第二章 认识数据

张静
（模式识别与智能数据研究室）
（Jingzhang@ecust.edu.cn）

主要内容

- 数据对象和属性类型
- 数据的统计描述
- 计算数据的相似度和不相似度
- 小结

数据集合的类型

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	pla y	ball	score	game	wi n	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

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

Important Characteristics of Structured Data

- Dimensionality（维数）
  - Curse of dimensionality
- Sparsity（稀疏性）
  - Only presence counts
- Resolution（分辨率）
  - Patterns depend on the scale
- Distribution（分布性）
  - Centrality and dispersion

数据对象

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples.*
- Data objects are described by **attributes.**
- Database rows -> data objects; columns ->attributes.

属性

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal（标称属性）
  - Binary（二元属性）
  - Ordinal attribute （序数属性）
  - Numeric: quantitative （数值属性）
    - Interval-scaled（区间标度）
    - Ratio-scaled（比例标度）

Attribute Types

- **Nominal** （标称）
  - *categories, states, or "names of things"*
  - *Hair_color = {black, brown, blond, red, auburn, grey, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**（二元）
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**（序数）
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- **Quantity（数值） (integer or real-valued)**
  - **Interval（区间标度属性）**
    - Measured on a scale of **equal-sized units**
    - Values have order
      - E.g., *temperature in C˚or F˚, calendar dates*
    - No true zero-point
  - **Ratio（比例标度属性）**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
      - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes （离散属性与连续属性）

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

主要内容

- 数据对象和属性类型
- 数据的统计描述
- 计算数据的相似度和不相似度
- 小结

数据的统计描述

- 数据的统计描述
  - 获得数据的总体印象
  - 识别数据的典型性质，凸显噪声或离群点
- 度量
  - 中心趋势度量
    - 均值（**mean**）
    - 中位数（**median**）
    - 众数（**mode**）
    - 中列数（**midrange**）
  - 离中心趋势度量
    - 四分位数（**quartiles**）
    - 四分位数极差（**interquartile range, IQR**）
    - 方差（**variance**）

度量数据的中心趋势

- 均值（Mean）：代数度量
  - 加权算术平均（Weighted arithmetic mean）：
  - 截断均值（Trimmed mean）：去除极端值

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

  - 注：
    - 分布式度量：是一种通过如下方法计算度量：将数据集划分成较小的子集，计算每个子集的度量，然后合并计算结果，得到原（整个）数据集的度量值。如sum（），count（）
    - 代数数量：可以通过应用一个代数函数于一个或多个分布度量计算的度量。如mean（）

度量数据的中心趋势

- 中位数（**Median**）：整体度量
  - 如果总数为奇数，则为中间那个数；如果为偶数，则为中间两个值的平均值
  - 对于已经按照某值划分的组数据，可以利用插值计算中位数的近似值：

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

  - ***L1***是中位数区间的下界，***N***是整个数据集的值的个数，$(\sum freq)_l$ 是低于中位数区间的所有区间的频率和，$freq_{median}$ 是中位数区间的频率，***width***是中位数区间的宽度。（ ）

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

度量数据的中心趋势

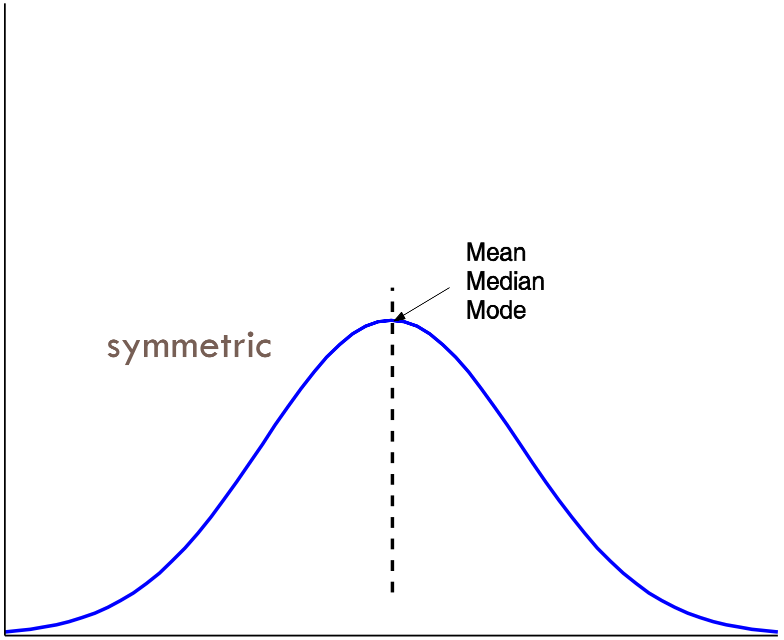
- <u>众数（**Mode**）：整体度量</u>
  - 数据集中出现频率最高的值
  - 单峰**Unimodal**, 双峰**bimodal**,三峰 **trimodal**
  - 对于适度倾斜（非对称）的单峰频率曲线，有如下经验关系:

