Week 6-2

Data Measures

Big Data

Prof. Hwanjo Yu
POSTECH

Evolution of sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a theoretical component.
    Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, computational branch
    (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

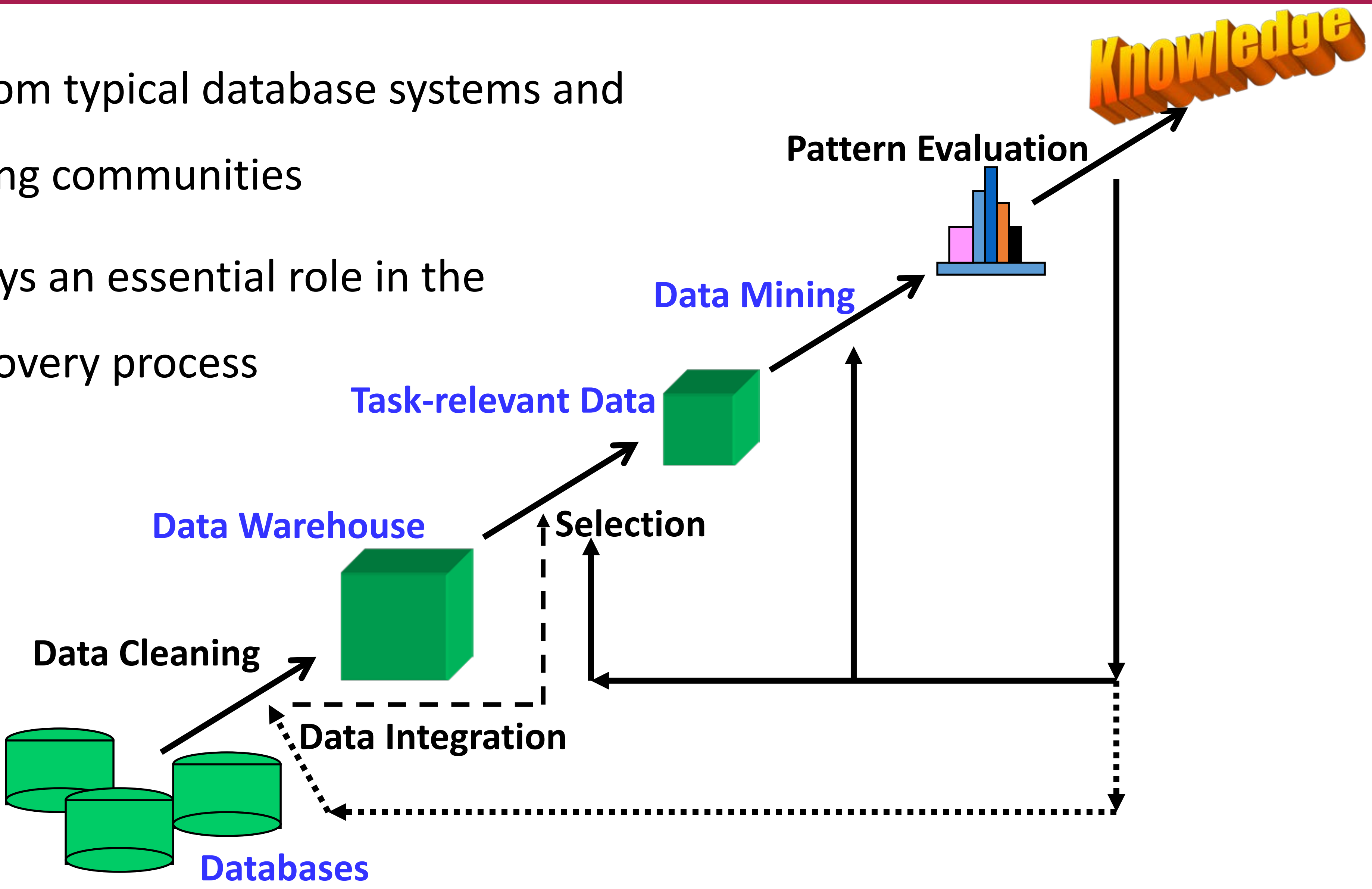
Jim Gray and Alex Szalay, The World Wide Telescope: An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of database technology

