Data Preprocessing

Young-Rae Cho, Ph.D.
Associate Professor
Division of Software / Division of Digital Healthcare
Yonsei University – Mirae Campus

Why Need Data Preprocessing?

- **Incomplete Data**
  - Missing values, or Lack of attributes of interest
- **Noisy Data**
  - Errors, or Outliers
- **Redundant Data**
  - Duplicate data, or Duplicate attributes

    e.g., Age = “47”, Birthday = “01/07/1968”
- **Inconsistent Data**
  - Containing discrepancies in format or name

    e.g., Rating by “1, 2, 3”, Rating by “A, B, C”
- **Huge Volume of Data**

Importance of Data Preprocessing

- **To Increase Data Quality**
  - Mining quality depends on data quality as well as mining techniques.
    (Lower Quality Data, Lower Quality Mining Results !!)

- **Majority of Data Mining**
  - Data pre-processing comprises the majority of the works for data warehousing and data mining.

Major Tasks of Data Preprocessing

- **Data Cleaning**
  - Fill in missing values, smooth noisy data, remove outliers, remove redundancy, and resolve inconsistency

- **Data Integration**
  - Integration of multiple databases or files

- **Data Transformation**
  - Normalization and aggregation

- **Data Reduction**
  - Reducing representation in volume with similar analytical results
  - Discretization of continuous data

Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. Data Integration
5. Data Transformation
6. Data Reduction

Data Type

- **Record**
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data, e.g., text documents
  - Transaction data
- **Ordered Data**
  - Sequential data, e.g., transaction sequences, biological sequences
  - Temporal data, e.g., time-series data
  - Spatial data, e.g., maps
- **Graph**
  - WWW, internet
  - Social or information networks
  - Biological networks

Attribute Type

- **Nominal**
  - e.g., ID number, profession, zip code
- **Ordinal**
  - e.g., ranking, grades, sizes
- **Binary**
  - e.g., medical test (positive or negative)
- **Interval**
  - e.g., calendar dates, temperature, height
- **Ratio**
  - e.g., population, sales

Discrete Attribute vs. Continuous Attribute

Discrete Attribute

- Finite set of values
- Sometimes, represented as integer values
- Binary attributes are a special case of discrete attributes

Continuous Attribute

- Real numbers as values
- Typically, represented as floating-point variables
- In practice, shown as a finite number of digits

Data Characteristics

- **Dimensionality**
  - Curse of dimensionality
- **Sparsity**
  - Lack of information
- **Resolution**
  - Patterns depending on the scale
- **Similarity**
  - Similarity measures for complex types of data

Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. Data Integration
5. Data Transformation
6. Data Reduction

Descriptive Data Mining

❑ Motivation

- To better understand the properties of data distributions,
  e.g., central tendency, spread and variation

❑ Measurements

- median, max, min, quantiles, outliers, etc.

❑ Analysis Process

- Folding the measures into numeric dimensions
- Graphic analysis on the transformed dimension space

Central Tendency Measures

❑ Mean

- 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

- Middle value if odd number of values
- Average of two middle values otherwise
- Estimation by interpolation for grouped data:

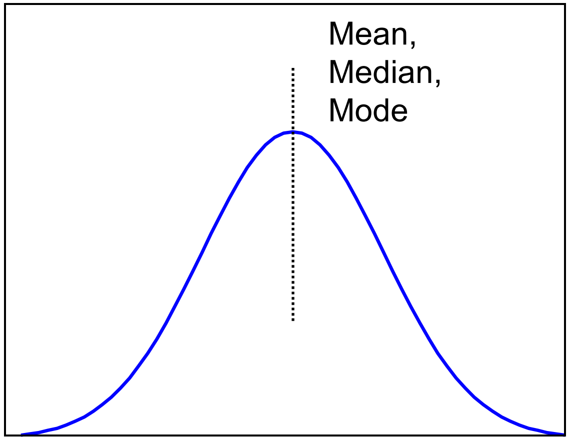
$$median = L_1 + \left(\frac{N/2 - (\sum freq_{low})}{freq_{median}}\right) width$$

