Data Mining:

Concepts and Techniques

— Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign
Simon Fraser University

Chapter 2: Getting to Know Your Data

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

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
- 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
  - Curse of dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Distribution
  - Centrality and dispersion

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**: a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Attributes (database term) also called:
  - dimensions (in data warehousing)
  - features (in machine learning)
  - variables (in statistics)
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attributes

- Observed values for a given attribute are called observations.
- A set of attributes used to describe a data object is called an attribute vector (or feature vector)
- A probability distribution (mass/density function) is univariate if it involves one attribute, and bivariate if it involves two attributes, and so on.
- The type of an attribute is determined by the set of possible values: nominal, binary, etc.

Attribute Types

- **Nominal:** categories, states, symbols or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
  - Sometimes we use numbers as symbols: for *hair_color*, we can assign 0 for *black*, 1 for *brown*, etc. Although numeric, these values are considered nominal (since not quantitative)
  - Some statistical measures do not apply to nominals, like the mean and median.
- **Binary** (a value absent/present)
  - Nominal attribute with only 2 states (0 and 1)
  - It is referred to as Boolean if the states indicate *true* and *false*. e.g. *smoker*: 1 if patient smokes, 0 otherwise.
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

Attribute Types

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings
  - Magnitude? Drink size from small to medium, how much larger?
  - e.g. *army ranks*: *private*, *private first class*, *specialist*, *corporal*, and *sergeant* (ordinal rather than nominal).
  - Ordinal values might be obtained from the discretization of numeric quantities, by splitting the value range into a finite number of ordered categories.

    e.g. weather temperature brackets:

    Hot: 80++

    Warm: 75-79

    Nice: 66-74

    Cool: 40-65

    Cold: --40

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval-scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - *Mean*, *median*, and *mode* measure are applicable.
  - No true zero-point, so we cannot say a value is a multiple of another, e.g. although 10C˚ is double of 5C˚ but we cannot say it is twice as warm as 5C˚. (semantic-wise)
- **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˚, 0 K˚ = -273.15 C˚, K:Kelvin temp).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- Attributes also can be classified into discrete and continues.
- **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
  - Values are countable if they can mapped one-one to natural numbers (whether finite or infinite)
  - 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

Chapter 2: Getting to Know Your Data

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

Basic Statistical Descriptions of Data

- <u>Motivation</u>
  - To better understand the data: central tendency, variation and spread
- <u>Measures of central tendency</u>
  - Given an attribute, where do most of its values fall?
  - Mean, median, mode, and midrange.
- <u>Dispersion of the data</u>
  - How are the data spread out?
  - Useful in identifying outliers.
  - Range, quartiles, and interquartile range, boxplots, variance and standard deviation.
- <u>Graphical display of statistical measures</u>
  - Bar charts, pie charts, and line graphs.
  - Data summaries: quantile plots, histograms, and scatter plots.

Measuring the Central Tendency

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

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

  - Arithmetic mean.

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

  - Weighted values? to indicate significance. Use weighted arithmetic mean.

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

  - Mean is sensitive to extreme values (outliers).

    Alternative: trimmed mean; chopping extreme values.

Measuring the Central Tendency

- Example: computing the mean

Example 2.6 **Mean.** Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is \$58,000. ■

Measuring the Central Tendency

- Median:
  - Skewed data? Median is a better measure of the center of data.
  - Middle value if odd number of values, or average of the middle two values otherwise

Example 2.7 **Median.** Let's find the median of the data from Example 2.6. The data are already sorted in increasing order. There is an even number of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is, $\frac{52+56}{2} = \frac{108}{2} = 54$. Thus, the median is $54,000.

Suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of $52,000. ■

Measuring the Central Tendency

- Median:
  - The median is expensive to compute when we have a large number of observations, because it requires sorting.
  - However, it can be easily approximated if the values are grouped into intervals with frequencies (number of values in each interval).
  - Approximated by interpolation (for *grouped data*):

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

Given table to the right: in total there are 3194 values. The median is supposed to be the average of the 1597 and the 1598 values, since, this value lies in the interval 21-50, it is the median interval.

Lower boundary of the median interval

number of values in entire dataset

sum of frequencies of all intervals lower than median interval

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

frequency of median interval

width of median interval

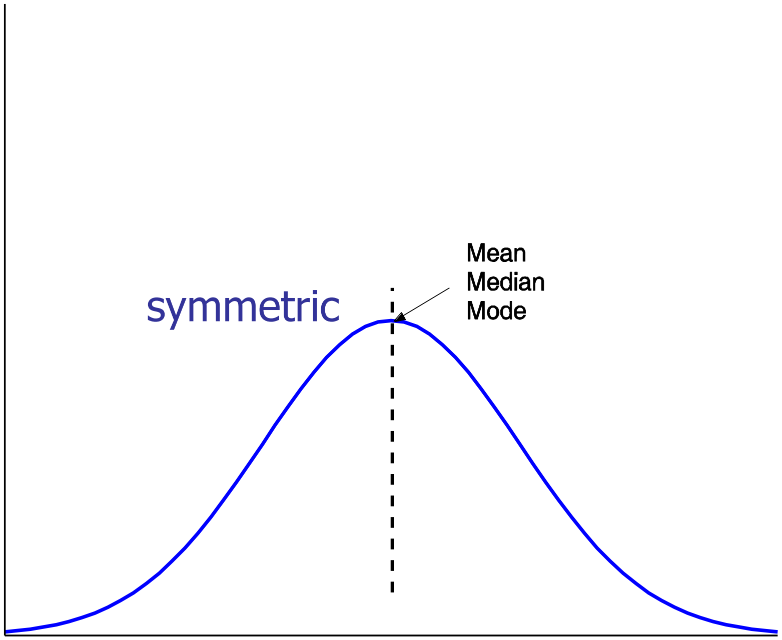
Measuring the Central Tendency

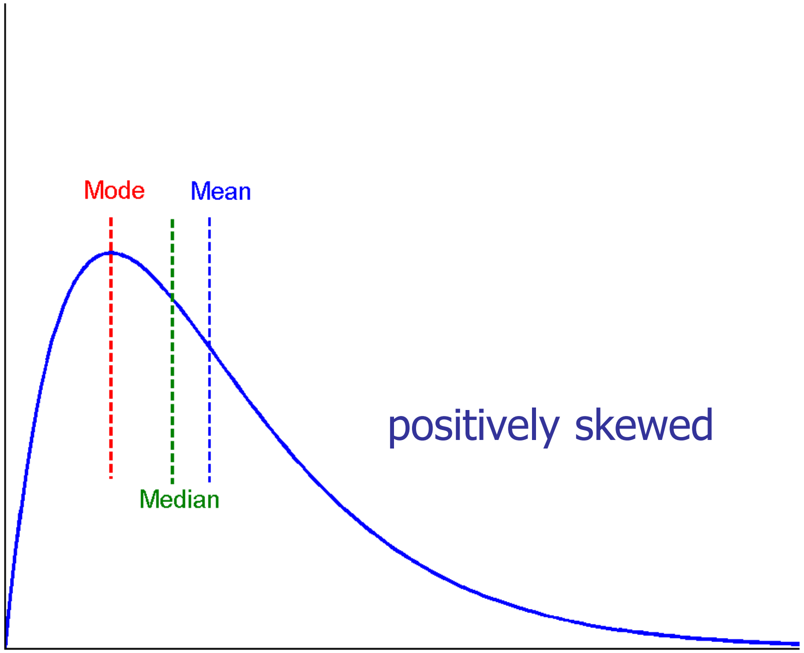
- Mode
  - Value that occurs most frequently in the data
  - It is applicable to qualitative and quantitative attributes.
  - It is possible for the greatest frequency to correspond to several different values:
    - One value: unimodal dataset
    - Two values: bimodal dataset
    - Three values: trimodal dataset
    - Multimodal: two or more
    - At the extreme when each distinct value occurs only once, there is no mode.

