

Information Processing Technology of Internet of Things

Chapter 1 Data Preprocessing


Wu Liu

Beijing Key Lab of Intelligent Telecomm. Software and Multimedia
Beijing University of Posts and Telecommunications

1.1 Data & its Characteristics



1.1 Data & its Characteristics

- Data Objects and Attribute Types 
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity



Types of Data Sets

■ 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

	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

■ 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:

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



Data Objects

- 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.

Attributes

- **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
 - Binary
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or “names of things”
 - *Hair_color* = {auburn, black, blond, brown, grey, red, white}
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - **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**
 - **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 (integer or real-valued)
- **Interval-scaled**
 - Measured on a scale of **equal-sized units**
 - The values of interval-scaled attributes have order and can be positive, 0, or negative.
 - Values have order
 - E.g., *temperature in C° or F°, calendar dates*
 - No true zero-point
- **Ratio-scaled**
 - 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

■ Discrete Attribute

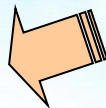
- **Has only a finite or countably infinite set of values**
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

■ 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



1.1 Data & its Characteristics

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data 
- Measuring Data Similarity and Dissimilarity



Basic Statistical Descriptions of Data

■ Motivation

- To better understand the data: central tendency, variation and spread
- basic data descriptions (e.g., measures of central tendency and measures of dispersion)
- graphic statistical displays (e.g., quantile plots, histograms, and scatter plots)
- provide valuable insight into the overall behavior of your data



Measuring the Central Tendency

■ Mean (algebraic measure) (sample vs. population):

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

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

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

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

■ Median:

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

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

■ Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula: $mean - mode = 3 \times (mean - median)$

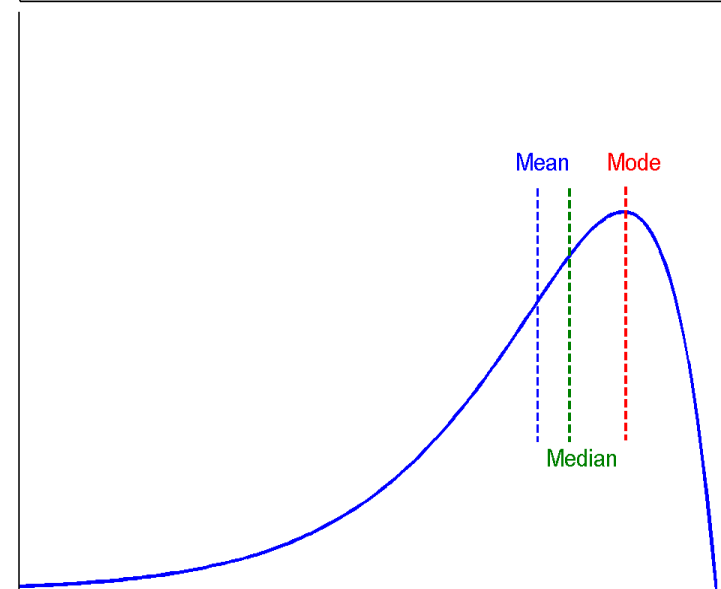
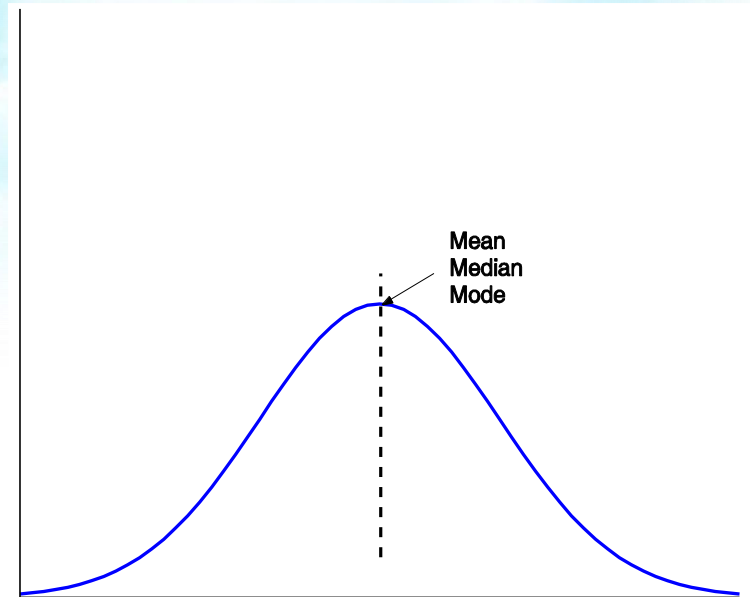
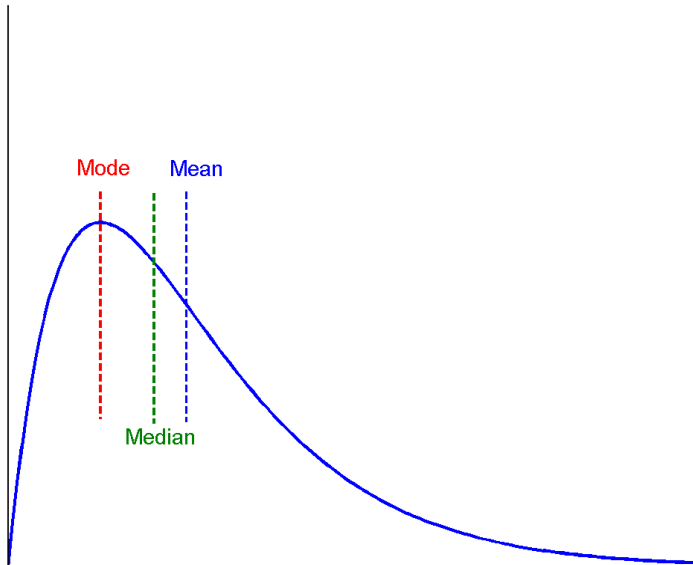
age	frequency
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700
81-110	44