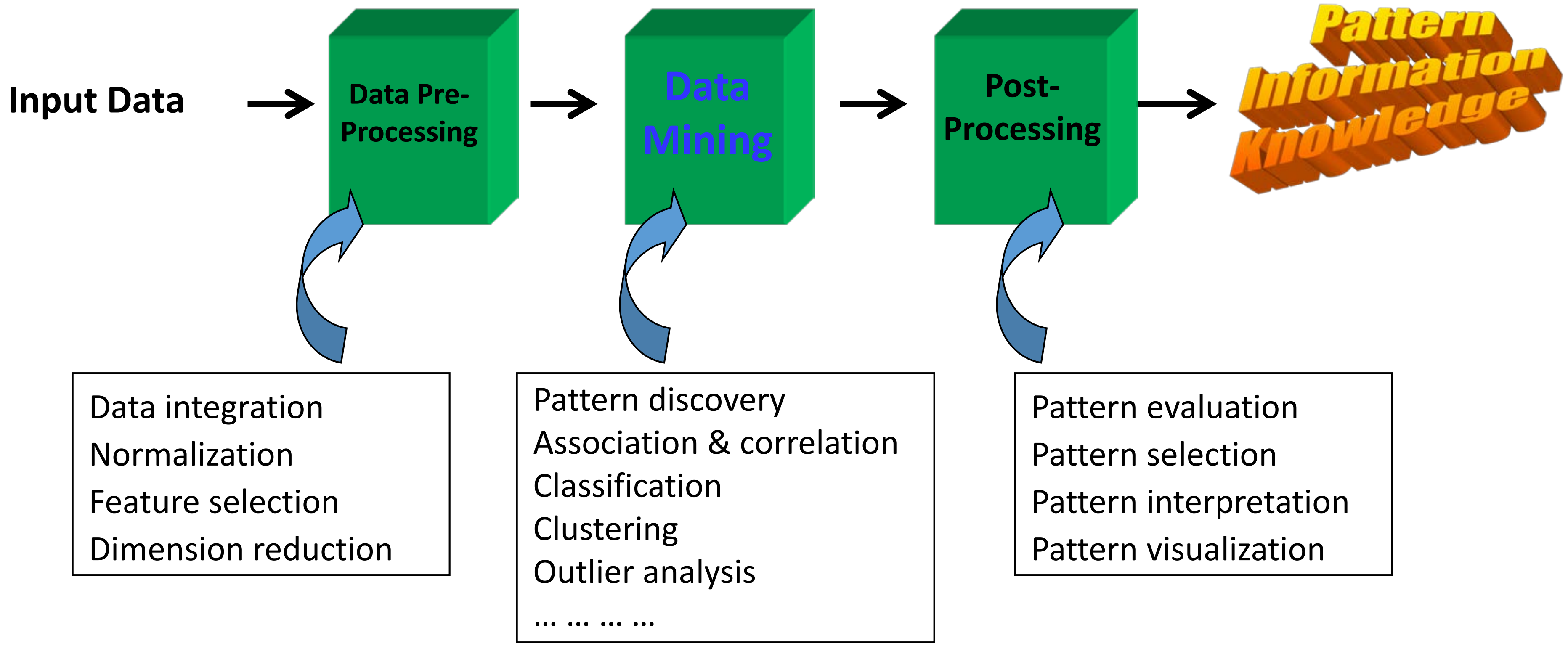
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

Knowledge discovery (KDD) process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



KDD process: View from ML and statistics



Multi-dimensional view of data mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-erformance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data mining: On what kinds of data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Data mining function: Association rule mining

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?

Data mining function: Classification and prediction

- Classification and label prediction (Supervised learning)
  - Construct models (functions) based on some training examples
  - Predict unknown class labels of data using the model
- Methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Applications
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data mining function: Cluster analysis

- Unsupervised learning (i.e. Class label is unknown)
- Group data based on their similarity (or distance)
- Principle: Maximizing intra-class similarity & minimizing interclass similarity

Data mining function: What else?

- Outlier analysis
- Sequential pattern analysis
- Trend and evolution analysis
- Structure and network analysis

Data type and representation

- Record
  - Relational records
  - Data matrix, e.g. numerical matrix, crosstabs
  - Text documents, e.g. term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - 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

Attribute type

- Can be categorized
  - ***Nominal (or Categorical)***, e.g. Type of car, Color name
  - ***Binary,*** e.g. Gender, Whether to have car or not
  - ***Ordinal***, e.g. Grade
  - ***Numerical***, e.g. Height, Temperature

or

  - ***Discrete***, e.g. Integer
  - ***Continuous***, e.g. Real

Measuring the central tendency

- **Mean (algebraic measure) (sample vs. population):**

  Note: n is **sample** size and N is **population** size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

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

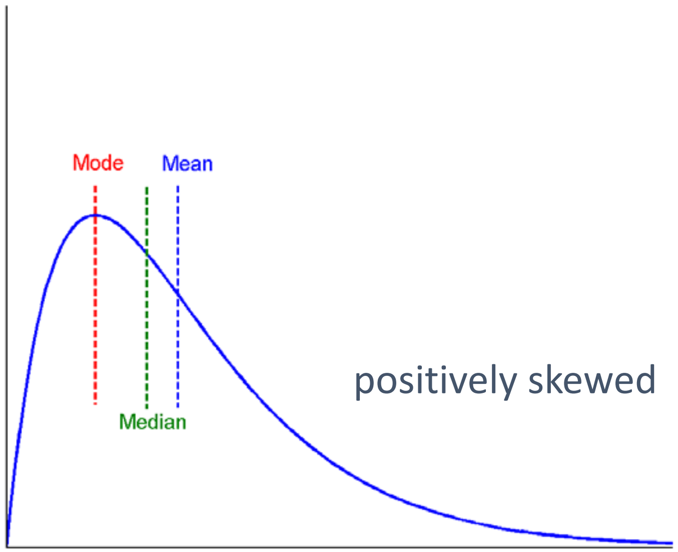
- **Median (holistic measure):**
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Computing it requires storing every data => Estimating it by one scan is an active research topic
- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $mean - mode = 3 \times (mean - median)$

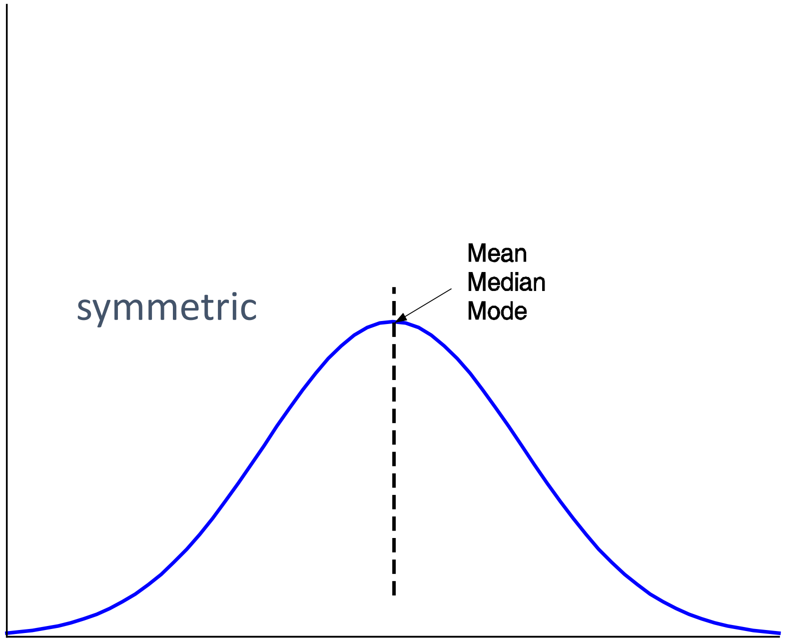
Distributive, algebraic, holistic measure