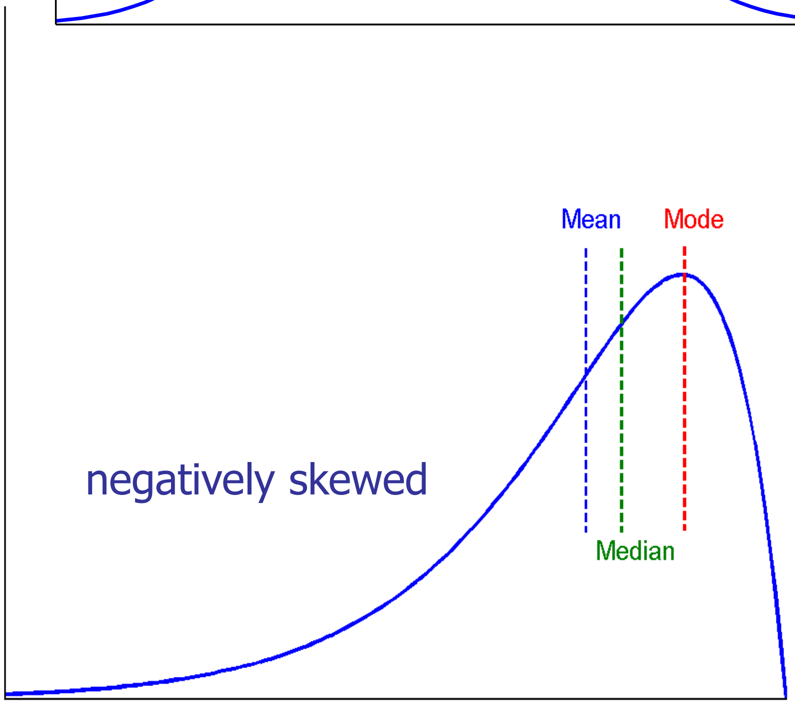
Example 2.8 **Mode.** The data from Example 2.6 are bimodal. The two modes are \$52,000 and \$70,000. ■

Symmetric vs. Skewed Data

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







Measuring the Central Tendency

- For unimodal numeric data, that are moderately skewed the following empirical formula can be used to estimate the mode in terms of the mean and median.

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

- Another less popular measure of central tendency is the midrange: given a set of values, it is the average of the largest and lowest values.

Measuring the Dispersion of Data

- Let $x_1, x_2, \ldots, xN$ be a set of observations for some numeric attribute , X.
- The range of the set is the difference between the largest and smallest values.
- Quartiles are data points that divide the sorted set of values (or the data distribution) into consecutive equal-sized subsets (define size as the count of values in the subset).
- Definition: The k^{th} q-quantile for a given data distribution is the value x such that at most k/q of the values are less than x and at most $(q-k)/q$ of the values are greater than x, where k is an integer such that $0 < k < q$.
- There are $(q-1)$ q-quantiles. e.g. there are four 5-quantiles.

Measuring the Dispersion of Data

- Most common quantiles:
  - The 2-quantile, say the point x , divides the set in half, half of the values are less than x, and the other half are more than x, i.e. x is the median.
  - The 4-quantiles are three points that divide the set into 4 parts, each contains ¼ of the values. 4-quantiles are also known as quartiles.
  - The 100-quantiles are commonly known as percentiles, they divide the distribution into 100 equal-sized consecutive sets.
  - Q_1 denotes the 1st quartile, which is also the 25th percentile. It cuts off the lowest 25% of the data, or the highest 75%.

Measuring the Dispersion of Data

- Median = 2-Quantile = 2nd Quartile = 50th Percentile)
- Plot of Median, Quartiles, and Percentiles:

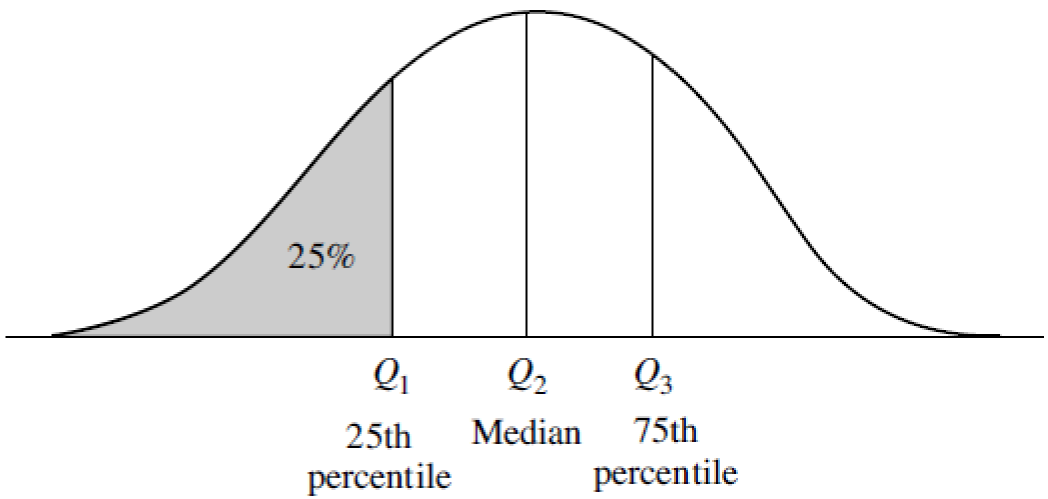


Figure 2.2 A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

Measuring the Dispersion of Data

- Quartiles can be used to figure out a distribution's center, spread, and shape.
  - Q_1 denotes the 1^{st} quartile, which is also the 25^{th} percentile. It cuts off the lowest 25% of the data, or the highest 75%.
  - Q_3 denotes the 3^{rd} quartile, which is also the 75^{th} percentile. It cuts off the lowest 75% of the data, or the highest 25%.
  - Q_2 is the 2^{nd} quartile, the 50^{th} percentile, and the median. It gives the center of the distribution.
  - The distance between the 1^{st} and 3^{rd} quartiles is a simple measure of spread, called the interquartile range (IQR), defined as:

$$IQR = Q_3 - Q_1$$

  - **Outlier**: as a rule of thumb, is a value higher than, or lower than, 1.5 x IQR of the Q_3 or Q_1 respectively.

Measuring the Dispersion of Data

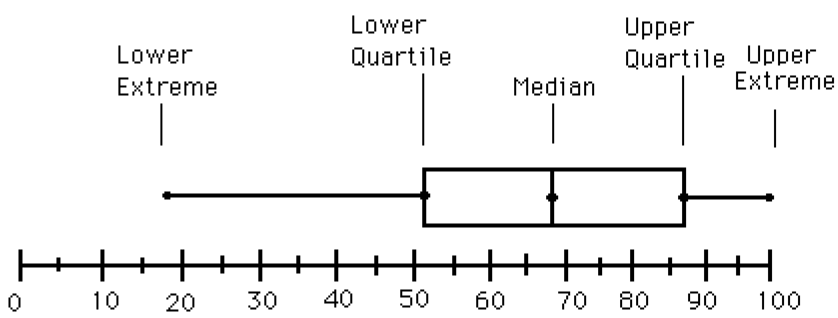
- Example: computing quartiles, and interquartile range
- 12 values in sorted order; the 1^{st} quartile is 12/4 = 3^{rd} value, the 2^{nd} quartile is 12/2= 6^{th} value, 3^{rd} quartile is the 36/4 = 9^{th} value.

Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Example 2.10 **Interquartile range.** The quartiles are the three values that split the sorted data set into four equal parts. The data of Example 2.6 contain 12 observations, already sorted in increasing order. Thus, the quartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, $Q_1 = \$47{,}000$ and Q_3 is \$63,000. Thus, the interquartile range is $IQR = 63 - 47 = \$16{,}000$. (Note that the sixth value is a median, \$52,000, although this data set has two medians since the number of data values is even.) ■

Measuring the Dispersion of Data

- No single measure of spread is very useful for describing skewed distributions.
- The **Five-number summary**: (min, Q_1, median, Q_3, max) does a better job descripting the shape of the distribution.
- The five-number summary is visualized using a **boxplot**: ends of the box are the quartiles; median is a line marked inside the box; two lines that extend from the box to min and max (called whiskers), outlier points added individually.



Measuring the Dispersion of Data
Boxplot Example

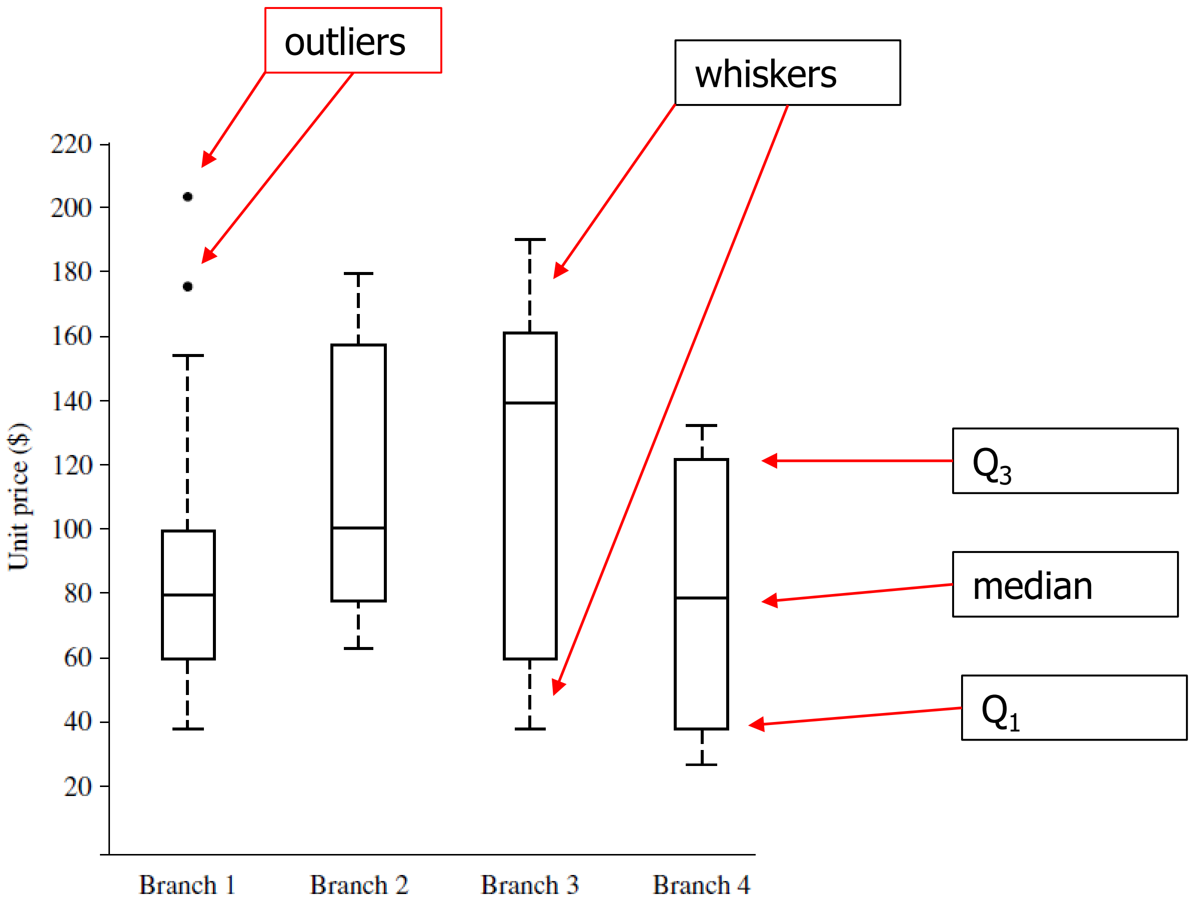
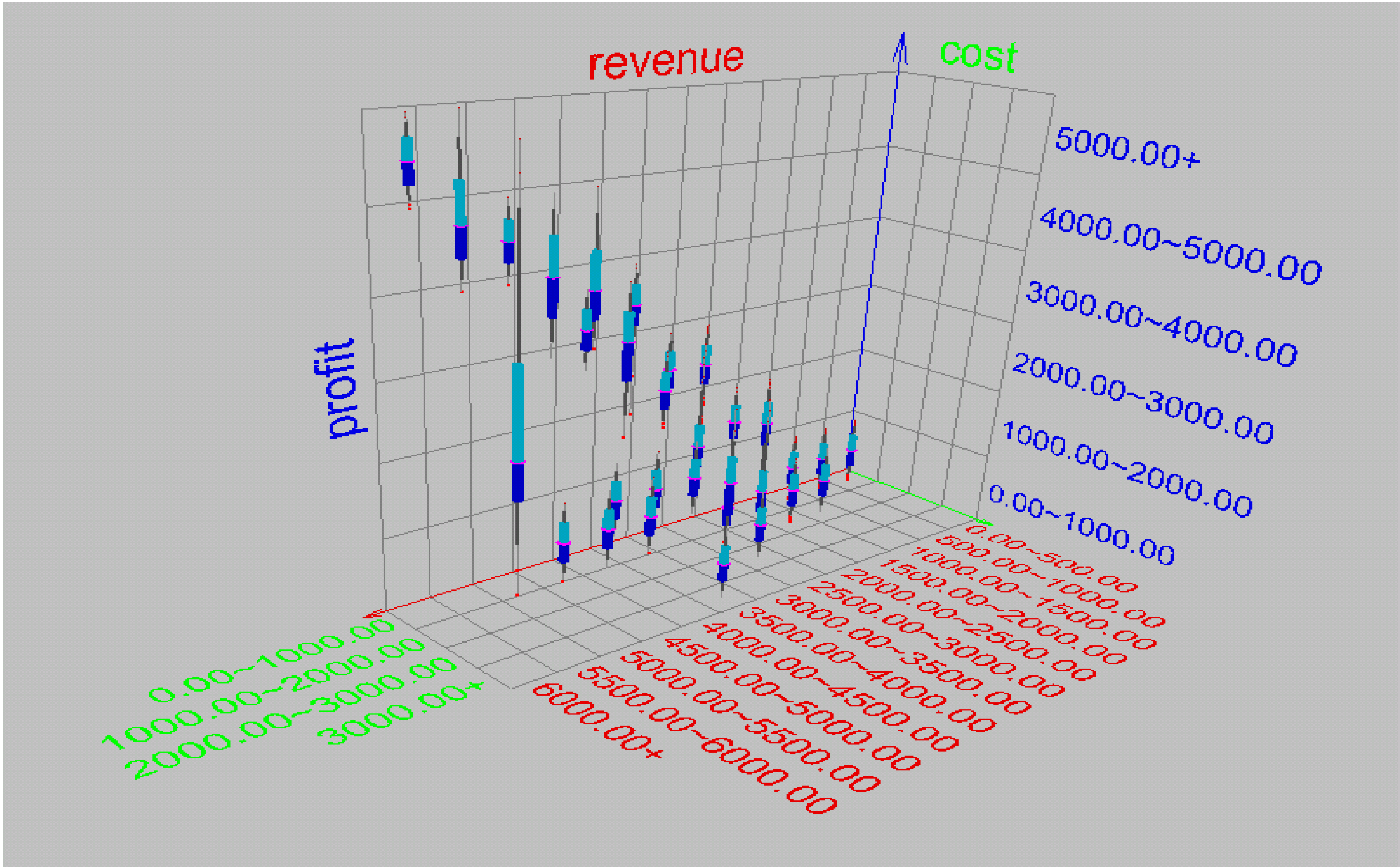