Median interval →



Symmetric vs. Skewed Data

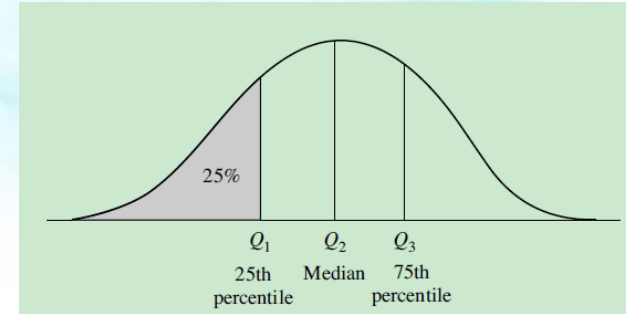
- Median, mean and mode of symmetric, positively and negatively skewed data



Measuring the Dispersion of Data

■ Quartiles, outliers and boxplots

- **Quartiles:** Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range:** $IQR = Q_3 - Q_1$
- **Five number summary:** min, Q_1 , median, Q_3 , 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 $1.5 \times IQR$



■ Variance and standard deviation (*sample: s , population: σ*)

- **Variance:** (algebraic, scalable computation)

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

- **Standard deviation s (or σ)** is the square root of variance s^2 (or σ^2)

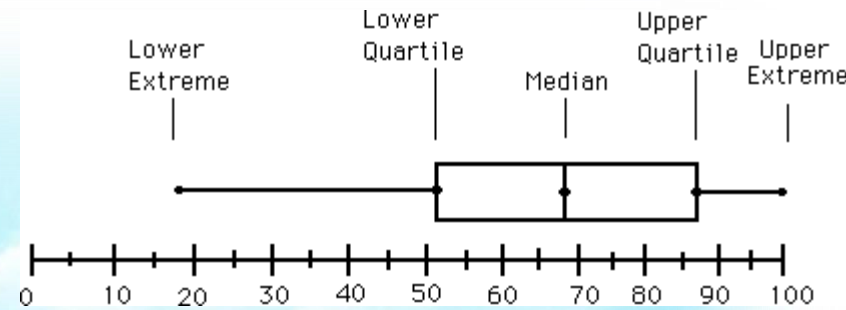


Graphic Displays of Basic Statistical Descriptions

- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis represents 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