❑ Mode

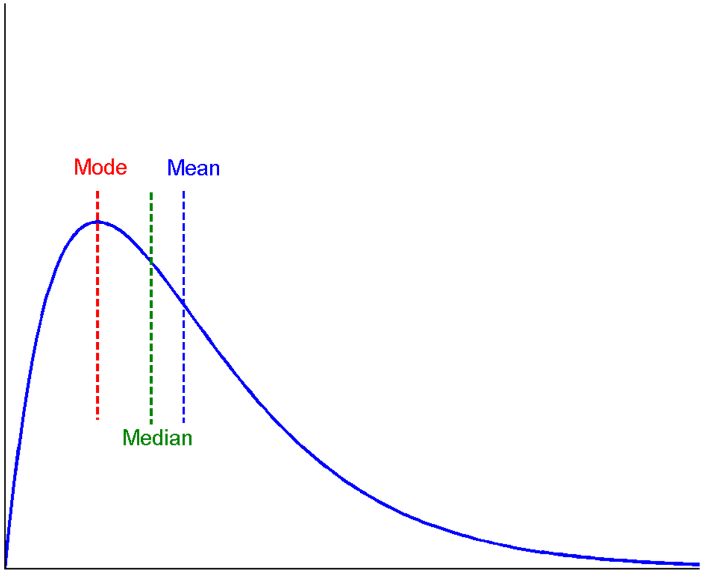
- The value that occurs the most frequently in the data
- Unimodal, bimodal, trimodal distribution

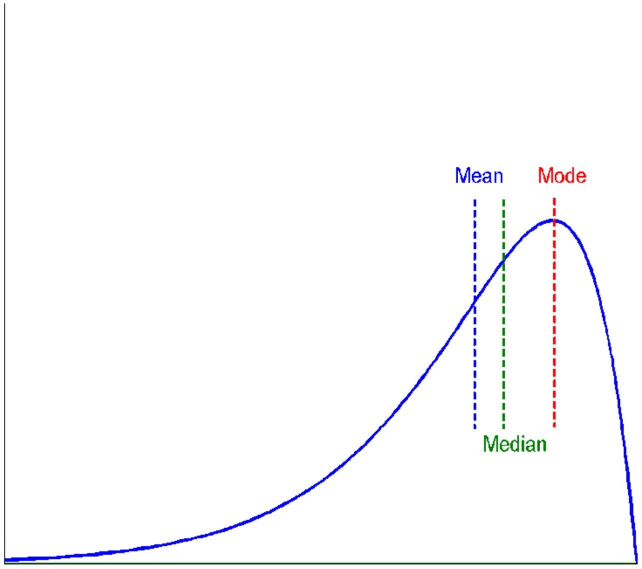
Central Tendency in Skewed Data

- ❑ **Symmetric Data**



- ❑ **Skewed Data**





Data Dispersion Measures

- **Quartiles and Outliers**
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: IQR = Q3 – Q1
  - Outliers: data with extreme low and high values

    usually, values lower/higher than Q1–1.5×IQR / Q3+1.5×IQR

- **Variance and Standard Deviation**
  - σ^2, σ in population:

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

  - s^2, s by sampling:

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

Degree of Freedom: # independent pieces of information

(= # independent measurement - # parameters)

Graphic Analysis

- **Boxplot**
  - Display of five-number summary

- **Histogram**
  - Display of tabulated frequencies

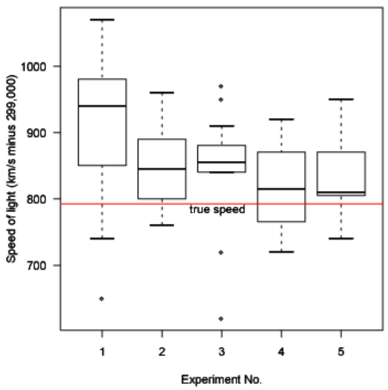
- **Quantile-Quantile (Q-Q) Plot**
  - Description of the relationship between two univariate distributions

- **Scatter Plot**
  - Description of the relationship between two attributes of a bivariate distribution

Boxplot Analysis

- **Five-number summary of a Distribution**
  - Minimum / Q1 / Median / Q3 / Maximum

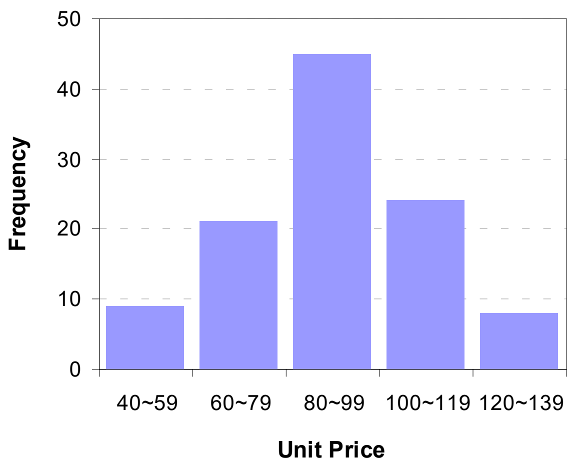
- **Boxplot**
  - Represented as a box
  - The bottom of the box is Q1
  - The top of the box is Q3
  - The median is marked by a line
  - Whiskers: two lines outside of the box
    extend to minimum and maximum