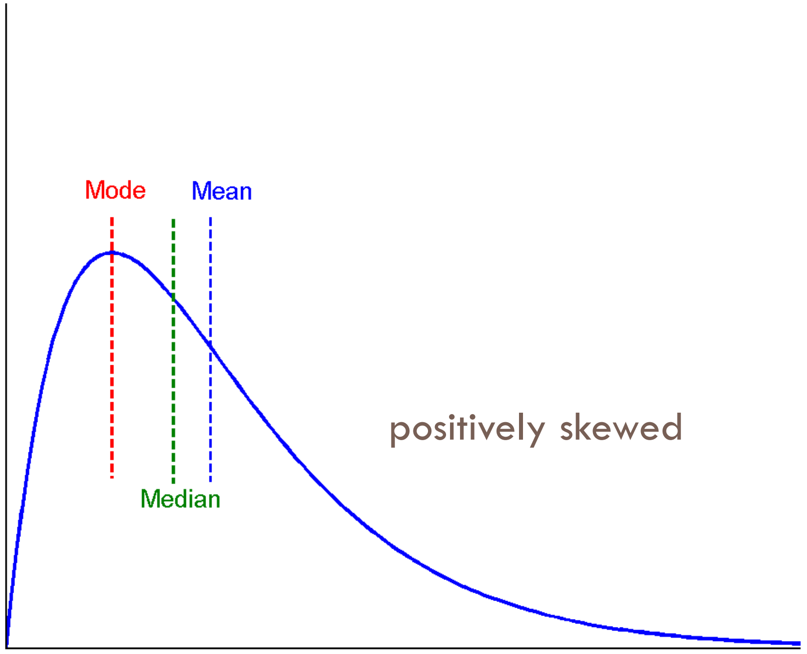
$$mean - mode = 3 \times (mean - median)$$

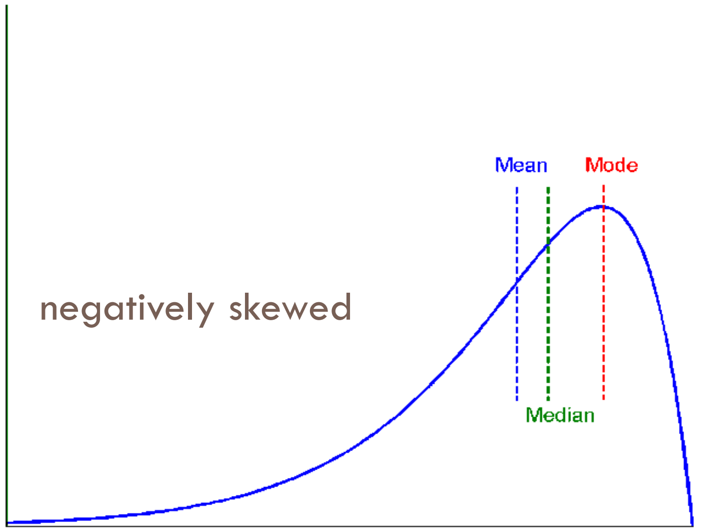
- <u>中列数（**Midrange**）：代数度量</u>
  - 数据集的最大和最小值的平均值

