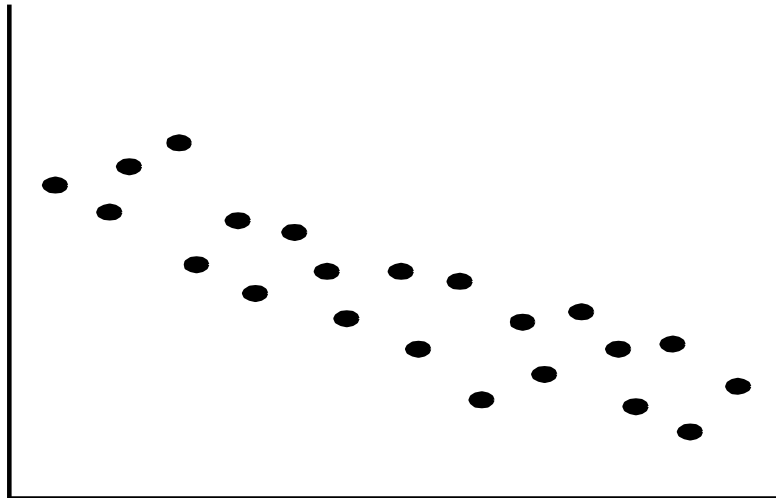
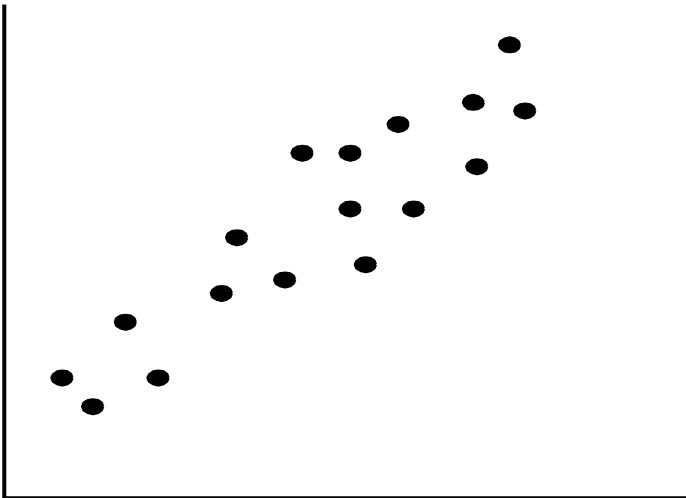
# Scatter Plot

- Απεικονίζει ζεύγη τιμών στο χώρο ώστε να διαπιστώσουμε την ύπαρξη συστάδων, ανωμαλιών, κ.λπ.
- Κάθε ζεύγος τιμών αντιμετωπίζεται σαν συντεταγμένες στο χώρο