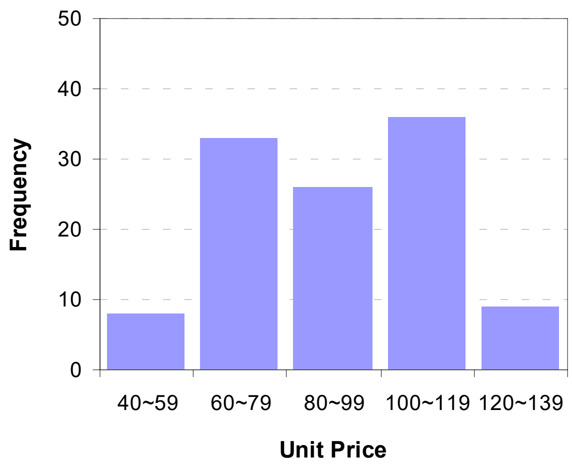


Histogram Analysis

- **Histogram**
  - Univariate graphic method
  - Represented as a set of bars reflecting the frequencies of the discrete values
  - Grouping data values into classes if they are continuous

- **Boxplot vs. Histogram**
  - Often, histogram gives more information than boxplot





Quantile Plot Analysis

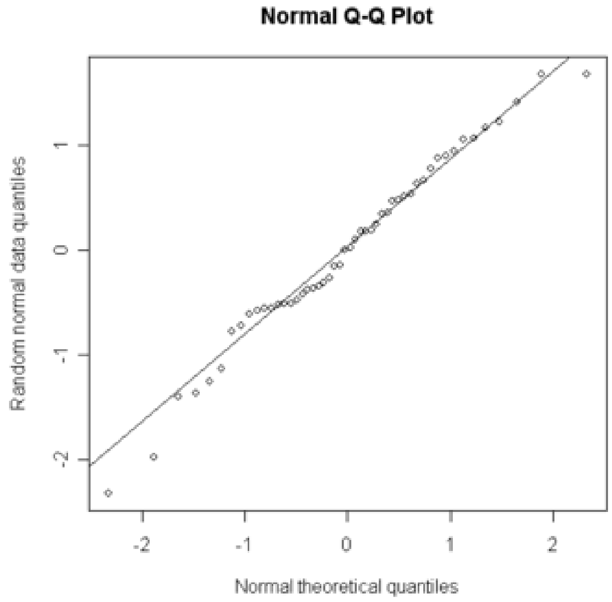
❑ Quantile Plot

- Plots quantile information of the data (sorted in an ascending order)
- Displays all the data

❑ Q-Q (Quantile-Quantile) Plot

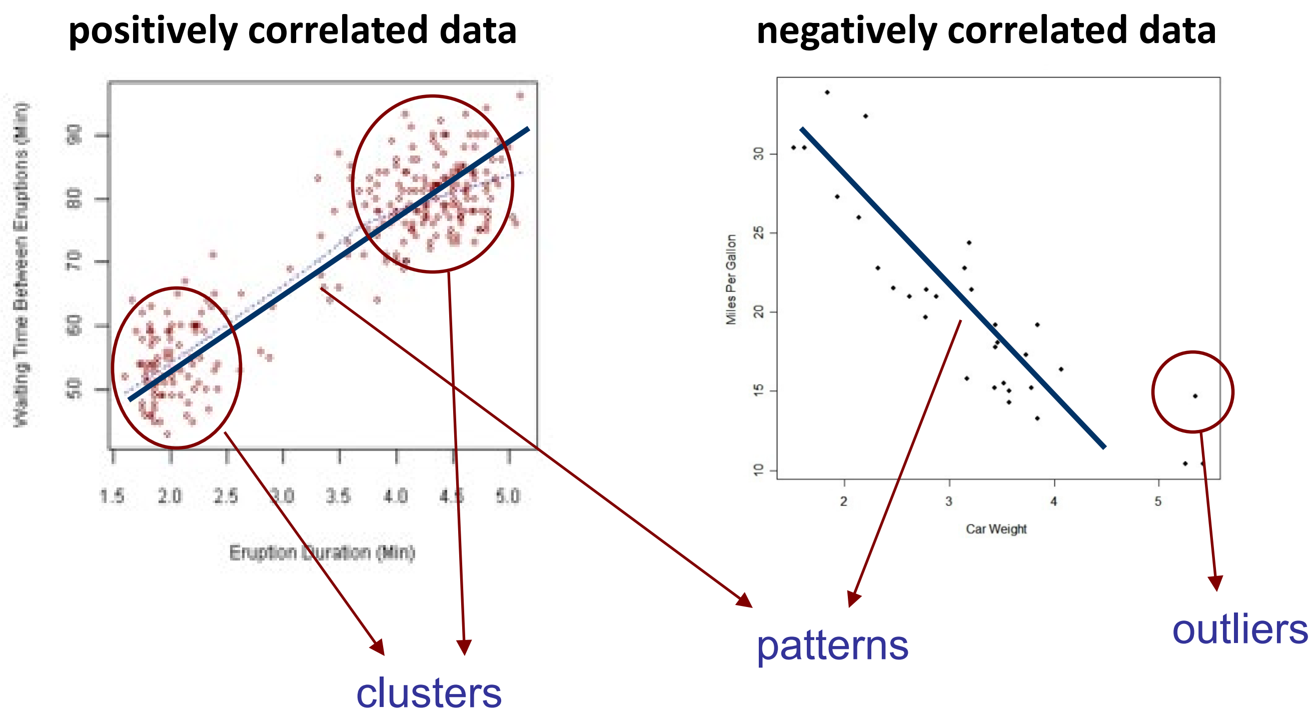
- Plots the quantiles of one univariate distribution against the quantiles of the other
- Describes the relationship