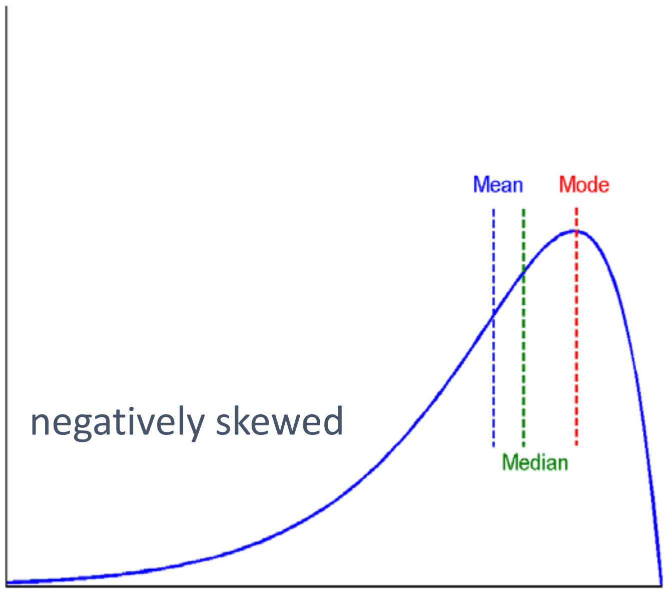
- A **distributive** measure can be computed by partitioning the data into smaller subsets (e.g. **sum and count**)
- An **algebraic** measure can be computed by applying an algebraic function to one or more distributive measures (e.g. **mean=sum/count**)
- A **holistic** measure must be computed on the entire data set (e.g. **median**)
  - Holistic measures are much more expensive to compute than distributive measures
  - Could be estimated

Symmetric vs. Skewed data

- Median, mean and mode of symmetric, positively and negatively skewed data







POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Measuring the dispersion of data

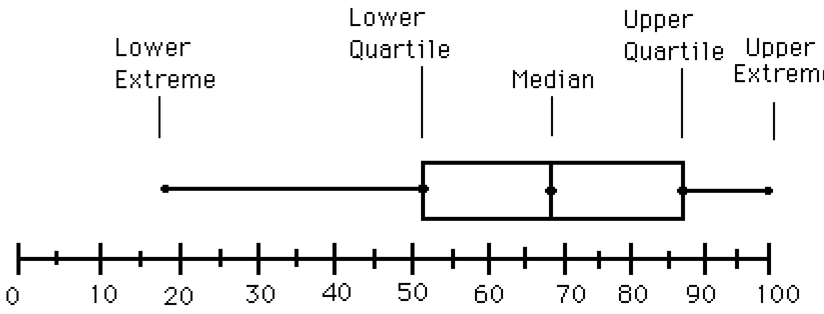
- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: IQR = Q3 – Q1
  - Five number summary: min, Q1, median, Q3, max
  - Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
  - Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

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

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

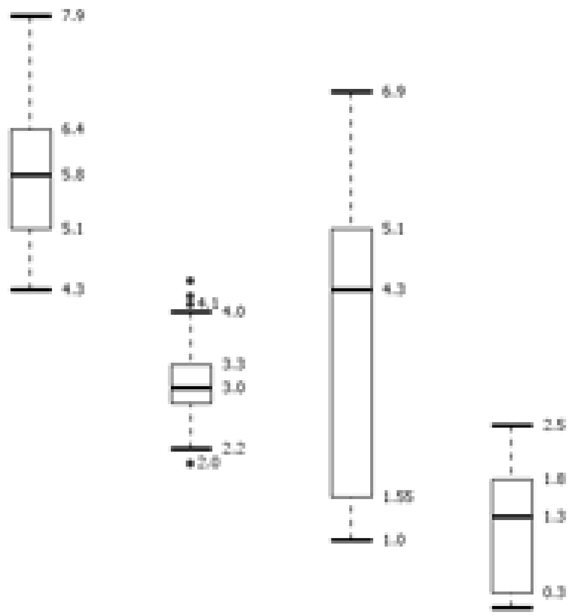
Boxplot analysis

- **Five-number summary** of a distribution
  - **Minimum, Q1, Median, Q3, Maximum**