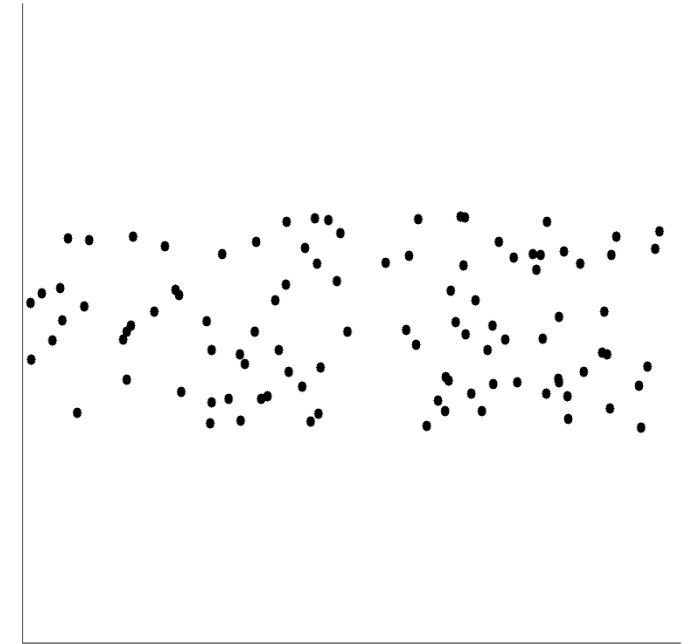
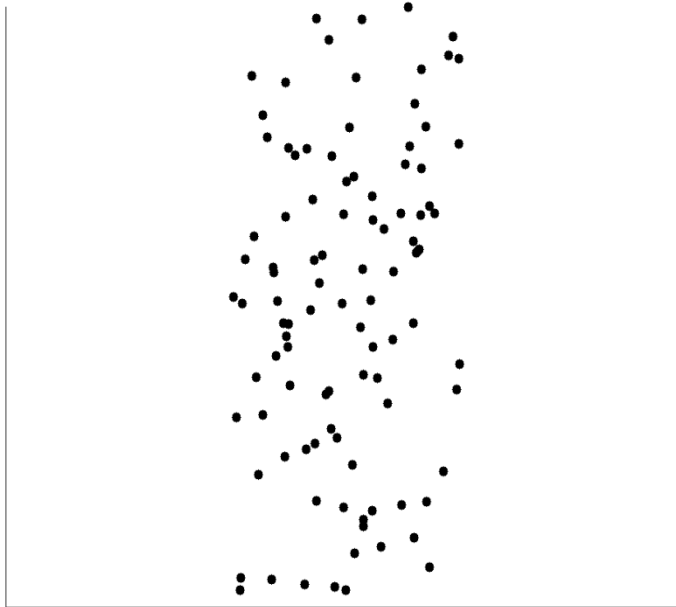
# Συσχέτιση

- Αριστερά: Θετική συσχέτιση – και οι δύο μεταβλητές αυξάνουν ταυτόχρονα
- Μέση: Αρνητική συσχέτιση – όταν η μια μεταβλητή αυξάνει η δεύτερη μειώνεται και το αντίστροφο
- Δεξιά: Εναλλασσόμενη συσχέτιση



# Συσχέτιση

- Παραδείγματα στα οποία δεν μπορούμε να αποφασίσουμε κάποιου είδους συσχέτιση



# Ομοιότητα – Ανομοιότητα Δεδομένων



# Ομοιότητα

- **Ομοιότητα (Similarity)**

- Αριθμητική τιμή που δείχνει όσο πολύ μοιάζουν αντικείμενα δεδομένων μεταξύ τους
- Η τελική είναι μεγαλύτερη όσο μεγαλύτερη είναι η ομοιότητα
- Συνήθως, η τελική τιμή ορίζεται στο διάστημα  $[0,1]$

- **Ανομοιότητα (Dissimilarity) ή απόσταση**

- Αριθμητική τιμή που δείχνει όσο πολύ διαφέρουν αντικείμενα δεδομένων μεταξύ τους
- Η τελική είναι μικρότερη όσο μεγαλύτερη είναι η διαφορά
- Η ελάχιστη τιμή είναι συνήθως ίση με 0
- Το άνω όριο της ανομοιότητας ποικίλει

# • Πίνακας Ανομοιότητας •

- Πίνακας Δεδομένων (Data matrix)

- η αντικείμενα δεδομένων με  $p$  διαστάσεις

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Πίνακας Ανομοιότητας (Dissimilarity matrix)

- η σημεία δεδομένων αλλά αποτυπώνει τη διαφορά ανάμεσα σε ζεύγη αντικειμένων
- Ο πίνακας είναι τριγωνικός αφού δεν χρειάζεται να απεικονίσουμε δύο φορές την ανομοιότητα για το ίδιο ζεύγος