Figure 2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

Visualization of Data Dispersion: 3-D Boxplots



Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

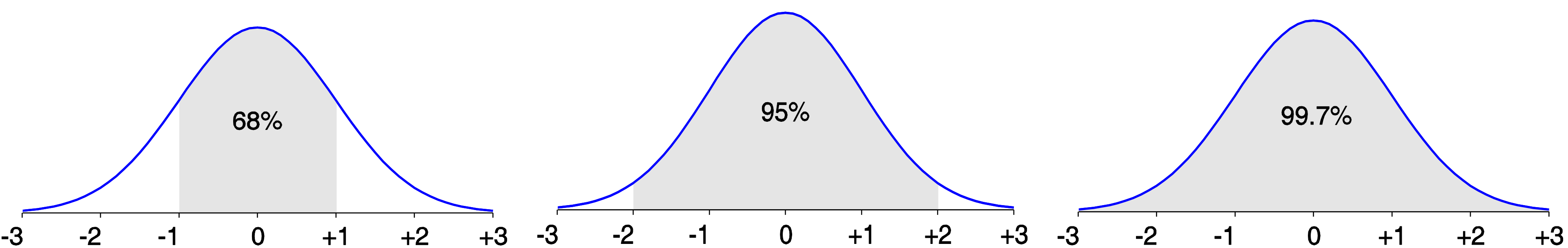
Sample: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

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

  - **Standard deviation** *s (or σ)* is the square root of the variance s^2 (or σ^2)
  - σ is a good measure of the spread of the data set, since an observation is unlikely to be more than several standard deviations from the mean.
  - $\sigma = 0$ when there is no spread, all values are identical. Otherwise, $\sigma > 0$.

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
- When $\mu=0$ and $\sigma=1$, it is called the standard normal distribution.



Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphical display of five-number summary.
- **Histogram**: summarizes a distribution by plotting its values against their frequencies. It can be applied to univariate distributions (one attribute X).
  - If X is numeric, then its range is partitioned into disjoint subranges. The histogram shows the counts of values in each subrange.
  - If X is nominal, the histogram is called a bar chart, shows the frequency for each value (no ranges).

Histogram Analysis

Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

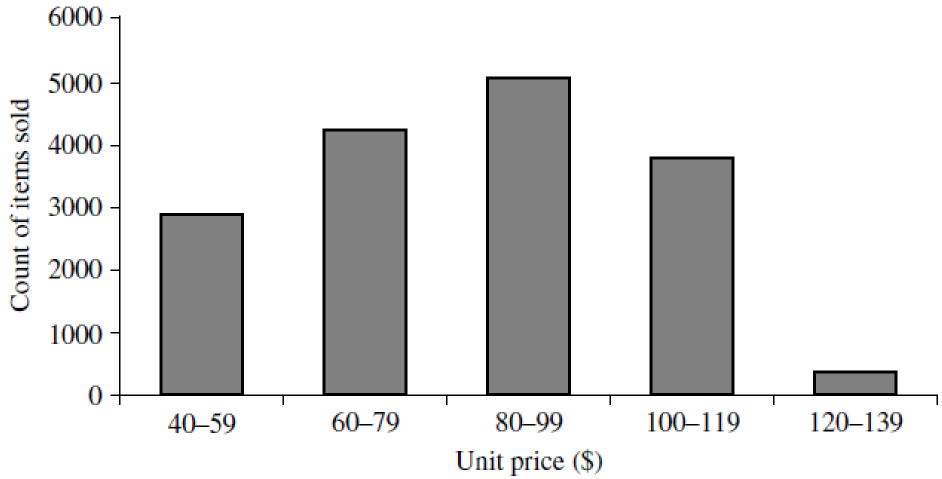
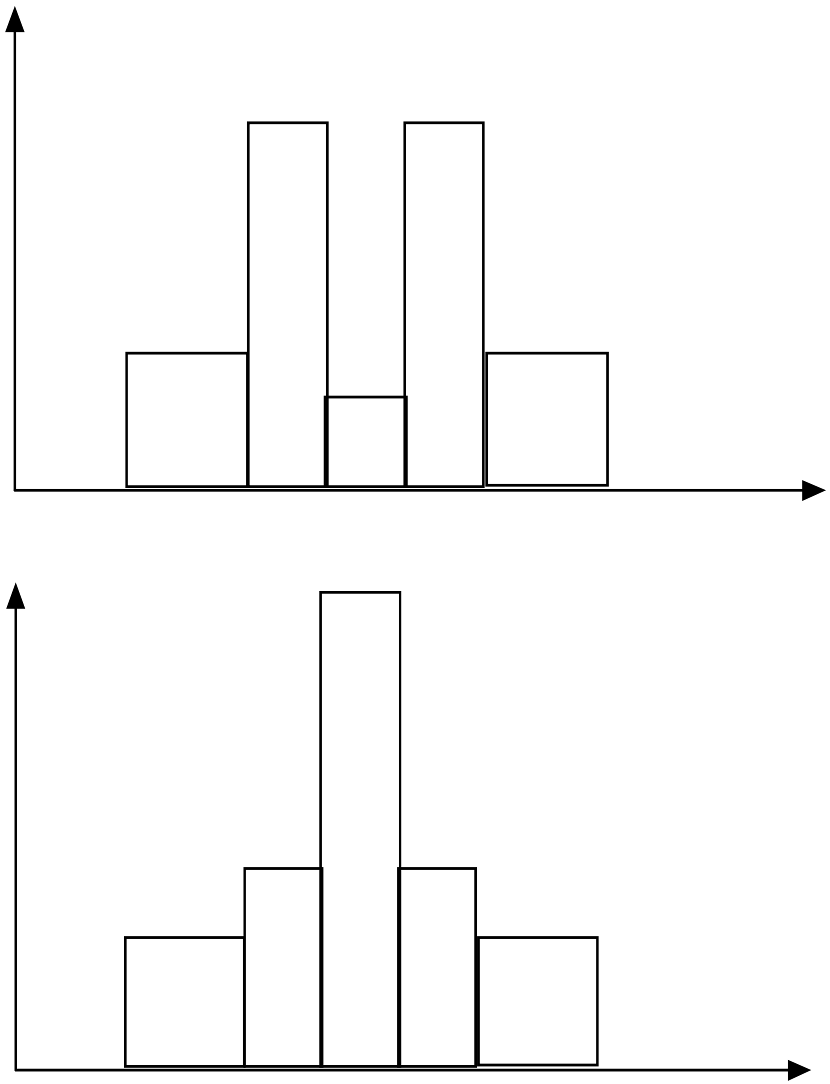


Figure 2.6 A histogram for the Table 2.1 data set.