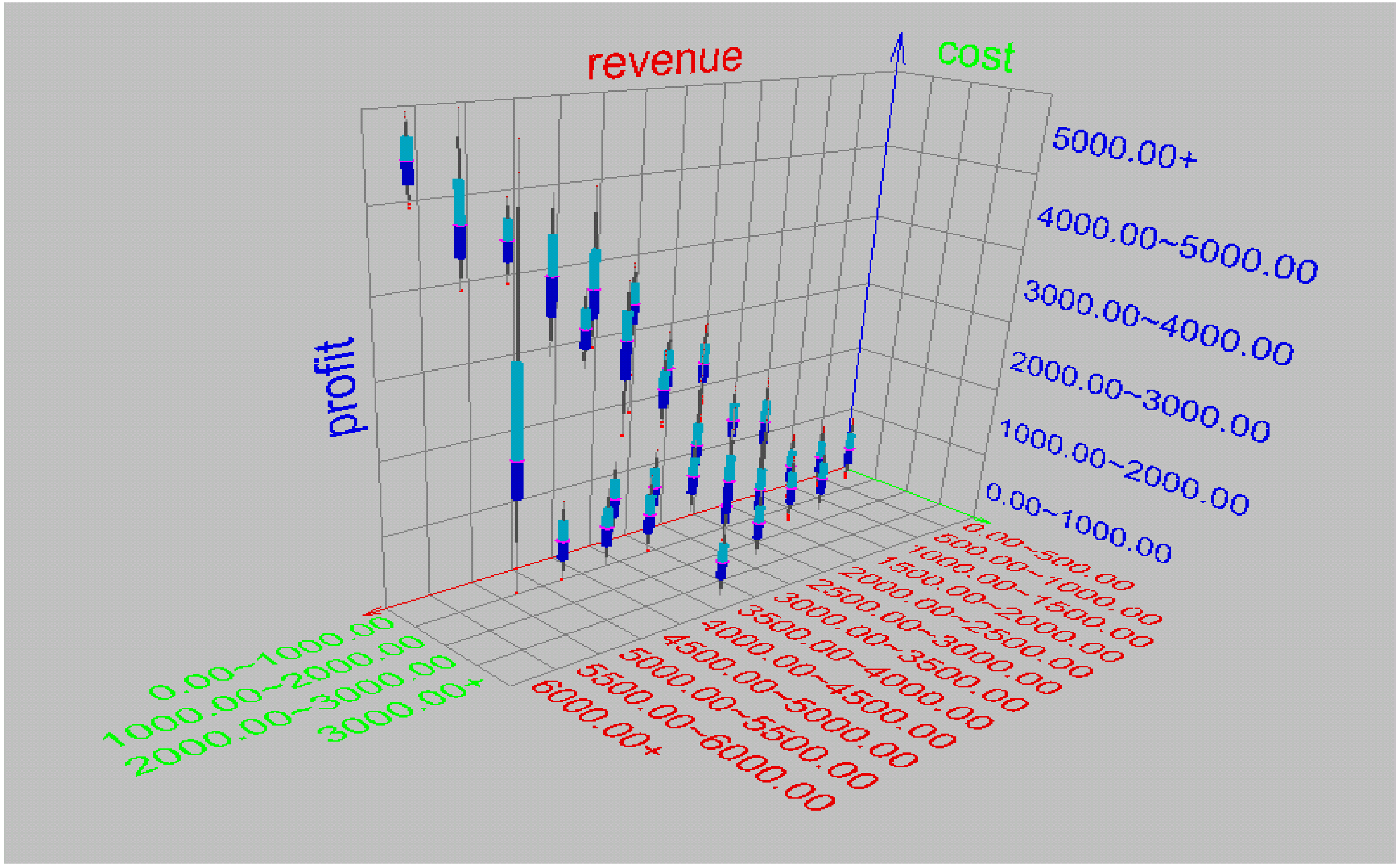


- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles,

    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

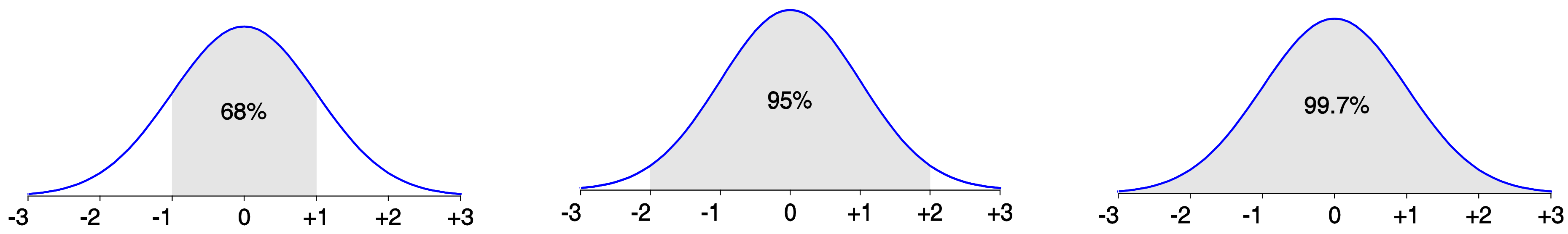


Visualization of data dispersion: 3-D boxplots



Properties of normal distribution curve

- **The normal (distribution) curve**
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it