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

# Nominal Data

- Μπορούν να πάρουν δύο ή περισσότερες τιμές, π.χ. red, yellow, blue, green (αποτελούν γενίκευση των δυαδικών μεταβλητών)
- 1<sup>η</sup> μέθοδος: Απλό ταίριασμα μεταξύ των τιμών
  - m: πλήθος των ταιριασμάτων, p: συνολικό πλήθος των μεταβλητών

$$d(i, j) = \frac{p - m}{p}$$

- 2<sup>η</sup> μέθοδος: Χρήση ενός μεγάλου αριθμού δυαδικών μεταβλητών για να δημιουργήσουμε ένα νέο δυαδικό χαρακτηριστικό για κάθε μια από τις M αλφαριθμητικές καταστάσεις

# Nominal Data

- Παράδειγμα

| Object Identifier | test-1<br>(nominal) | test-2<br>(ordinal) | test-3<br>(numeric) |
|-------------------|---------------------|---------------------|---------------------|
| 1                 | code A              | excellent           | 45                  |
| 2                 | code B              | fair                | 22                  |
| 3                 | code C              | good                | 64                  |
| 4                 | code A              | excellent           | 28                  |

- Πίνακας ανομοιότητας με  $p=1$

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix} \rightarrow$$

$$d(i,j) = \frac{p-m}{p}$$

- $\text{sim}(i,j) = 1-d(i,j) = m/p$

|            | test1   | test2   | test3   |   |   |            |         |         |         |   |   |  |  |  |
|------------|---------|---------|---------|---|---|------------|---------|---------|---------|---|---|--|--|--|
| 1          | a       | v       | q       |   |   |            |         |         |         |   |   |  |  |  |
| 2          | a       | b       | r       |   |   |            |         |         |         |   |   |  |  |  |
| 3          | b       | b       | q       |   |   |            |         |         |         |   |   |  |  |  |
| 4          | c       | v       | t       |   |   |            |         |         |         |   |   |  |  |  |
| 5          | a       | b       | q       |   |   |            |         |         |         |   |   |  |  |  |
|            |         |         |         |   |   |            |         |         |         |   |   |  |  |  |
|            |         |         |         |   |   |            |         |         |         |   |   |  |  |  |
| Dis Matrix | 1       | 2       | 3       | 4 | 5 | Sim Matrix | 1       | 2       | 3       | 4 | 5 |  |  |  |
| 1          |         |         |         |   |   | 1          |         |         |         |   |   |  |  |  |
|            | 0,66666 |         |         |   |   |            | 0,33333 |         |         |   |   |  |  |  |
| 2          | 7       |         |         |   |   | 2          | 3       |         |         |   |   |  |  |  |
|            | 0,66666 | 0,66666 |         |   |   |            | 0,33333 | 0,33333 |         |   |   |  |  |  |
| 3          | 7       | 7       |         |   |   | 3          | 3       | 3       |         |   |   |  |  |  |
|            | 0,66666 |         |         |   |   |            | 0,33333 |         |         |   |   |  |  |  |
| 4          | 7       | 1       | 1       |   |   | 4          | 3       | 0       | 0       |   |   |  |  |  |
|            | 0,33333 | 0,33333 | 0,33333 |   |   |            | 0,66666 | 0,66666 | 0,66666 |   |   |  |  |  |
| 5          | 3       | 3       | 3       | 1 |   | 5          | 7       | 7       | 7       | 0 |   |  |  |  |