Boxplot Analysis

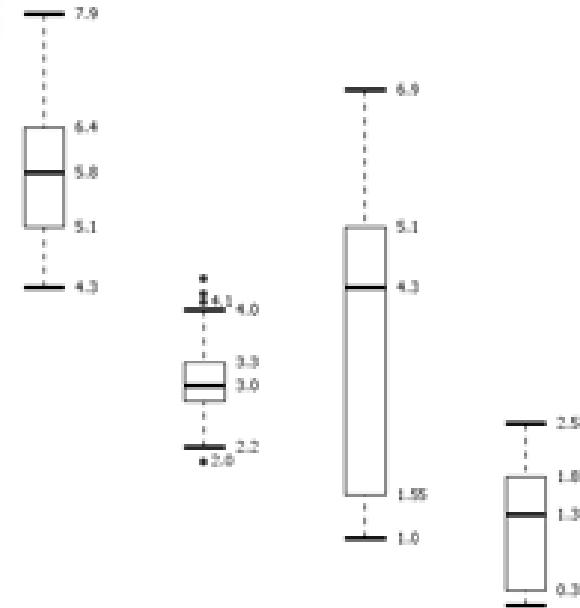


- **Five-number summary** of a distribution

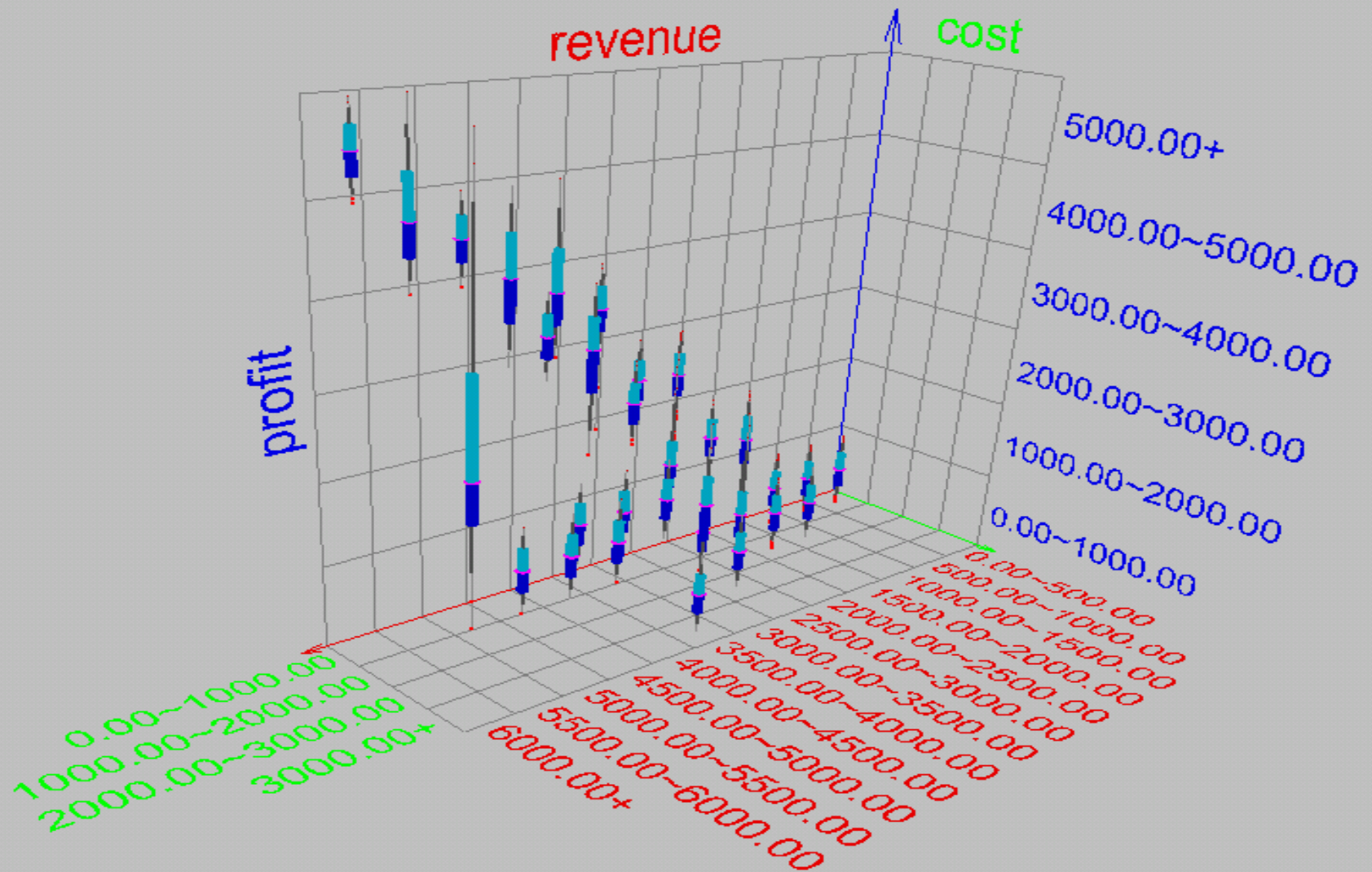
- Minimum, Q1, Median, Q3, Maximum

- **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

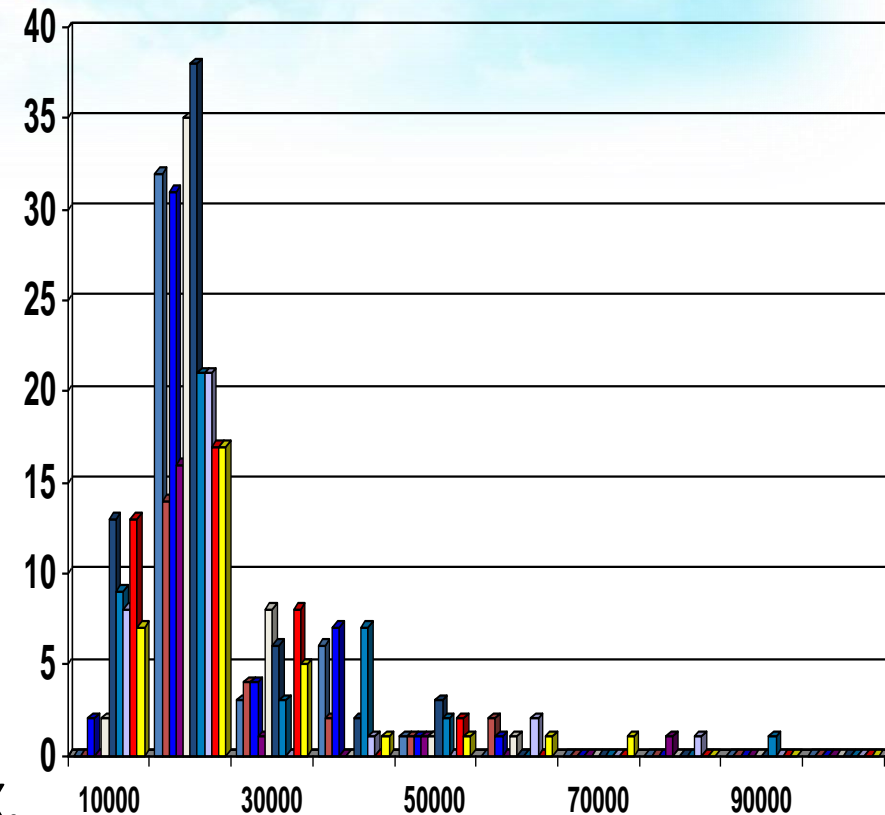


Visualization of Data Dispersion: 3-D Boxplots

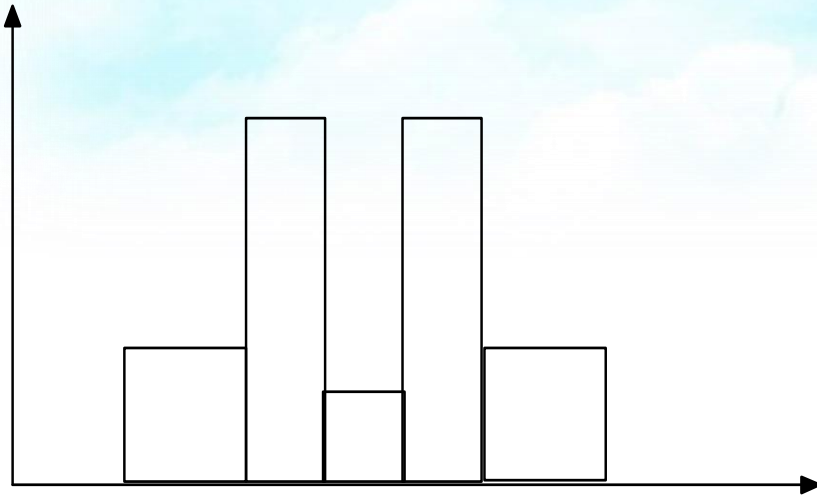


Histogram Analysis