对称数据 vs. 倾斜数据

- 对称与正倾斜和负倾斜数据的中位数、均值和众数



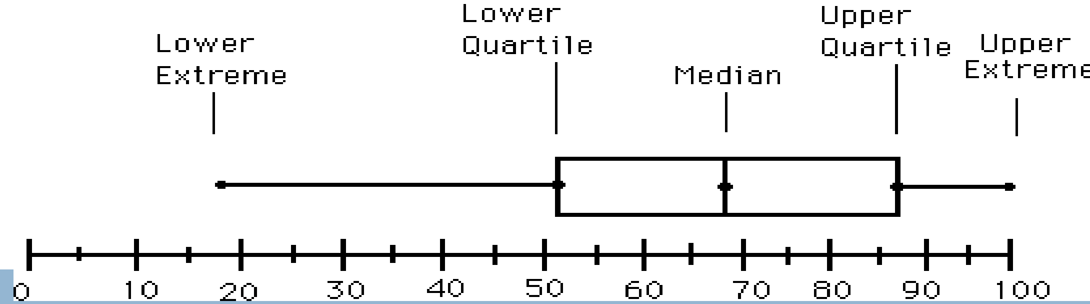




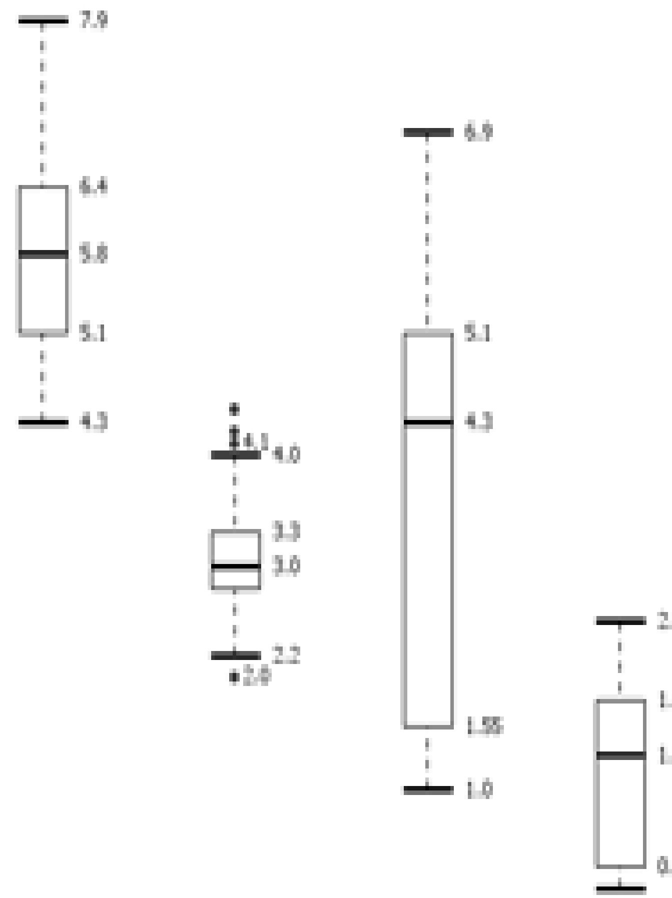
度量数据的离散程度

- 极差（range），四分位数（Quartiles），离群点（outliers）和盒图（ boxplots）
  - Range（极差）: max()-min()
  - Quartiles（四分位数）: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range（中间四分位数极差）: $IQR = Q_3 - Q_1$
  - Five number summary（五数概括）: min, Q_1, median, Q_3, max
  - Boxplot（盒图）: 盒的端点是四分位数；中位数用盒内的线标记；仅当最小最大观测值超过四分位数不到1.5 x IQR时，盒外的两条线延伸到最小和最大观测值，否则，胡须出现在四分位数的1.5 x IQR之内的最极端的观测值处终止；离群点单独表示。
  - Outlier（离群值）: 通常为高于/低于 1.5 x IQR的值。

盒图分析



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



度量数据的离散程度

- 方差（**Variance**）和标准差（ **standard deviation**）
  - **Variance: (algebraic, scalable computation)**

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

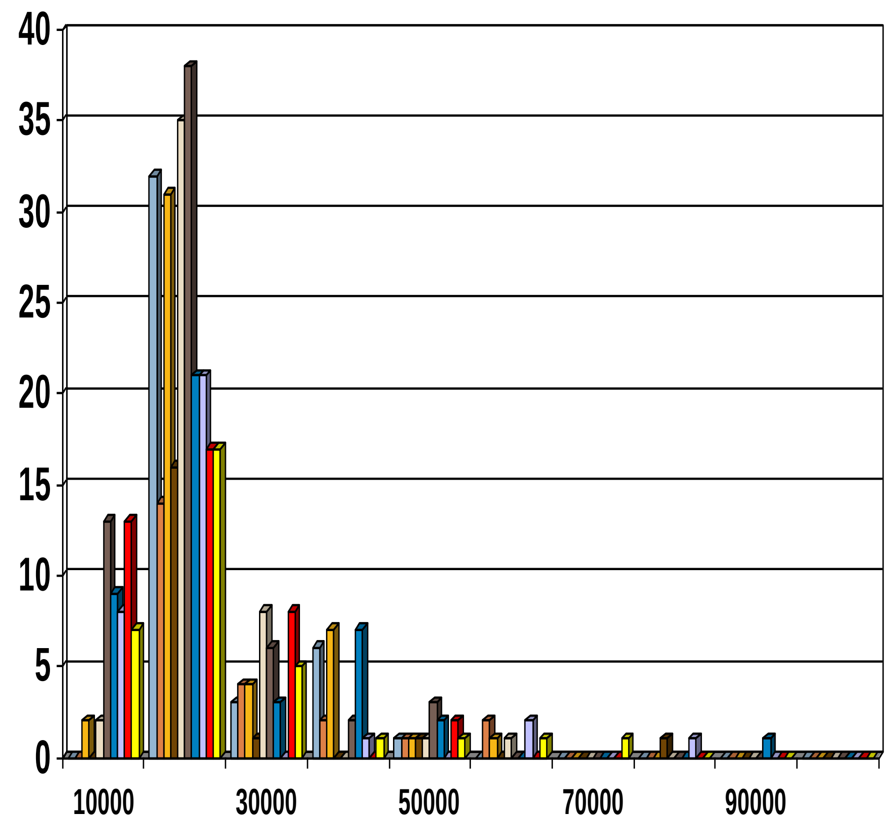
  - **Standard deviation *σ* is the square root of variance *σ*²**
- 作为发散性度量，标准差的基本性质如下
  - **σ**是关于均值的发散，仅当选择均值作为中心度量时使用。
  - 仅当不存在发散时，即当所有的观测值具有相同值时，**σ=0**，*否则***σ>0**。