# Nominal Data

- Πίνακας ανομοιότητας με  $p=4$

- $\text{sim}(i,j) = 1 - d(i,j) = m/p$

$$d(i, j) = \frac{p - m}{p}$$

|            | test1 | Test2 | test3 | Test4 |   |            |     |     |     |   |   |  |
|------------|-------|-------|-------|-------|---|------------|-----|-----|-----|---|---|--|
| 1          | a     | v     | q     | d     |   |            |     |     |     |   |   |  |
| 2          | a     | b     | r     | w     |   |            |     |     |     |   |   |  |
| 3          | b     | b     | q     | e     |   |            |     |     |     |   |   |  |
| 4          | c     | v     | t     | w     |   |            |     |     |     |   |   |  |
| 5          | a     | b     | q     | d     |   |            |     |     |     |   |   |  |
|            |       |       |       |       |   |            |     |     |     |   |   |  |
|            |       |       |       |       |   |            |     |     |     |   |   |  |
|            |       |       |       |       |   |            |     |     |     |   |   |  |
| Dis Matrix | 1     | 2     | 3     | 4     | 5 | Sim Matrix | 1   | 2   | 3   | 4 | 5 |  |
| 1          |       |       |       |       |   | 1          |     |     |     |   |   |  |
| 2          | 3/4   |       |       |       |   | 2          | 1/4 |     |     |   |   |  |
| 3          | 3/4   | 3/4   |       |       |   | 3          | 1/4 | 1/4 |     |   |   |  |
| 4          | 3/4   | 3/4   | 1     |       |   | 4          | 1/4 | 1/4 | 0   |   |   |  |
| 5          | 1/4   | 1/2   | 1/2   | 1     |   | 5          | 3/4 | 1/2 | 1/2 | 0 |   |  |

# Binary Data

- Χρήση ενός πίνακα συνάφειας (contingency table) για τα δυαδικά δεδομένα

|          |     | Object j |         |         |
|----------|-----|----------|---------|---------|
|          |     | 1        | 0       | sum     |
| Object i | 1   | $q$      | $r$     | $q + r$ |
|          | 0   | $s$      | $t$     | $s + t$ |
|          | sum | $q + s$  | $r + t$ | $p$     |

- Μετρική απόστασης για συμμετρικά δυαδικά δεδομένα
- Μετρική απόστασης για μη συμμετρικά δυαδικά δεδομένα
- Συντελεστής Jaccard (Jaccard coefficient) για την εξαγωγή ομοιότητας μη συμμετρικών δυαδικών δεδομένων

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

# Παράδειγμα

| Name | Gender | Fever | Cough | Test-1 | Test-2 | Test-3 | Test-4 |
|------|--------|-------|-------|--------|--------|--------|--------|
| Jack | M      | Y     | N     | P      | N      | N      | N      |
| Mary | F      | Y     | N     | P      | N      | P      | N      |
| Jim  | M      | Y     | P     | N      | N      | N      | N      |

- Το Gender θεωρείται ότι είναι συμμετρικό
- Οι υπόλοιπες μεταβλητές είναι μη συμμετρικές
- Υποθέτουμε ότι τα Y & P απεικονίζονται με 1, ενώ το N με 0

|     |         |         |         |
|-----|---------|---------|---------|
|     | 1       | 0       | sum     |
| 1   | $q$     | $r$     | $q + r$ |
| 0   | $s$     | $t$     | $s + t$ |
| sum | $q + s$ | $r + t$ | $p$     |

$$d(\text{jack}, \text{mary}) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(\text{jack}, \text{jim}) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(\text{jim}, \text{mary}) = \frac{2 + 1}{1 + 2 + 1} = 0.75$$

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

$$d(i, j) = \frac{r + s}{q + r + s}$$

# Αριθμητικά Δεδομένα

- Απόσταση Minkowski: Αποτελεί μια δημοφιλή μετρική

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

- όπου  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  &  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  είναι δύο p-διαστάσεων αντικείμενα δεδομένων και h είναι η τάξη (η απόσταση ορίζεται σαν η L-h norm)
- Ιδιότητες
  - $d(i, j) > 0$ , αν  $i \neq j$ , και  $d(i, i) = 0$  (Positive definiteness)
  - $d(i, j) = d(j, i)$  (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$  (τριγωνική ανίσωση - Triangle Inequality)
- Μια συνάρτηση απόστασης που ικανοποιεί τις παραπάνω ιδιότητες αποτελεί μια αποδεκτή μετρική

# Περιπτώσεις

- $h = 1$ : Manhattan (city block, L1 norm) distance

π.χ., η απόσταση Hamming: πλήθος bits που διαφέρουν ανάμεσα σε δύο δυαδικά διανύσματα

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$ : (L2 norm) Euclidean distance

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$ . “supremum” ( $L_{\max}$  norm,  $L_{\infty}$  norm) distance.