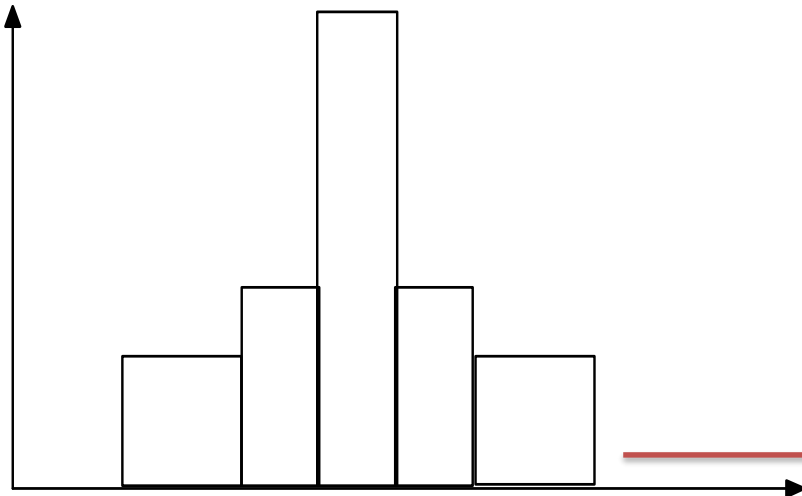
- **Histogram:** Graph display of tabulated frequencies, shown as bars. The height of the bar indicates the frequency (i.e., count) of a given attribute value.
- It shows what proportion of cases fall into each of several categories
- The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as **buckets or bins**, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width.