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

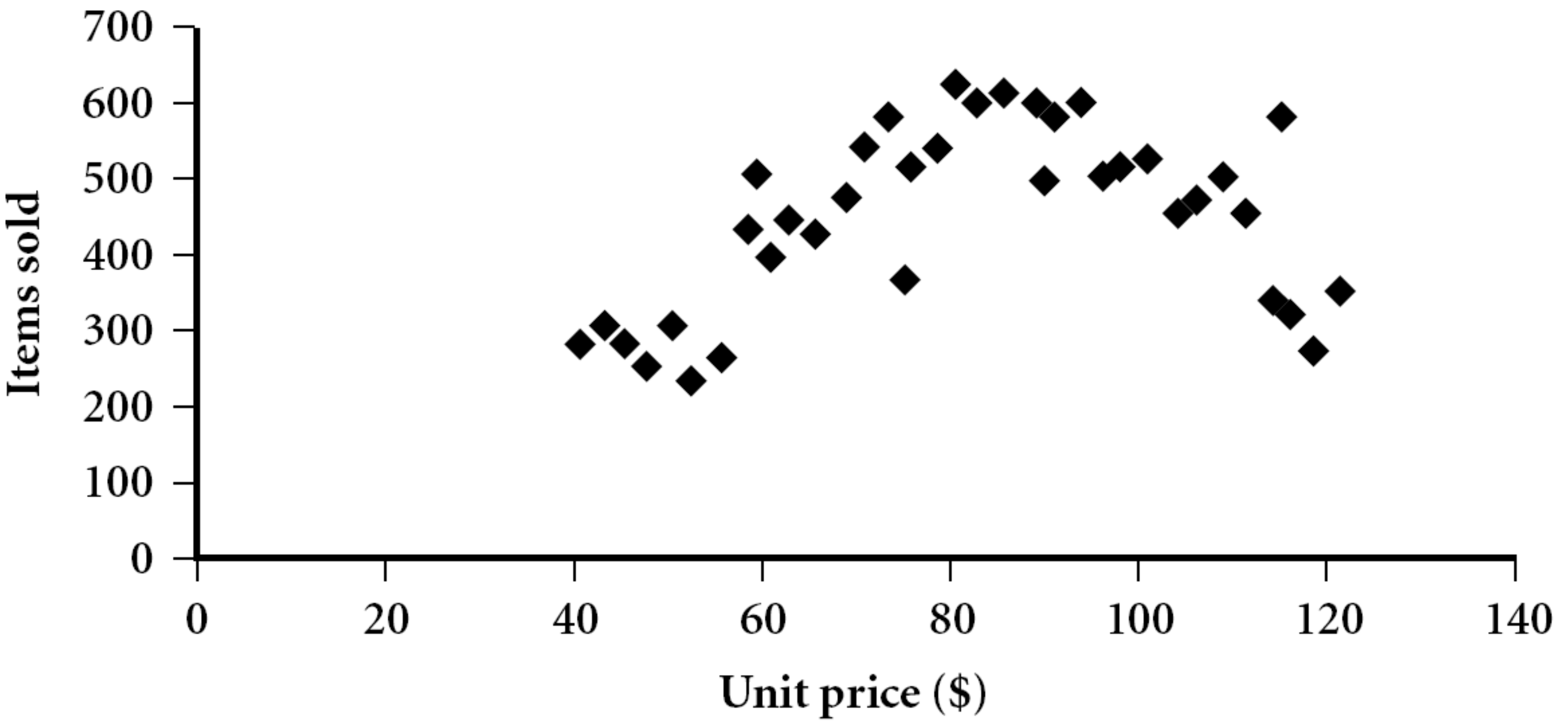
Graphic Displays of Basic Statistical Descriptions

- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane.

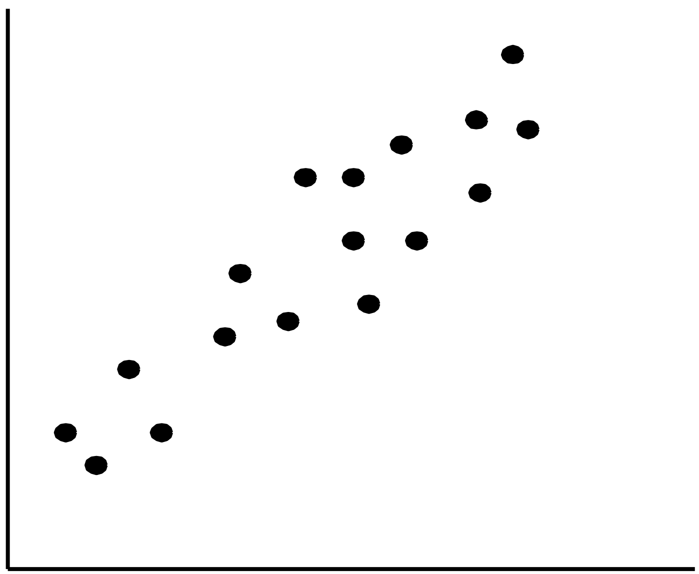
- It is a useful method to see if there are any clusters, outliers, or correlation in data.

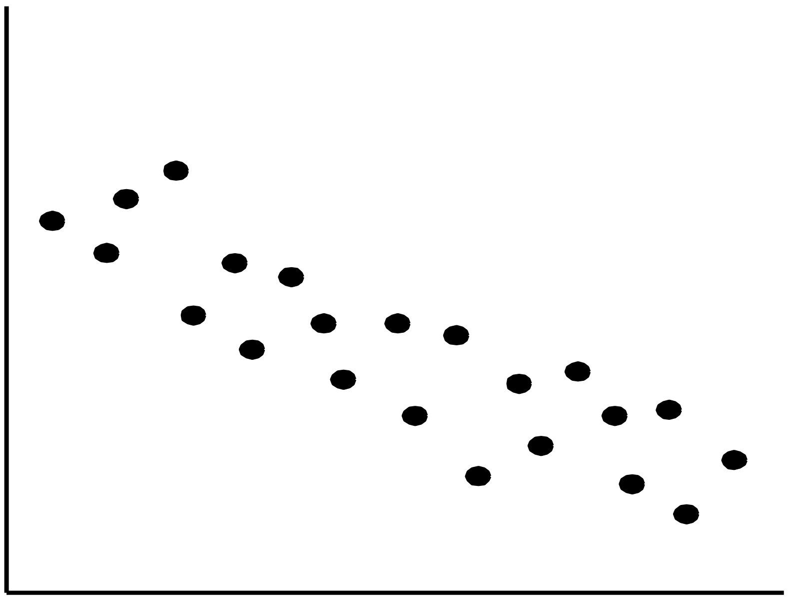
Scatter plot

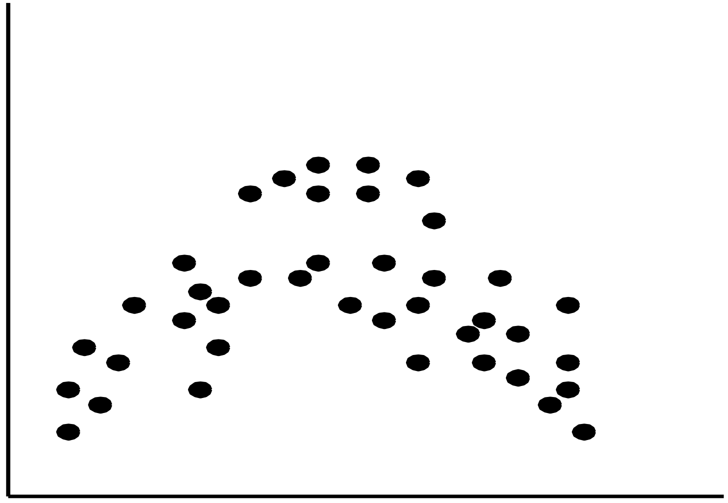
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Positively and Negatively Correlated Data