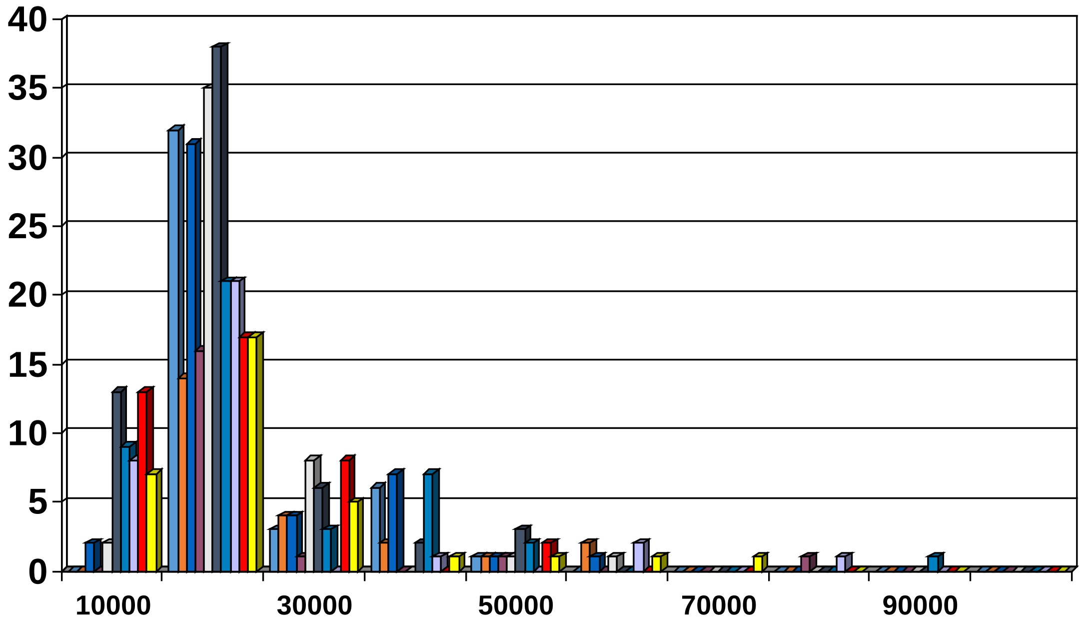


Graphic displays of basic statistical descriptions

- **Boxplot:**
  - graphic display of five-number summary
- **Histogram:**
  - x-axis => values, y-axis => frequencies
- **Quantile plot:**

  each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:**
  - graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot:**
  - each pair of values is a pair of coordinates and plotted as points in the plane

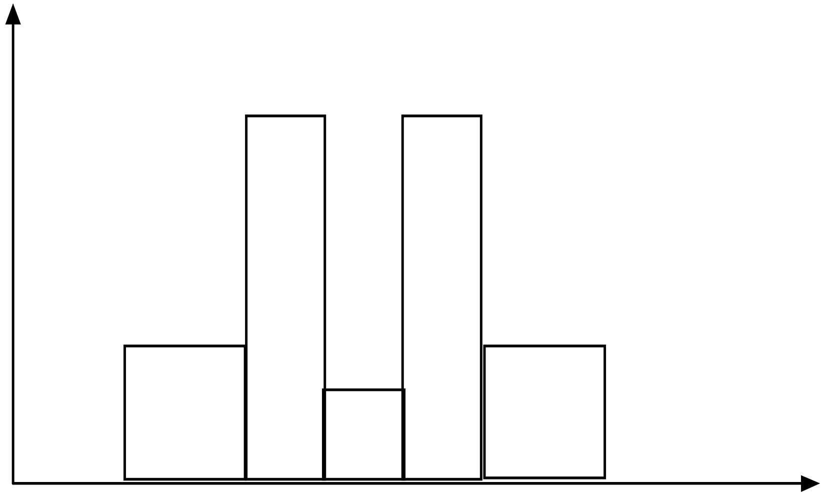
Histogram analysis

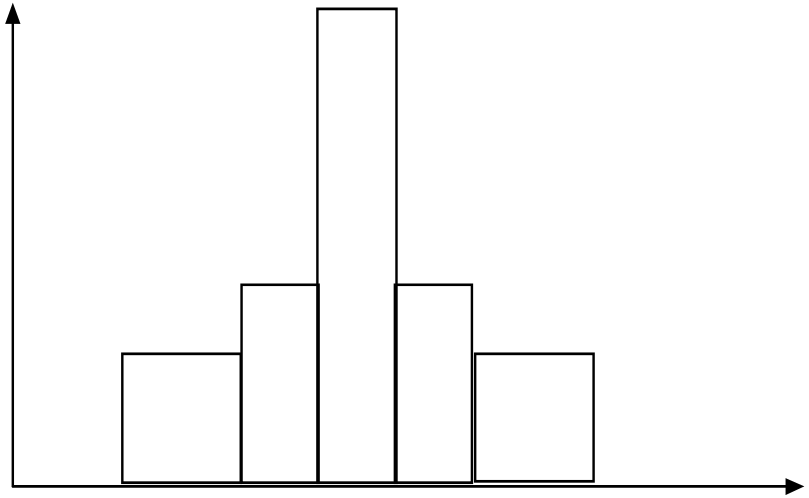


Histogram

- x-axis => values
- y-axis => frequencies
- Show overall distribution of **one dimensional** data