数据的基本统计描述的图形显示

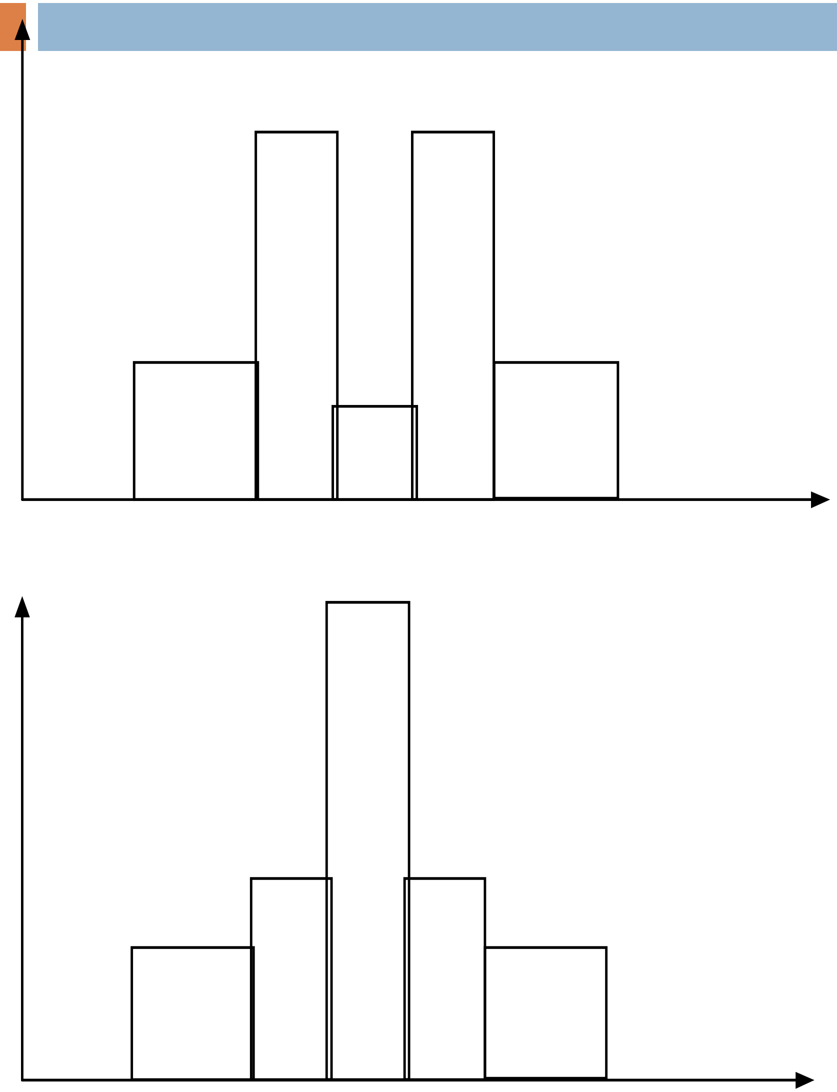
- **Boxplot**（盒图）：graphic display of five-number summary
- **Histogram**（直方图）：x-axis are values, y-axis repres. frequencies
- **Quantile plot**（分位数图）： each value x_i is paired with f_i indicating that approximately $f_i * 100$ % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**（分位数-分位数图）：graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**（散点图）：each pair of values is a pair of coordinates and plotted as points in the plane