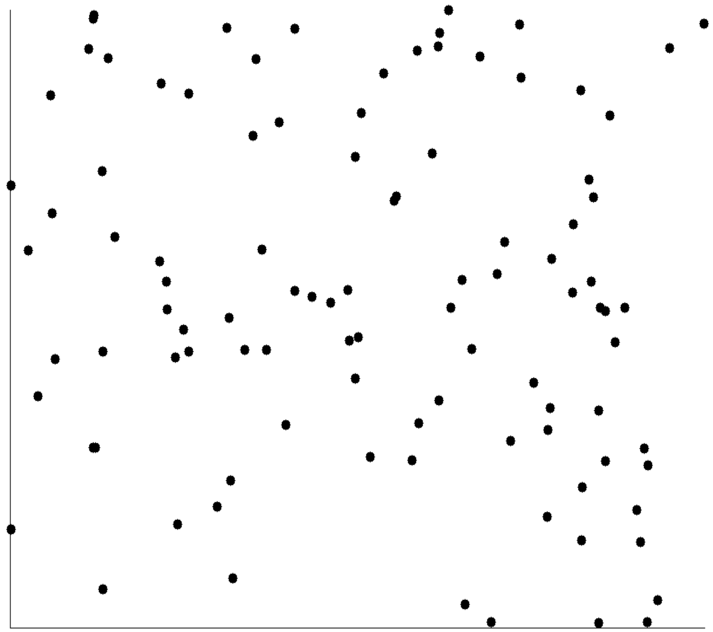


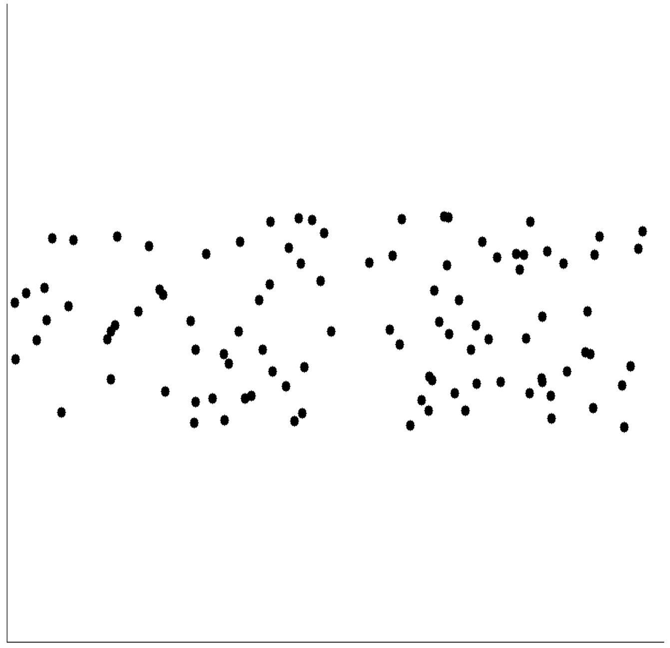


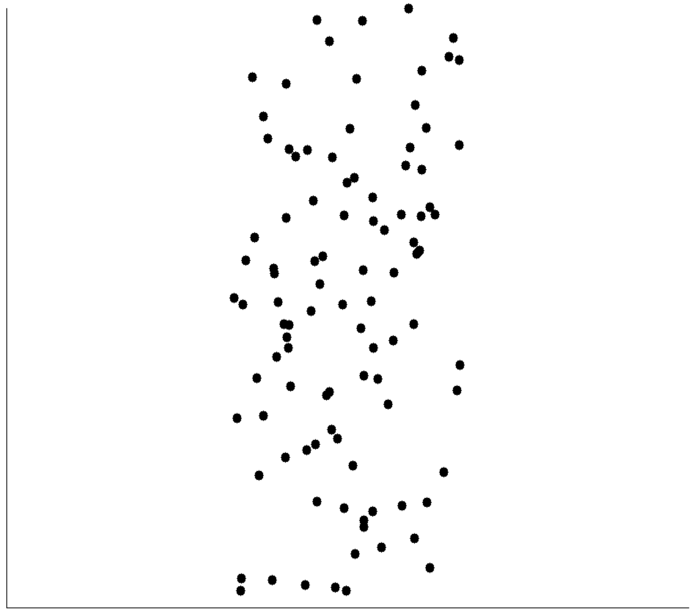


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data

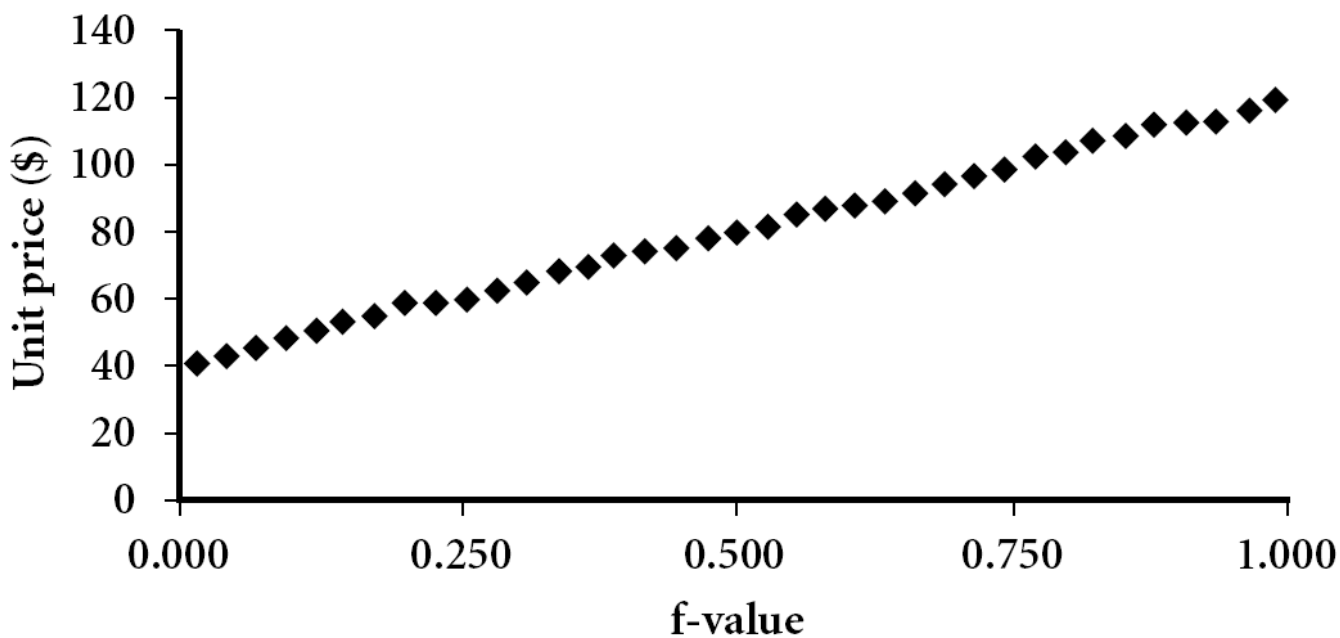






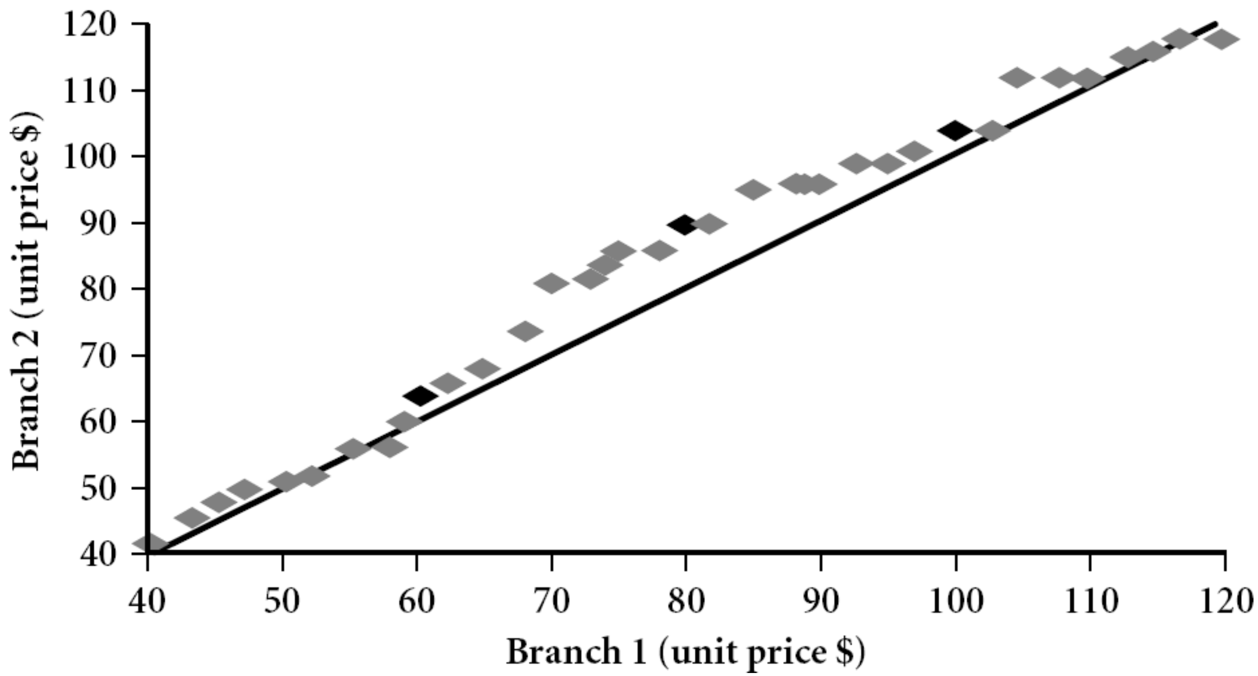
Quantile Plot

- 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.



Chapter 2: Getting to Know Your Data

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

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 data points with p dimensions
  - n-by-p matrix
  - Two modes

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

- Dissimilarity matrix: object-by-object structure
  - n data points, but registers only the distance
  - n-by-n matrix
  - A triangular matrix (symmetry)
  - Single mode

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

Data Matrix and Dissimilarity Matrix

- The dissimilarity matrix:
  - $d(i,j)$ is the difference between objects (data points) i and j.
  - Usually $\mathrm{d}(i,j) \geq 0$
  - $d(i,j) = d(j,i)$
  - Measures of similarity are usually expressed in terms of measures of dissimilarity, for example,

$$sim(i,j) = 1 - d(i,j)$$

assuming that any dissimilarity value is at most 1

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., the attribute map_color may have 5 states: *red*, *yellow*, *pink*, *blue*, *green* (generalization of a binary attribute)
- Method 1: Simple matching
  - *m* : # of matches, *p* : total # of attributes

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

  - weights can be assigned to increase the effect of matches, or to assign greater weight to matches in attributes with a large number of values.

Proximity Measure for Nominal Attributes

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

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

Let's compute the dissimilarity matrix with respect to the nominal attribute test-1 (p = 1)

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

Proximity Measure for Nominal Attributes

- If similarity is required instead, it can be computed by:

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

- Method 2: given a nominal attribute with M states (values), for each state introduce a binary attribute, then use measures that apply to binary attributes, e.g., map_color can be replaced by 5 binary attributes *isRed*, *isYellow*, etc. One attribute will be 1, while the rest are set to 0's.

Proximity Measure for Binary Attributes

- A contingency table for binary data

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

  - q is the number of binary attributes that are 1 for both objects i and j