Histograms often tell more than boxplots

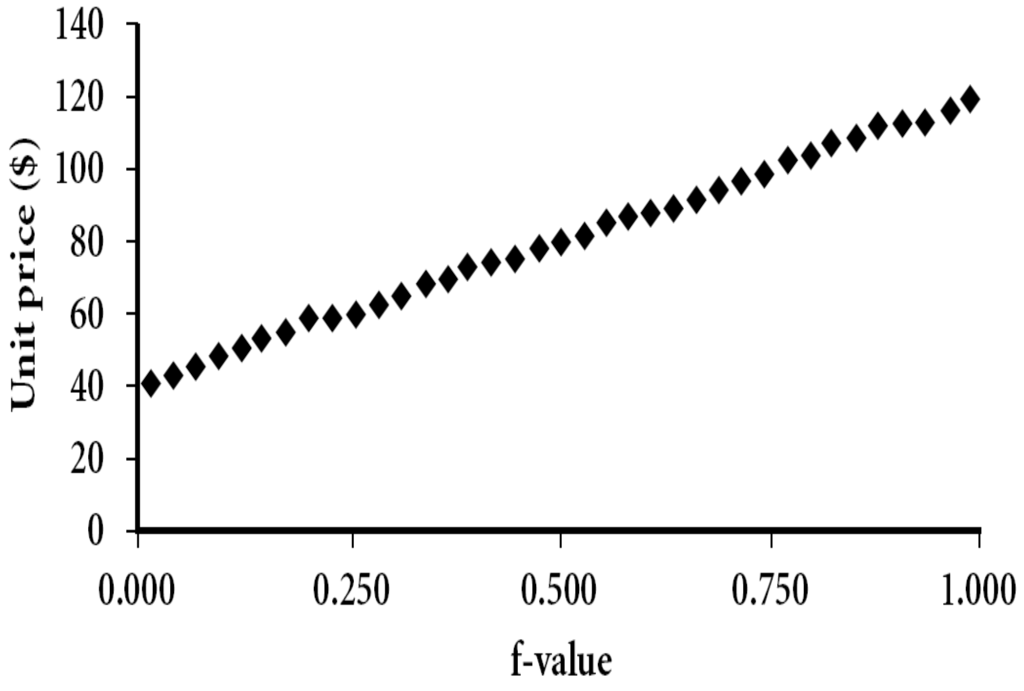




- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

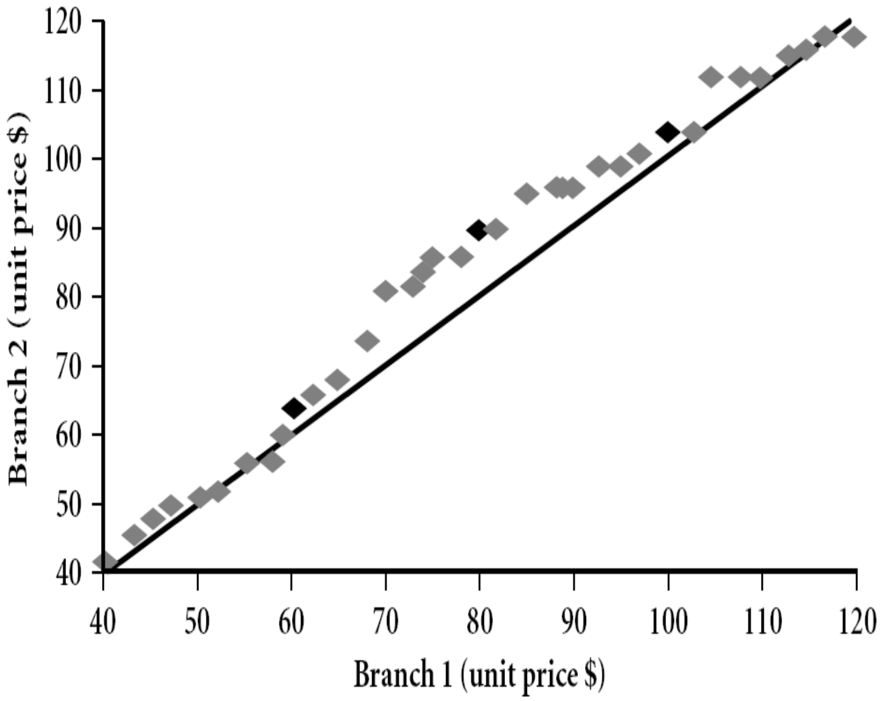
Quantile plot (generalization of quartile)

- Plots **quantile** information
  - For a data x_i sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i
  - Quartiles are 4-quantiles.
- Displays all the data for the given attribute
- Can see both the **overall behavior** and **unusual occurrences**



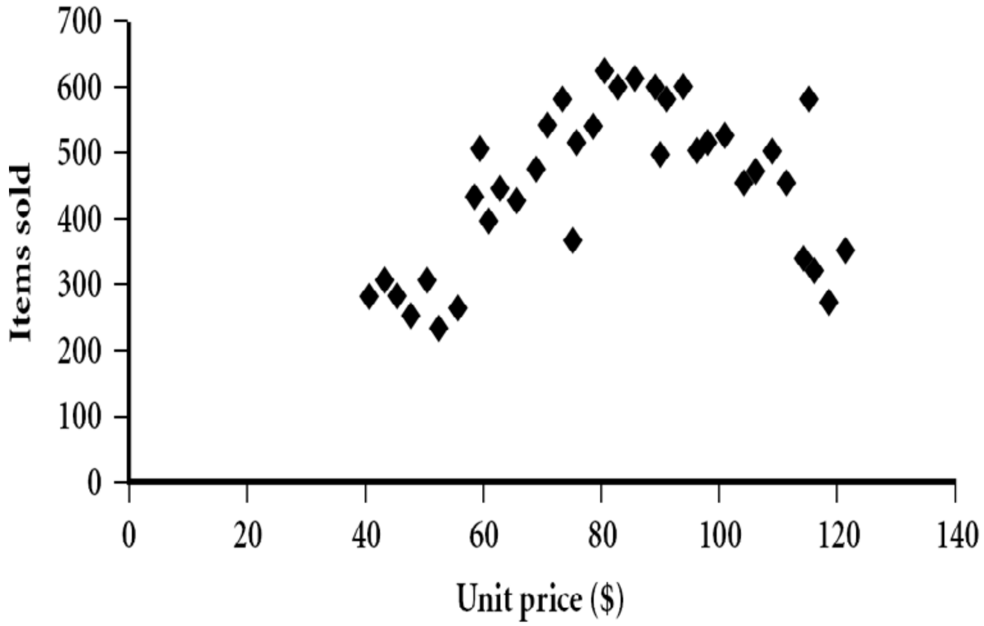
Quantile-Quantile (Q-Q) plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of 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.