Histograms Often Tell More than Boxplots

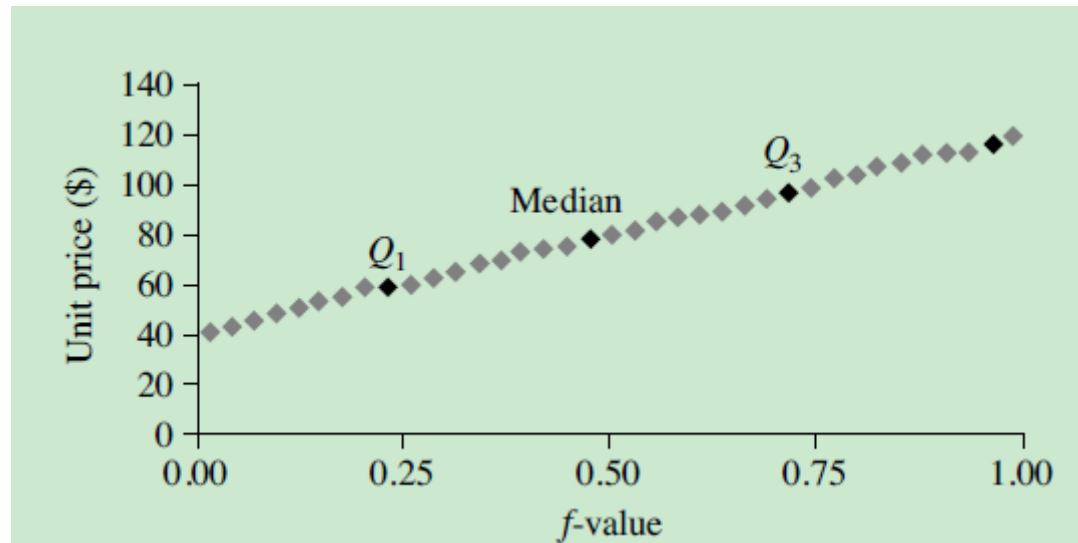


- The two histograms shown in the left may have the same boxplot representation
 - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



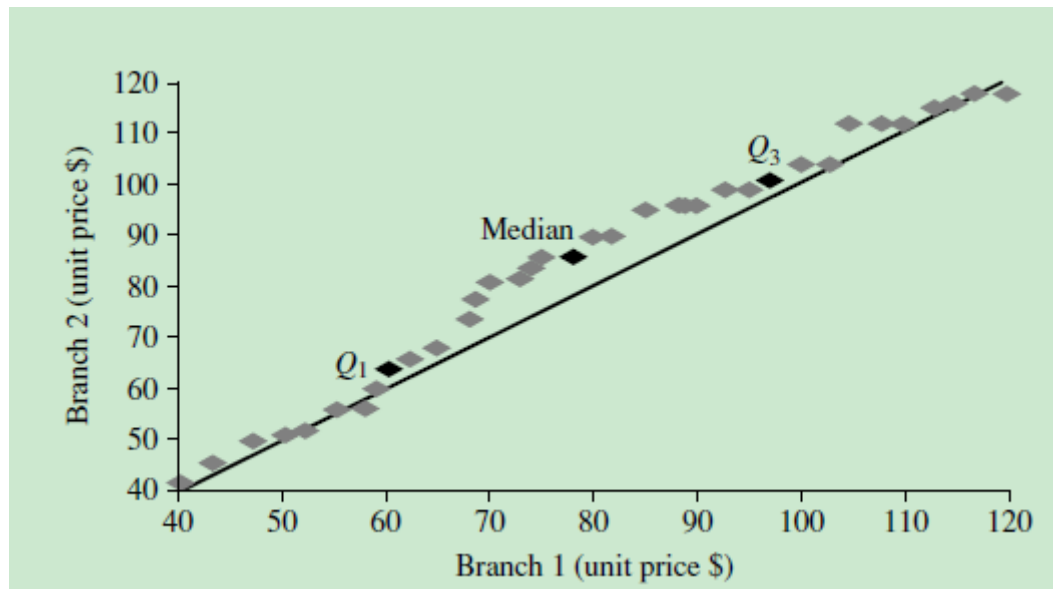
Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a univariate data distribution.
- 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