  - r is the number of attributes that are 1 for object i and 0 for object j.
  - …
- Distance measure for symmetric binary variables:

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

- Distance measure for asymmetric binary variables:
  - 1 is more significant than 0, e.g., 1 indicates the test for cancer turned to positive.
  - In the formula we consider matches and mismatched on the value 1, and discards matches on zeros (i.e. discard t)

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

The total number of attributes is $p = q + r + s + t$

Proximity Measure for Binary Attributes

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

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

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

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

Note: $sim_{Jaccard}(i,j) = 1 - d(i,j) = 1 - \frac{r+s}{q+r+s}$

Dissimilarity between Binary Variables

- Example

Table 2.4 Relational Table Where Patients Are Described by Binary Attributes

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

  - Gender is a symmetric attribute: male/female
  - The remaining attributes are asymmetric binary
  - For fever & cough, set Y=1 & N=0: they indicate Yes and No,
  - For the rest (the 4 tests), set P=1 and N=0: they indicate positive and negative.

Dissimilarity between Binary Variables

- Example

Table 2.4 Relational Table Where Patients Are Described by Binary Attributes

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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

0 (#atts: 1 for Jack & 0 for Mary) +1 (0 for Jack & 1 for Mary)

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

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

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

Note: gender was ignored in the computations

Jim and Mary are unlikely to have a similar disease while Jack and Mary are the most likely to have a similar disease.

Distance on Numeric Data

- In some cases, we normalize the data before computing the dissimilarity of the objects on their numeric attributes, Why?

  Measuring an attribute in smaller units will expand its range of values, thus giving it a higher weight in distance computation.

  Distance: in miles vs in inches

  Attributes: Salary (5 or 6 figure) vs height (in feet)

  Using normalization: each attribute assumes values in [-1,1] or in [0.0,1.0]. (same weight)

- 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

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 $p-\mathrm{dimesnional}$ data objects, and h is the order.