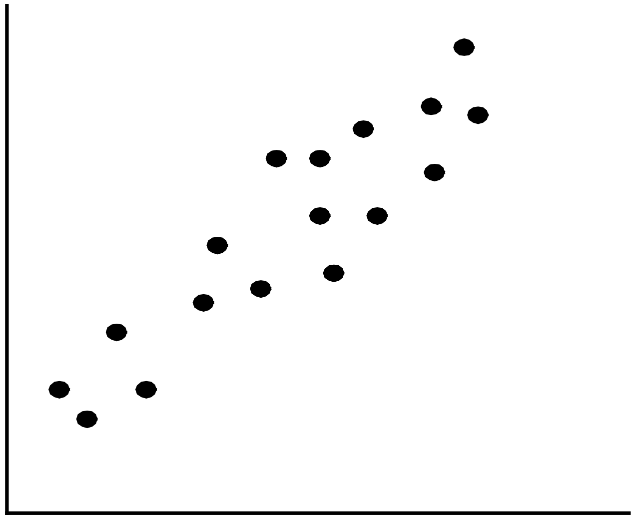
Scatter plot

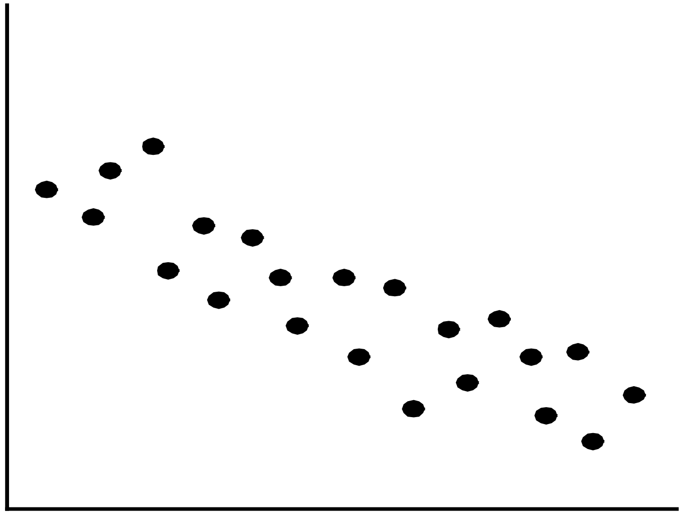
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

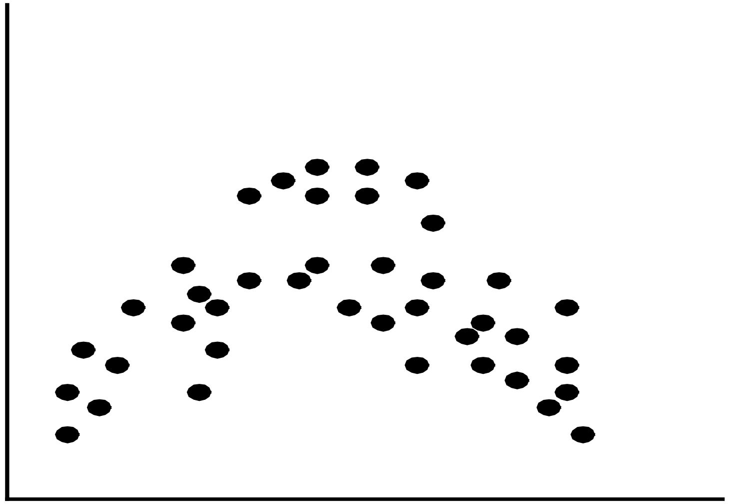


POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Positively and negatively correlated data