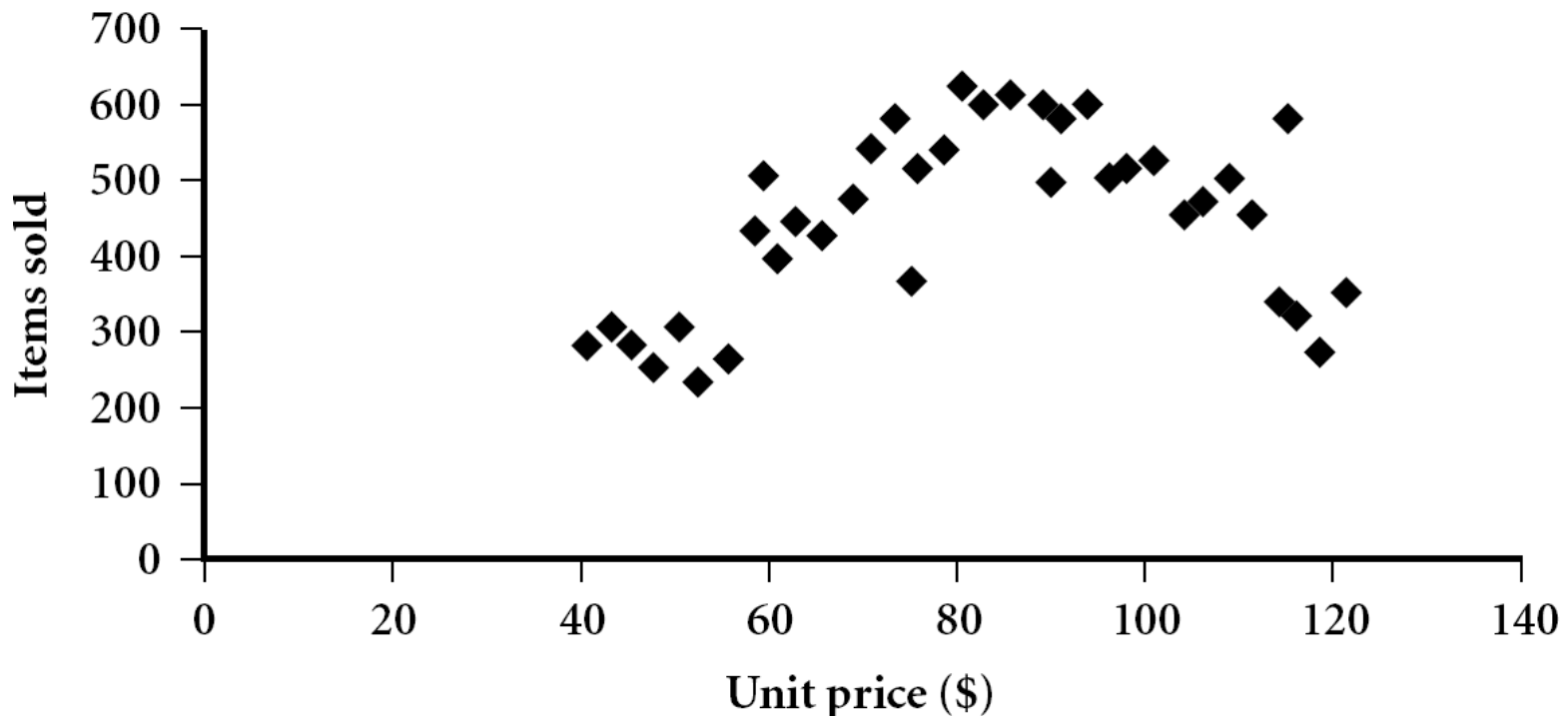
Quantile-Quantile (Q-Q) Plot

- 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?
- **Each point corresponds to the same quantile for each data set** and shows the unit price of items sold at branch 1 versus branch 2 for that quantile.
- 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.

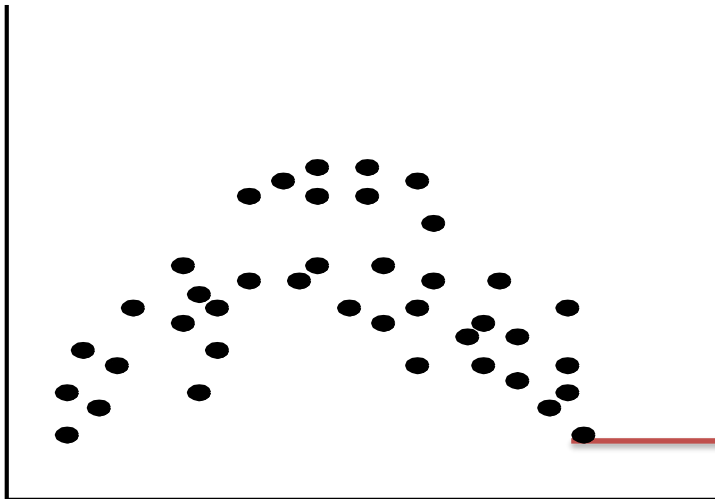
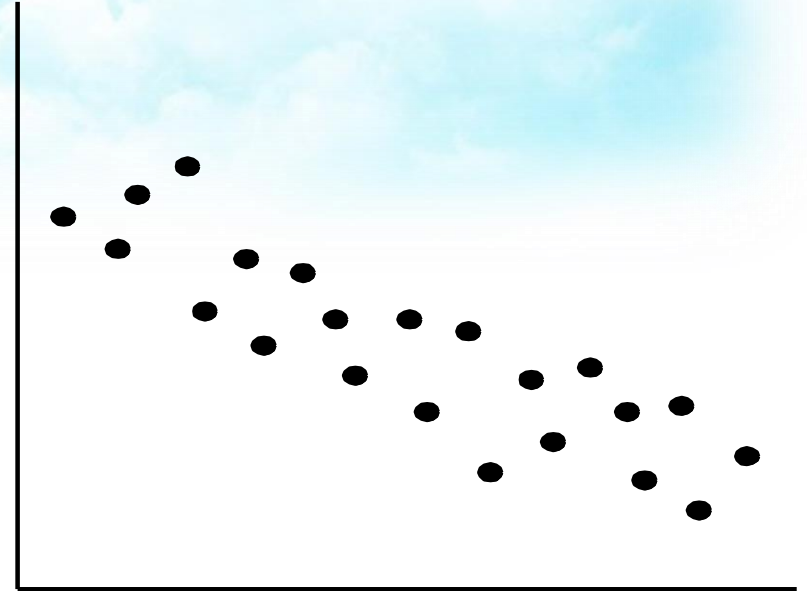
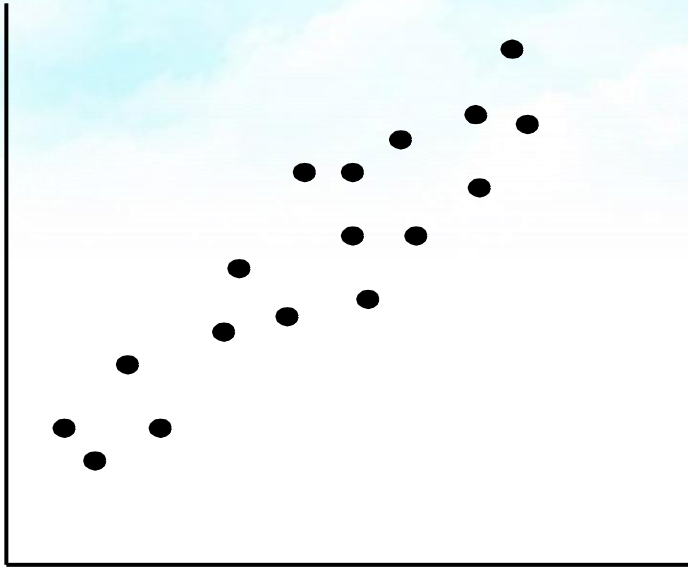