Histogram Analysis（直方图分析）

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



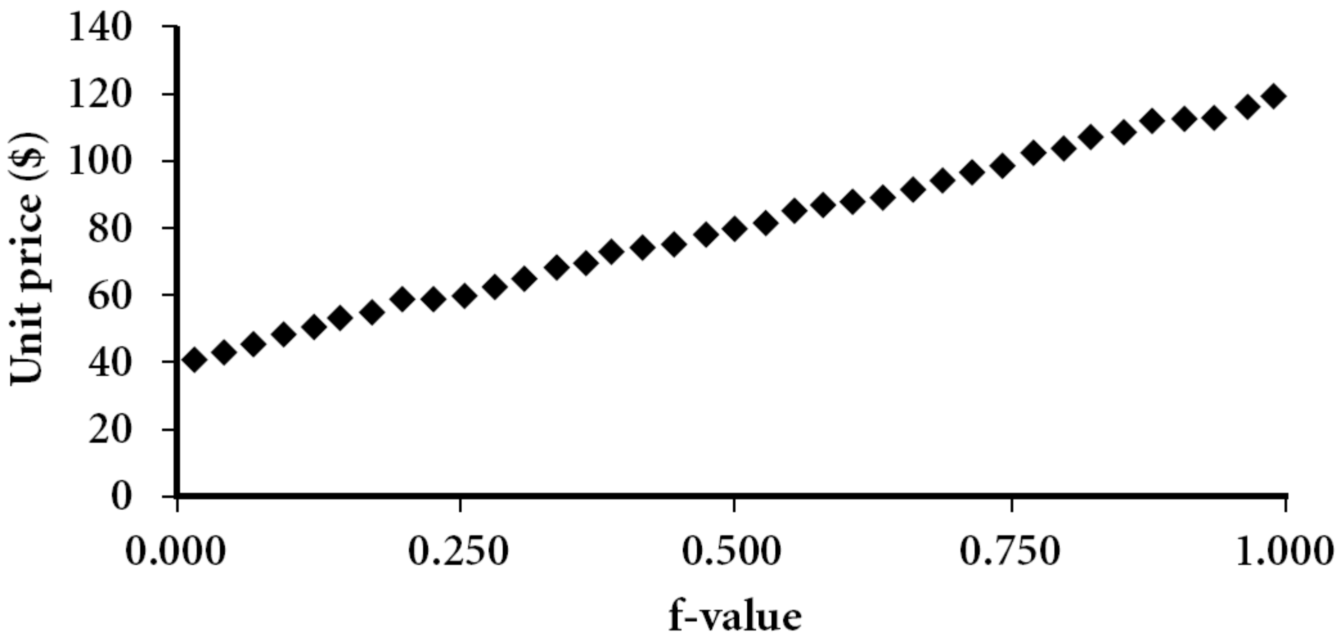
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