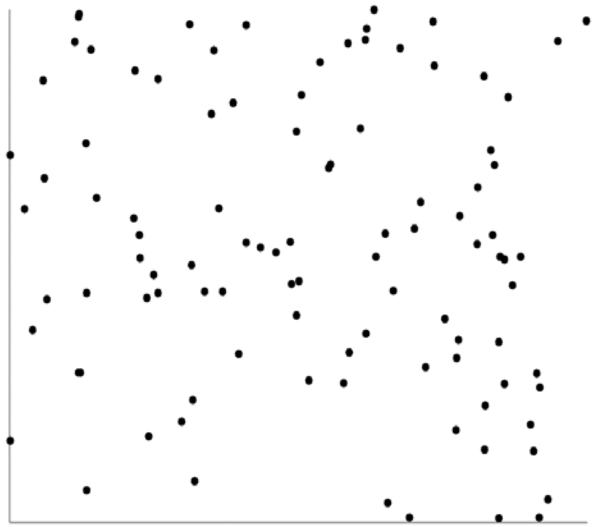


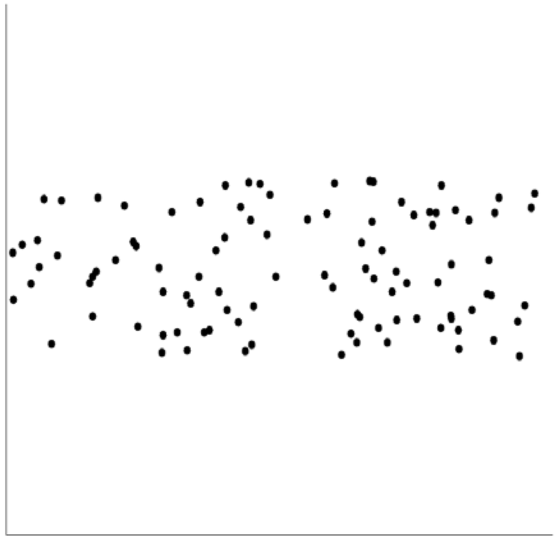


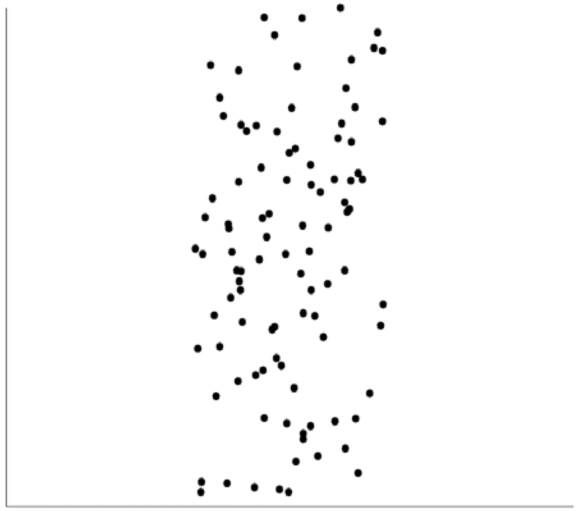


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated data







Similarity and dissimilarity

- Similarity
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (e.g. distance)
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0, and upper limit varies
- Proximity refers to a similarity or dissimilarity

Data matrix and dissimilarity matrix

- Data matrix
  - n-by-p matrix
  - n instances in p dimensions

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

- Dissimilarity matrix
  - n-by-n (triangular) matrix
  - distances between every pair of instances

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

Proximity measure for nominal attributes

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

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

- Transform to a large number of binary attributes
  - creating a new binary attribute for each of the M nominal states

Proximity measure for binary attributes

- A **contingency table** for binary data
- Distance measure for **symmetric** binary variables:
- Distance measure for **asymmetric** binary variables:
- Jaccard coefficient (similarity measure for asymmetric binary variables):

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

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

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

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

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

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

Asymmetric 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

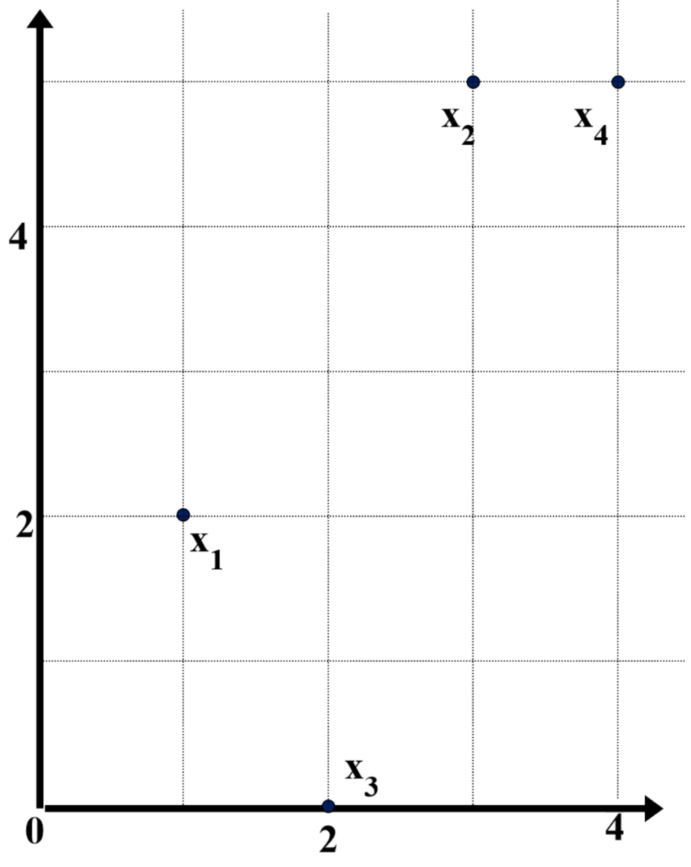
  - Gender is a symmetric attribute
  - The remaining attributes are asymmetric binary
  - Let the values Y and P be 1, and the value N be 0

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

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

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

Example: Data matrix and dissimilarity matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

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 (the distance so defined is also called L-*h* norm)
- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Special cases of Minkowski distance

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

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

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

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

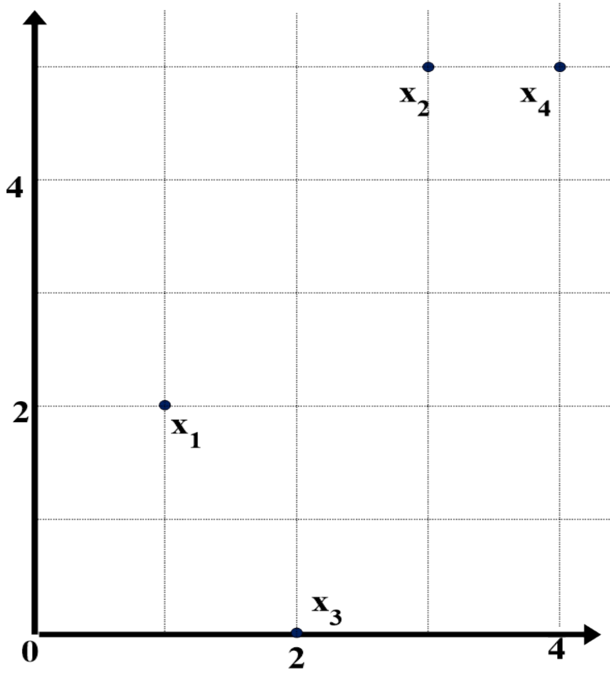
- $h \to \infty$: "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component (attribute) of the vectors

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

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

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

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
    - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{\max\_r_f - 1}$$

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

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 $\bullet$ 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

- Example: Find the similarity between documents 1 and 2.

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

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