είναι η μέγιστη απόσταση ανάμεσα σε οποιοδήποτε στοιχείο των διανυσμάτων

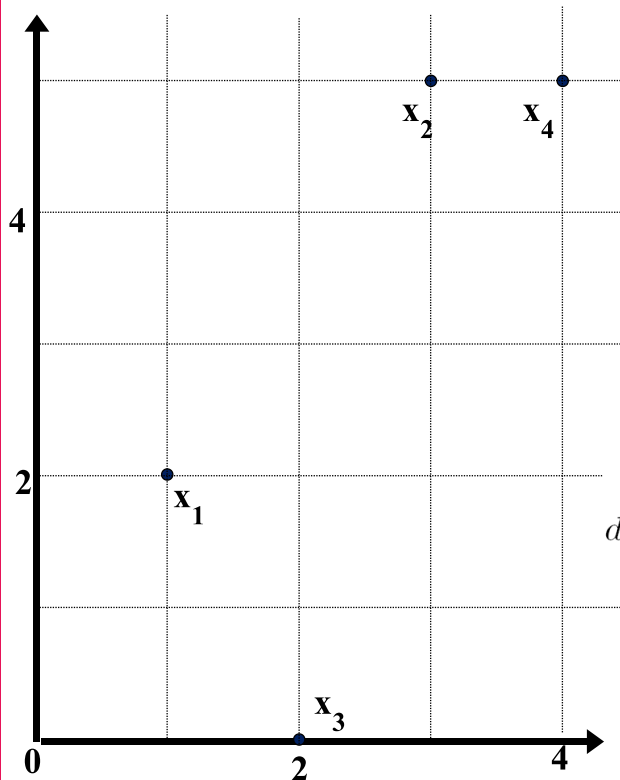
$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h}$$

# Παραδείγματα

| point | attribute 1 | attribute 2 |
|-------|-------------|-------------|
| x1    | 1           | 2           |
| x2    | 3           | 5           |
| x3    | 2           | 0           |
| x4    | 4           | 5           |

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$



$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f^p |x_{if} - x_{jf}|$$

Manhattan ( $L_1$ )

| L  | x1 | x2 | x3 | x4 |
|----|----|----|----|----|
| x1 | 0  |    |    |    |
| x2 | 5  | 0  |    |    |
| x3 | 3  | 6  | 0  |    |
| x4 | 6  | 1  | 7  | 0  |

Euclidean ( $L_2$ )

| L2 | x1   | x2  | x3   | x4 |
|----|------|-----|------|----|
| x1 | 0    |     |      |    |
| x2 | 3.61 | 0   |      |    |
| x3 | 2.24 | 5.1 | 0    |    |
| x4 | 4.24 | 1   | 5.39 | 0  |

Supremum

| $L_\infty$ | x1 | x2 | x3 | x4 |
|------------|----|----|----|----|
| x1         | 0  |    |    |    |
| x2         | 3  | 0  |    |    |
| x3         | 2  | 5  | 0  |    |
| x4         | 3  | 1  | 5  | 0  |

# Παραδείγματα

|    |   |   |   |
|----|---|---|---|
| X1 | 3 | 4 | 1 |
| X2 | 6 | 2 | 1 |
| X3 | 1 | 0 | 4 |
| X4 | 5 | 5 | 2 |

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Manhattan ( $L_1$ )

|   |    |    |   |
|---|----|----|---|
| - |    |    |   |
| 5 | -  |    |   |
| 9 | 10 | -  |   |
| 4 | 5  | 11 | - |

Euclidean ( $L_2$ )

|      |      |      |   |
|------|------|------|---|
| -    |      |      |   |
| 3.61 | -    |      |   |
| 5.38 | 6.16 | -    |   |
| 2.44 | 3.31 | 6.70 | - |