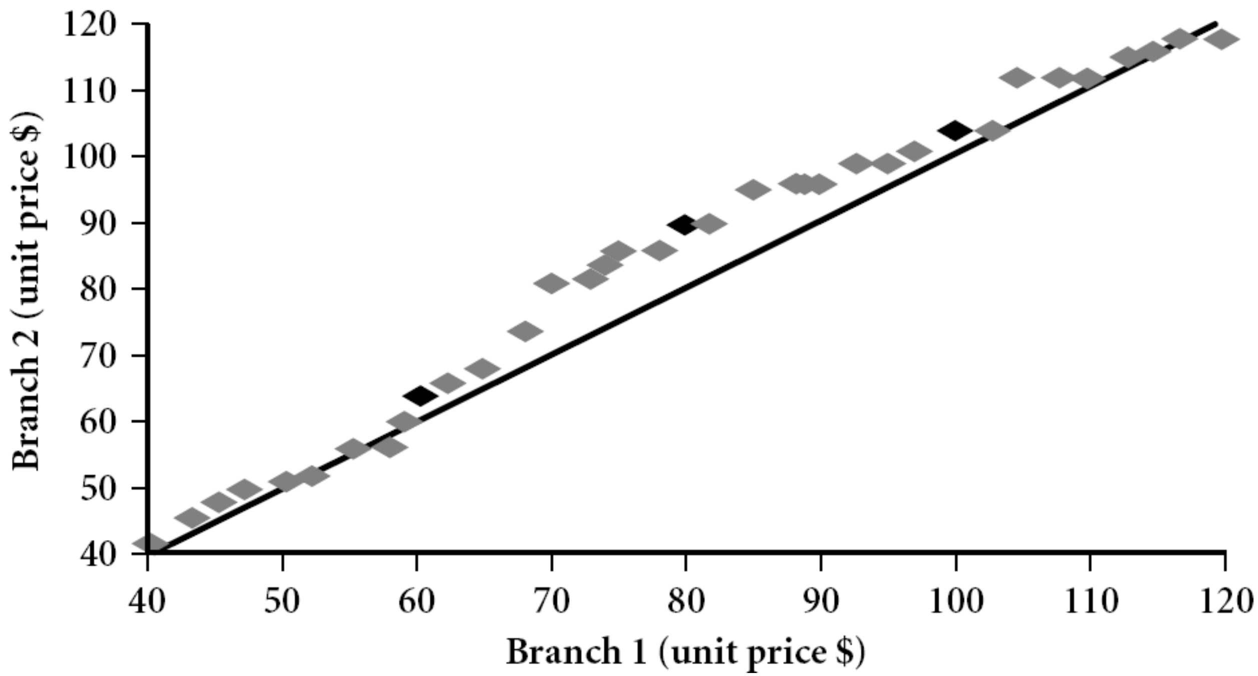
Quantile Plot（分位数图）

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



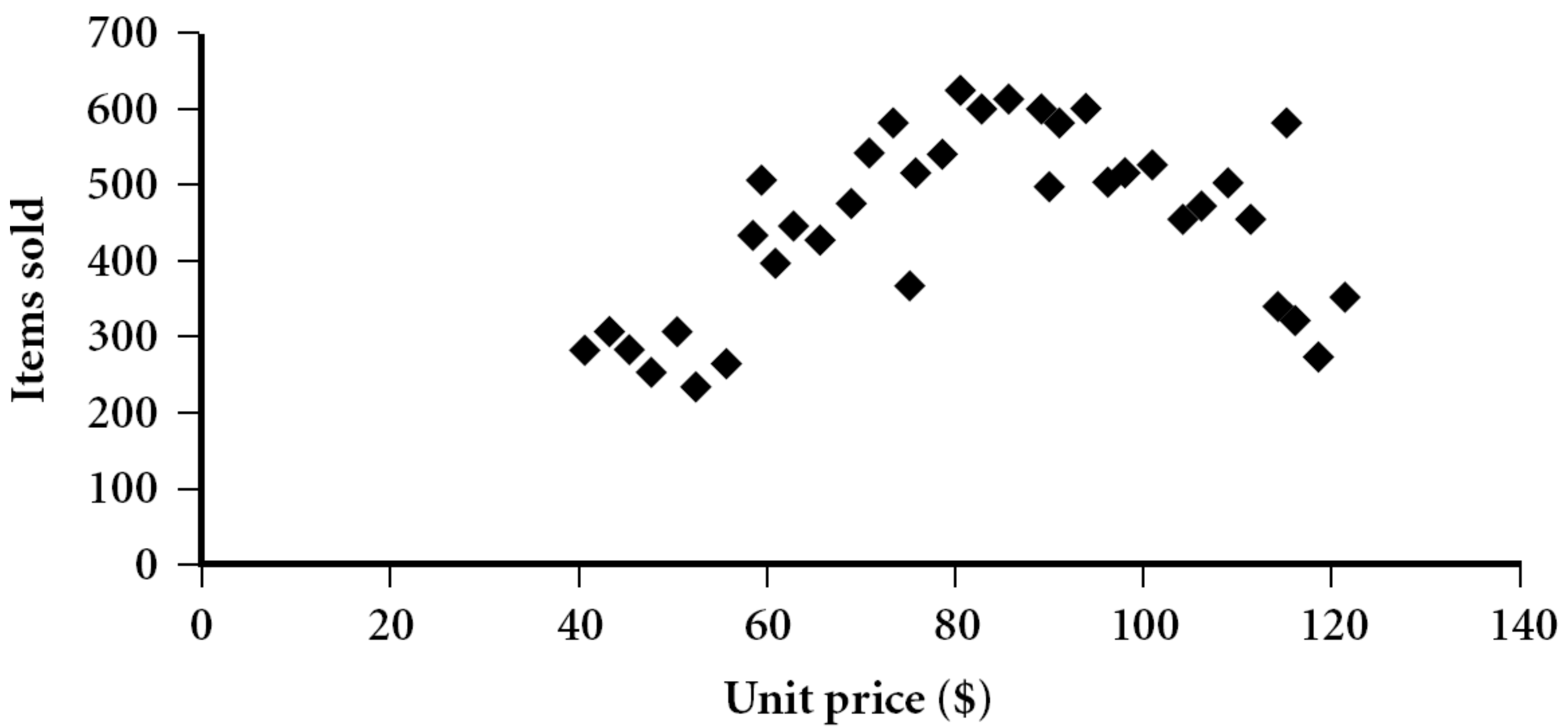
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