Scatter plot

- 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



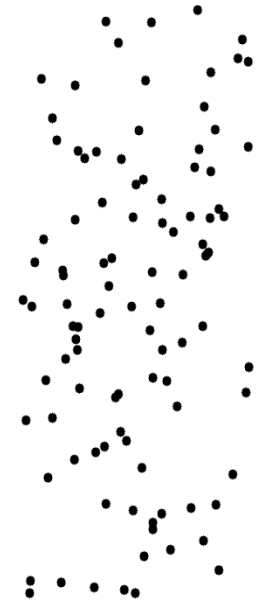
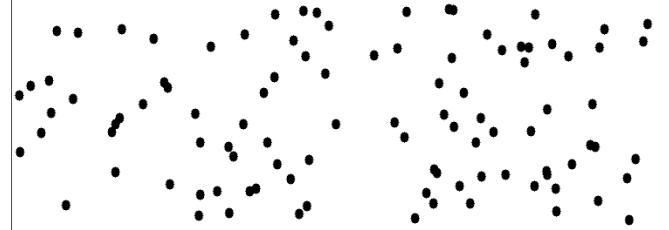
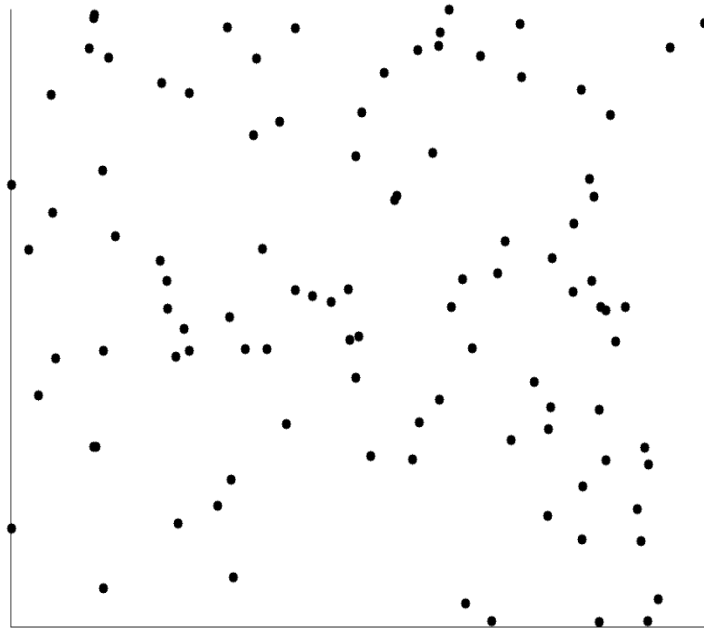
Positively and Negatively Correlated Data



- Correlations of two attributes can be positive, negative, or null (uncorrelated).
- The left half fragment is positively correlated
- The right half is negative correlated

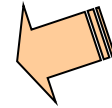


Uncorrelated Data



1.1 Data & its Characteristics

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity



Similarity and Dissimilarity

■ **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** refers to a similarity or dissimilarity



Data Matrix and Dissimilarity Matrix

- Data matrix (object-by-attribute structure)