Supremum

|   |   |   |   |
|---|---|---|---|
| - |   |   |   |
| 3 | - |   |   |
| 4 | 5 | - |   |
| 2 | 3 | 5 | - |

# • Ομοιότητα Συνημιτόνου •

- Ένα έγγραφο μπορεί να αναπαρασταθεί με χιλιάδες ιδιότητες κάθε μια από τις οποίες καταγράφει τη συχνότητα ενός όρου ή φράσης στο έγγραφο

| <i>Document</i> | <i>team</i> | <i>coach</i> | <i>hockey</i> | <i>baseball</i> | <i>soccer</i> | <i>penalty</i> | <i>score</i> | <i>win</i> | <i>loss</i> | <i>season</i> |
|-----------------|-------------|--------------|---------------|-----------------|---------------|----------------|--------------|------------|-------------|---------------|
| Document1       | 5           | 0            | 3             | 0               | 2             | 0              | 0            | 2          | 0           | 0             |
| Document2       | 3           | 0            | 2             | 0               | 1             | 1              | 0            | 1          | 0           | 1             |
| Document3       | 0           | 7            | 0             | 2               | 1             | 0              | 0            | 3          | 0           | 0             |
| Document4       | 0           | 1            | 0             | 0               | 1             | 2              | 2            | 0          | 3           | 0             |

- Άλλα παραδείγματα: στοιχεία γονιδίων σε μικροπίνακες, κ.λπ.
- Εφαρμογές: Ανάκτηση πληροφοριών, βιολογικές ταξινομίες, κ.λπ.
- Μετρική συνημιτόνου (Cosine measure): Αν τα  $d_1$  και  $d_2$  είναι δύο διανύσματα (π.χ. διανύσματα συχνοτήτων), τότε

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\| ,$$

- όπου η πράξη  $\cdot$  είναι το εσωτερικό γινόμενο των διανυσμάτων και  $\|d\|$  είναι το μήκος διανύσματος

$$\text{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)| |\vec{V}(d_2)|} = \frac{\sum_{i=1}^M x_i y_i}{\sqrt{\sum_{i=1}^M \vec{V}_i^2(d)}}$$



# Παράδειγμα

- $\cos(d1, d2) = (d1 \cdot d2) / \|d1\| \|d2\|$  , όπου το  $\cdot$  δείχνει το εσωτερικό γινόμενο και  $\|d\|$  είναι το μήκος του διανύσματος  $d$

- Να βρεθεί η ομοιότητα ανάμεσα στα έγγραφα 1 & 2 καθώς και στα 1 & 3

$$d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d1 \cdot d2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$$

$$\|d1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d2\| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d1, d2) = 25 / (6.481 \times 4.12) = 0.94$$

$$d1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d3 = (0, 7, 0, 2, 1, 0, 0, 3, 0, 0)$$

$$d1 \cdot d3 = 5*0+0*7+3*0+0*2+2*1+0*0+0*0+2*3+0*0+0*0 = 8$$

$$\|d1\| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d3\| = (0*0+7*7+0*0+2*2+1*1+0*0+0*0+3*3+0*0+0*0)^{0.5} = (59)^{0.5} = 7.681$$

$$\cos(d1, d3) = 8 / (6.481 \times 7.681) = 0.16$$

| Document  | team | coach | hockey | baseball | soccer | penalty | score | win | loss | season |
|-----------|------|-------|--------|----------|--------|---------|-------|-----|------|--------|
| Document1 | 5    | 0     | 3      | 0        | 2      | 0       | 0     | 2   | 0    | 0      |
| Document2 | 3    | 0     | 2      | 0        | 1      | 1       | 0     | 1   | 0    | 1      |
| Document3 | 0    | 7     | 0      | 2        | 1      | 0       | 0     | 3   | 0    | 0      |
| Document4 | 0    | 1     | 0      | 0        | 1      | 2       | 2     | 0   | 3    | 0      |