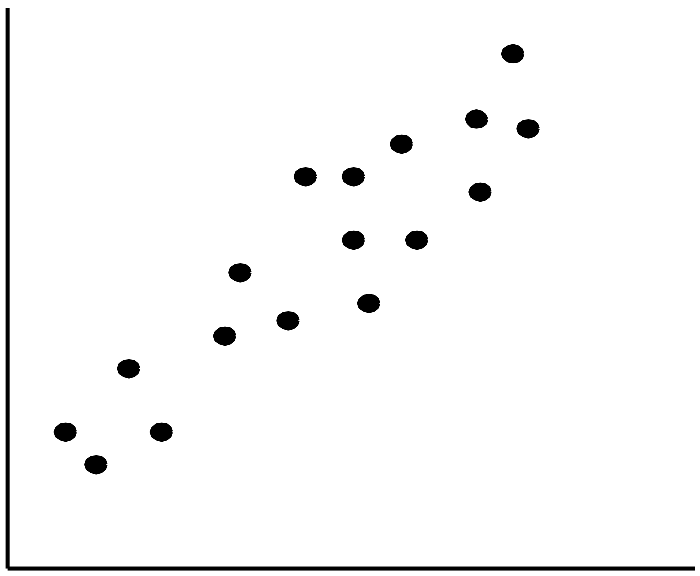


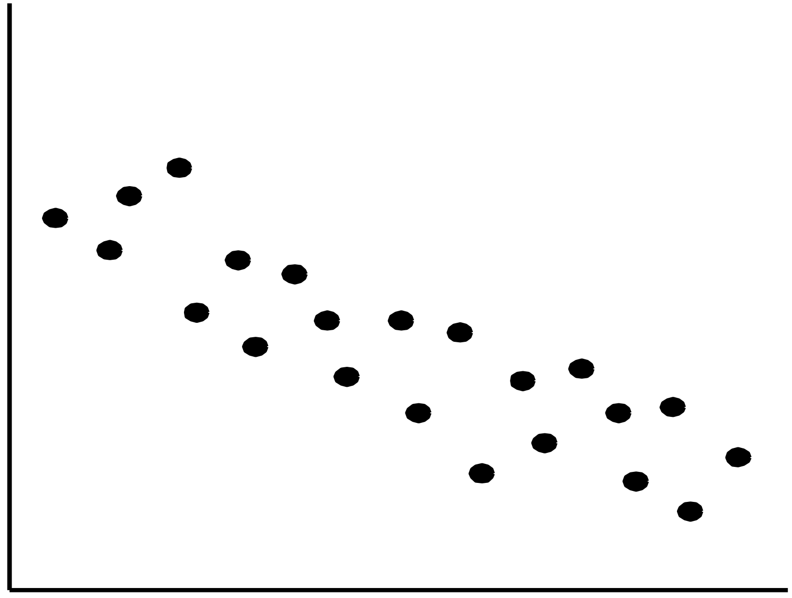
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



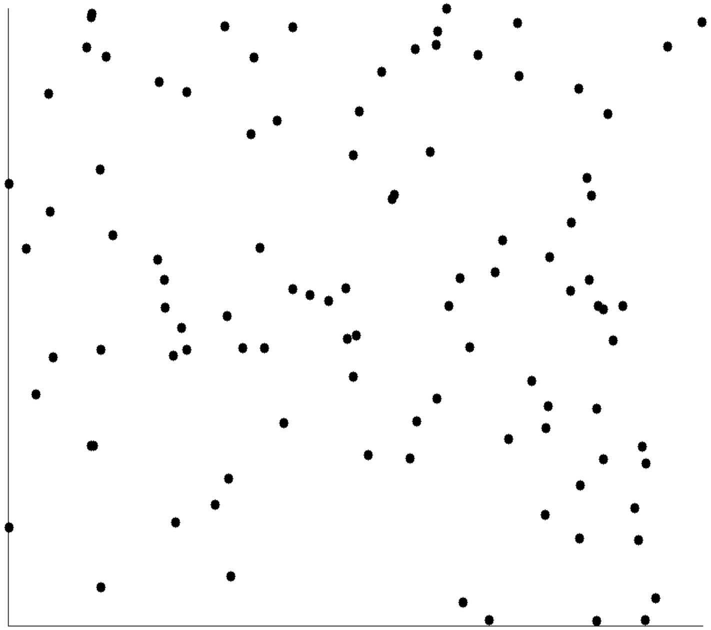
Positively and Negatively Correlated Data