- n objects with p attributes
- n data points with p dimensions
- Two modes

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

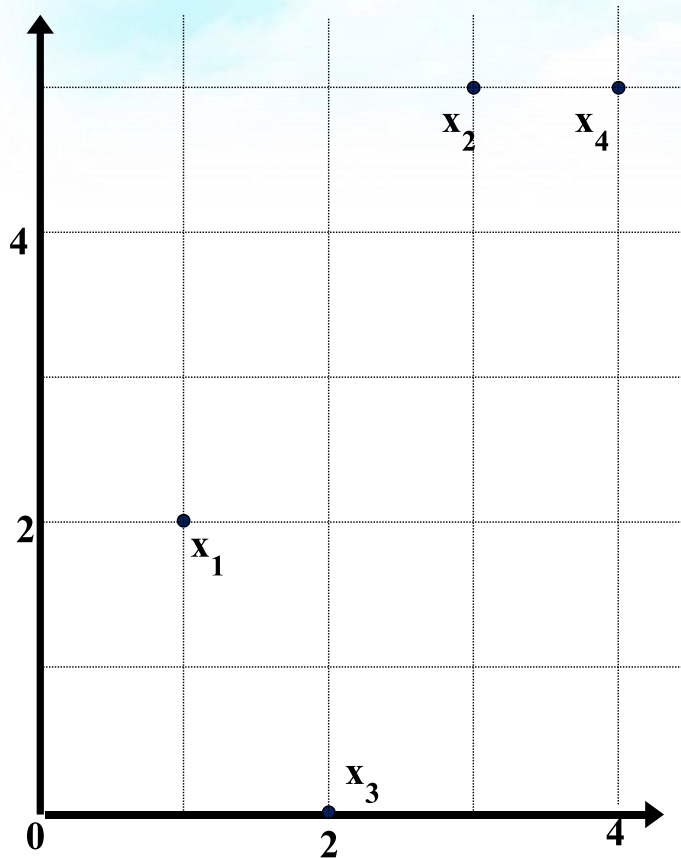
- Dissimilarity matrix (object-by-object structure)

- $d(i, j)$ is the measured dissimilarity or “difference” between objects i and j.
- n data points, but registers only the distance
- A triangular matrix
- Single mode

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



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		
$x3$	2.24	5.1	0	
$x4$	4.24	1	5.39	0



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 : number of matches, p : total number of variables

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

Object Identifier	test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

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



$$\begin{bmatrix} 0 \\ 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Since here we have one nominal attribute, *test-1*, we set $p = 1$



Proximity Measure for Binary Attributes

- A contingency table for binary data
- Distance measure for symmetric binary variables:
- Distance measure for asymmetric binary variables: the number of negative matches, t , is considered unimportant and is thus ignored
- 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

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

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

$$sim(i, j) = \frac{q}{q + r + s} = 1 - d(i, j)$$



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 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

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

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 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** (L_1 norm) **distance**
 - E.g., the Hamming distance: the number of bits that are different between two binary vectors

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

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

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

- $h \rightarrow \infty$. **“supremum”** (L_{\max} norm, L_{∞} norm, Chebyshev distance) distance.
 - This is the maximum difference between any component (attribute) of the vectors

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



Example: Minkowski Distance

Dissimilarity Matrices

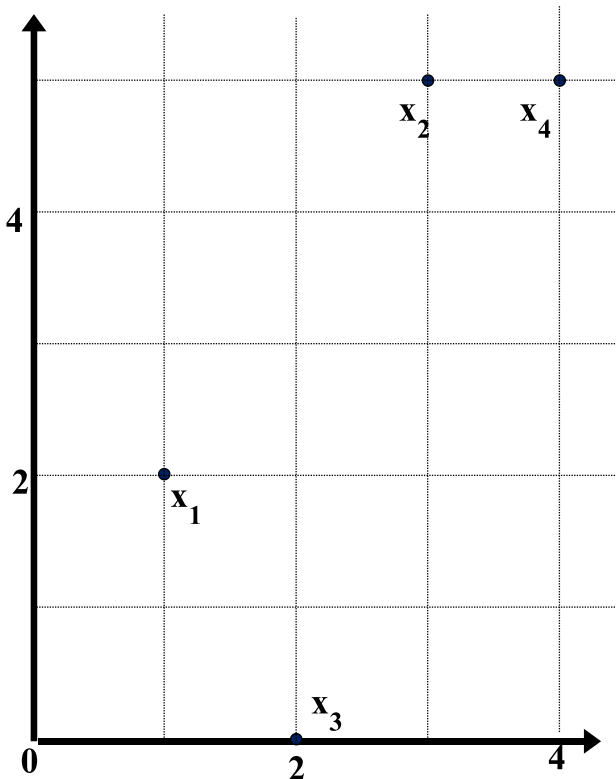
point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

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



Proximity Measure for Ordinal Attributes

- 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 $r_{if} \in \{1, \dots, M_f\}$
 - 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



Example: Ordinal Attributes

- **Dissimilarity between ordinal attributes.**
test-2: There are three states for test-2: **fair, good, and excellent**, that is, $M_f = 3$.
- For step 1, if we replace each value for test-2 by its rank, the four objects are assigned the ranks 3, **1, 2, and 3**, respectively.
- Step 2 normalizes the ranking by mapping **rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0**.
- For step 3, we can use, say, the Euclidean distance, which results in the following dissimilarity matrix:

Object Identifier	test-2 (ordinal)
1	excellent
2	fair
3	good
4	excellent

0			
1.0	0		
0.5	0.5	0	
0	1.0	0.5	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 $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
 - 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	teamcoach		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, $\|d\|$: is the Euclidean norm 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.

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

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

$$d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$\|d_1\| = (5*5 + 0*0 + 3*3 + 0*0 + 2*2 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (3*3 + 0*0 + 2*2 + 0*0 + 1*1 + 1*1 + 0*0 + 1*1 + 0*0 + 1*1)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = 0.94$$