(the distance defined above is also called L_h norm)

Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm) distance

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

  - E.g., the Hamming distance: the number of bits (symbols) that are different between two binary vectors
- $h = 2$: (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.
  - This is the maximum difference between any component (attribute) of the vectors

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

Special Cases of Minkowski Distance

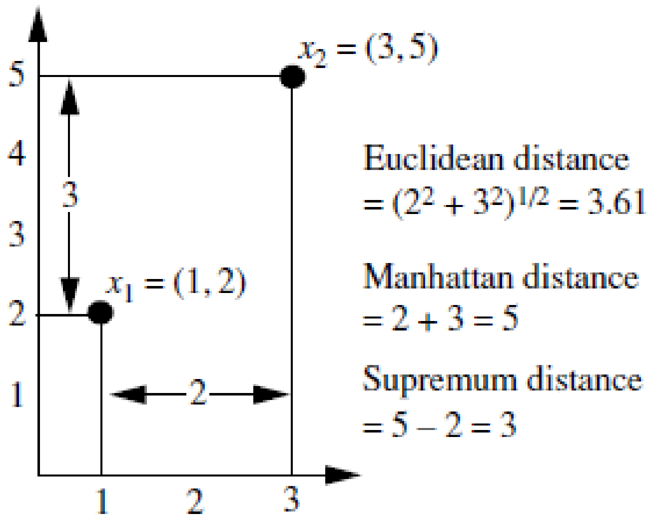
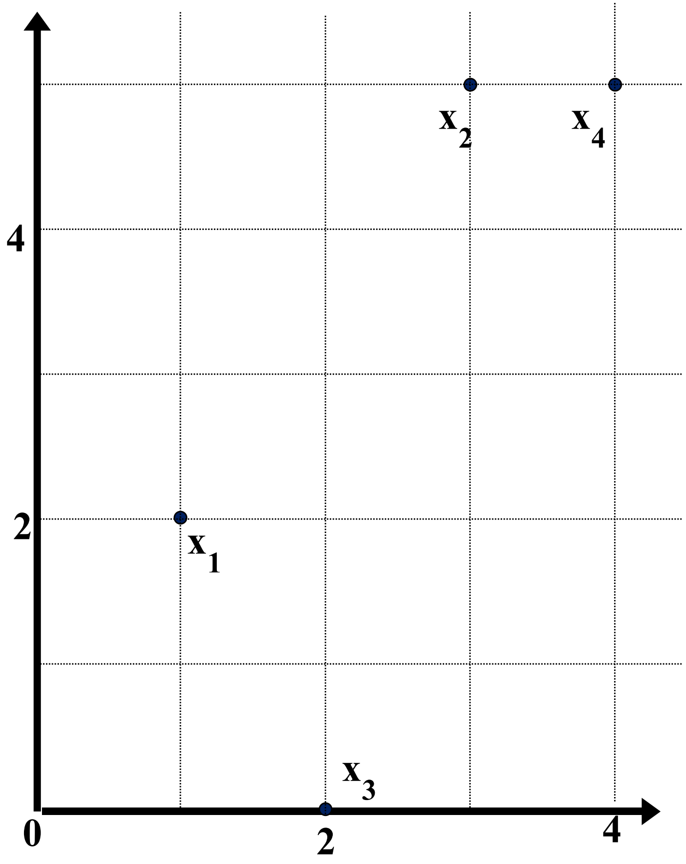


Figure 2.23 Euclidean, Manhattan, and supremum distances between two objects.

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

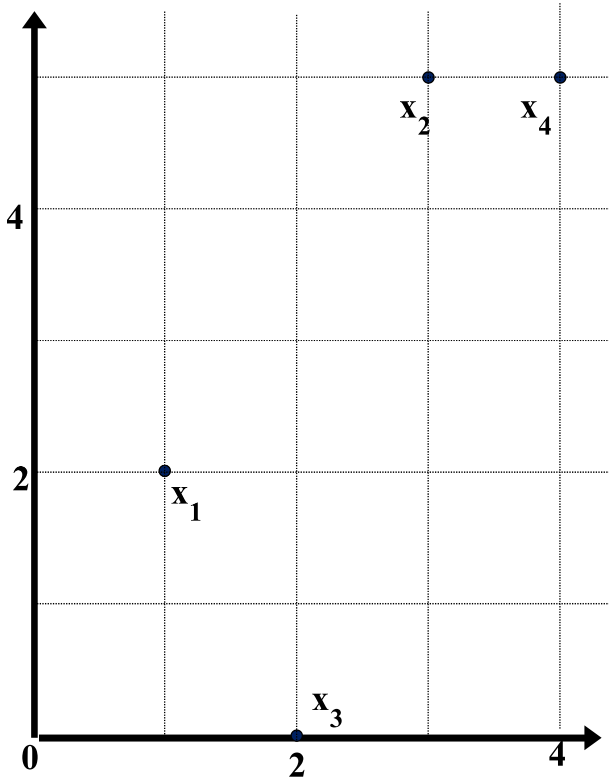
(with Euclidean Distance)

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

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

Ordinal Variables

- An ordinal variable have a meaningful order, yet, the magnitude between two successive values is unknown, e.g., the difference between small & large drink.
- An ordinal attributes can be obtained from the discretization of a continues numeric attribute.

Example:

The interval-scaled attribute *temperature* (in Celsius) can be organized into the following states:

Category	**Range**
cold temperature	-30 to -10
moderate temperature	-10 to 10
warm temperature	10 to 30

Ordinal Variables

- How to handle ordinal attributes in dissimilarity computation? By ranking
- Let M be the number of possible states that an ordinal attribute f can have. The states define a ranking over f : $1, 2, \ldots, M_f$

Example:

Category	Rank: $M_f = 3$
cold temperature	1
moderate temperature	2
warm temperature	3

- After ranking, ordinal attributes can be treated as numeric (Euclidean, etc.)

Ordinal Variables

- Suppose that f is an attribute among other ordinal attributes describing n objects. Steps:

1. Replace each x_{if} by its corresponding rank,

   $$r_{if} \in \{1,2,\ldots,M_f\}$$

2. Since ordinal attributes can differ in the number of states; normalization is necessary: map each range of ranks onto $[0.0,1.0]$. Replace each r_{if} by z_{if} such that

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

3. Dissimilarity can then be computed using any of the distance measures for numeric attributes, using z_{if} to represent the f value for the i^{th} object.

Ordinal Variables

Table 2.2 A Sample Data Table Containing Attributes of Mixed Type

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

Let's compute the dissimilarity matrix with respect to the ordinal attribute test-2:

- 3 states/ranks: fair = 1, good = 2, excellent = 3 (so M = 3 for test-2)
- Normalized (not necessary since we have only one ordinal):

  fair: $z = \frac{1-1}{3-1} = 0.0$, good: $z = \frac{2-1}{3-1} = 0.5$, excellent: $z = \frac{3-1}{3-1} = 1.0$

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}.$$

Euclidean

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- Method 1:
  - Group the attributes by type (nominal, ordinal, numeric, etc.)
  - Perform separate analysis on each group (distance, clustering, etc.)
  - This is feasible if each separate analysis per attribute type generates compatible results, otherwise, use method 2 we process all the attributes together.

Attributes of Mixed Type

- Method 2:

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

  - The indicator $\delta_{ij}^{(f)}=0$, if either
    - x_{if} or x_{jf} is missing, or
    - $x_{if}=x_{if}=0$ and attribute f is asymmetric binary

    otherwise $\delta_{ij}^{(f)}=1$
  - If f is numeric: $d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{max_f-min_f}$; the difference normalized by the range of attribute f
  - If f is nominal or binary: $d_{ij}^{(f)}=0$ if $x_{if}=x_{jf}$, 1 otherwise.
  - If f is ordinal: compute the ranks r_{if} and z_{if} and treat z_{if} as numeric.

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 (Euclidean norm)

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
  The cosine value is approximately 1 when the angel between the vectors is very small (hence considered similar)
- 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)

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

Chapter 2: Getting to Know Your Data

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

Summary

- 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
  - Data visualization: map data onto graphical primitives
  - 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