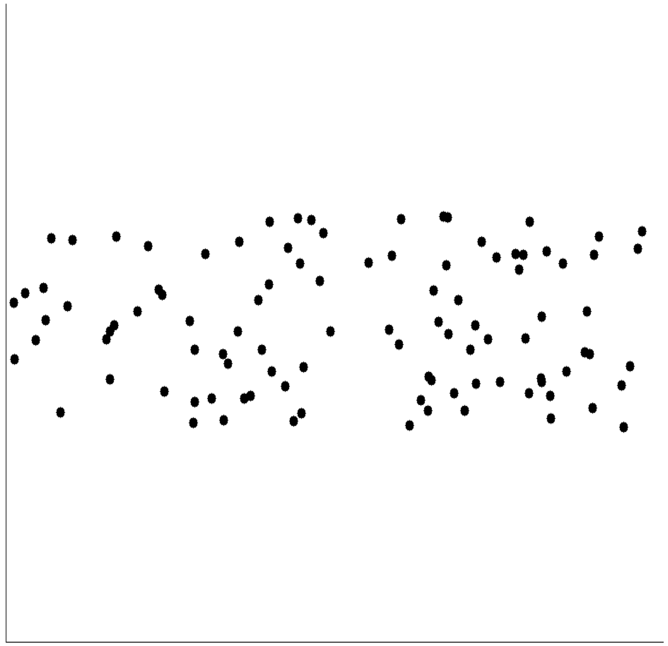


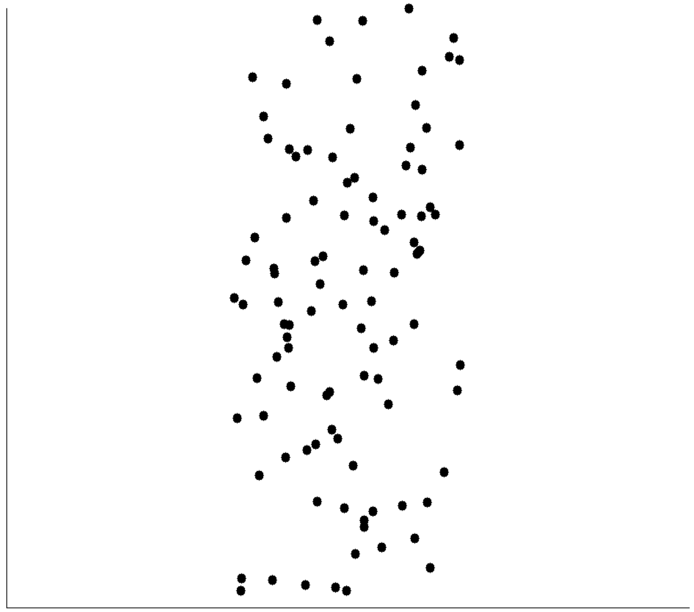


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







主要内容

- 数据对象和属性类型
- 数据的统计描述
- 计算数据的相似度和不相似度
- 小结

Similarity and Dissimilarity （相似性和相异性）

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity**（邻近性） refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

- Data matrix
  - n data points with p dimensions
  - Two mode (二模)

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

- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix
  - One mode (单模)

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

Nominal Attributes （标称属性）

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- <u>Method 1</u>: Simple matching
  - *m*: # of matches, *p*: total # of variables

不匹配率：

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

- <u>Method 2</u>: Use a large number of binary attributes
  - creating a new binary attribute for each of the *M* nominal states
  - 编码

Binary Attributes （二元属性）

- A contingency table for binary data （列联表）

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

- Distance measure for symmetric binary variables（对称的二元相异性）：

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

- Distance measure for asymmetric binary variables（非对称的二元相异性）：

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

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

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

- Note: Jaccard coefficient is the same as "coherence":

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q+r) + (q+s) - q}$$

Dissimilarity between Binary Variables

- Example

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

  - gender is a symmetric attribute
  - the remaining attributes are asymmetric binary
  - let the values Y and P be set to 1, and the value N be set to 0

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

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

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

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

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

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order

- Properties
  - $d(i, j) > 0$ if $i \neq j$ 非负性
  - $d(i, i) = 0$ (Positive definiteness) 同一性
  - $d(i, j) = d(j, i)$ (Symmetry) 对称性
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality) 三角不等式
- A distance that satisfies these properties is a metric（度量）

Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm范数) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: (L_2 norm) Euclidean distance

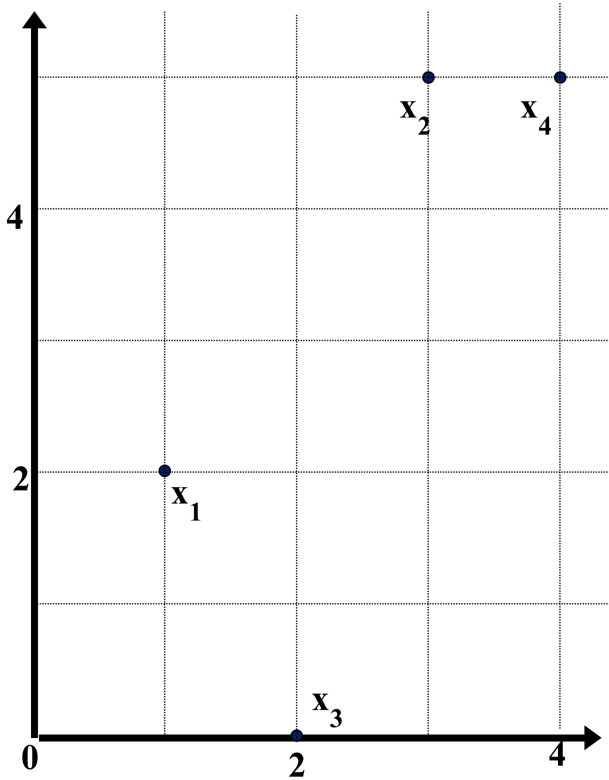
$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance. （上确界距离）
  - This is the maximum difference between any component (attribute) of the vectors

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

Example: Minkowski Distance

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



Dissimilarity Matrices

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_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables 序数变量

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - f is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - f is numeric: use the normalized distance
  - f is ordinal
    - Compute ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - Treat z_{if} as interval-scaled

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

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

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \; ||d_2|| ,$$

where • indicates vector dot product, $||d||$: the length of vector d

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \; ||d_2||$,
  where $\bullet$ indicates vector dot product, $||d|$: the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5} = 4.12$
$\cos(d_1, d_2) = 0.94$

小结

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al, Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009

作业：

- 第二章课后习题P53
  - 习题2.2
  - 习题2.3
  - 习题2.4
  - 习题2.6

思考题

- 试分析分布度量、代数度量，以及整体度量对于数据库的增量计算有何区别?

结束