  between two distributions



Scatter Plot Analysis

- **Scatter Plot**
  - Displays the points of bivariate data
  - Describes the relationship between two attributes (variables)



Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. Data Integration
5. Data Transformation
6. Data Reduction

Missing Data

- **Data is not always available**
  - e.g., many tuples have no record value for several attributes

- **Missing data may be due to**
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

- **Missing data may need to be inferred**

How to Handle Missing Data

- **Ignore the missing values**
  - Not effective

- **Fill in the missing values manually**
  - Tedious, infeasible?

- **Fill in the missing values automatically with**
  - "unknown": not effective
  - The attribute mean
  - The attribute mean of all samples belonging to the same class
  - The most probable value by inference or classification techniques

Noisy Data

- **Noise**
  - Random error or variance in a measured variable
- **Incorrect data may be due to**
  - faulty data collection instruments
  - data transmission problem
  - technology limitation
  - inconsistency in data conversion
- **Other Data Problems**
  - Duplicate records
  - Incomplete data
  - Inconsistent data

How to Handle Noisy Data

- **Binning**
  - Sort data and partition into bins
  - Smooth by bin means, smooth by bin median, smooth by bin boundaries

- **Regression**
  - Smooth by fitting the data into regression functions

- **Clustering**
  - Detect and remove outliers

- **Inspection Semi-automatically**
  - Detect suspicious values and check by human

Partitioning for Binning

- **Equal-Width (Distance) Partitioning**
  - Divides the range into N intervals of equal distance (uniform grid)
  - If A and B are the lowest and highest values of the attribute,
    then the width of intervals will be (B-A)/N.
  - Straightforward
  - Problem:
    1. Outliers may dominate the partitions.
    2. Skewed data is not handled well.

- **Equal-Depth (Frequency) Partitioning**
  - Divides the range into N intervals of equal frequency, i.e., each containing approximately same number of samples.
  - Problem: Not possible for categorical attributes

Data Smoothing for Binning

- **Example**
  - Sorted data of price (in dollars): 4,8,9,15,21,21,24,25,26,28,29,34
  - Partition into three equal-frequency bins

Bin 1: 4, 8, 9, 15
Bin 2: 21, 21, 24, 25
Bin 3: 26, 28, 29, 34

Bin 1: 9, 9, 9, 9
Bin 2: 23, 23, 23, 23
Bin 3: 29, 29, 29, 29

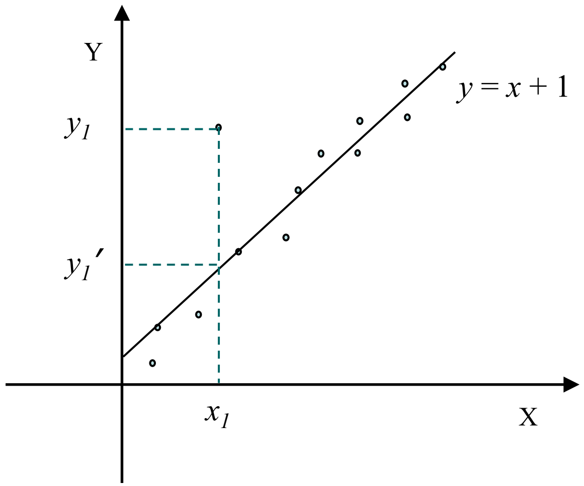
smoothing by bin means

Bin 1: 4, 4, 4, 15
Bin 2: 21, 21, 25, 25
Bin 3: 26, 26, 26, 34

smoothing by bin boundaries

Regression

- **Linear Regression**
  - Modeled as a linear function of one variable,
    $Y = w X + b$
  - Often, uses a least-square method.

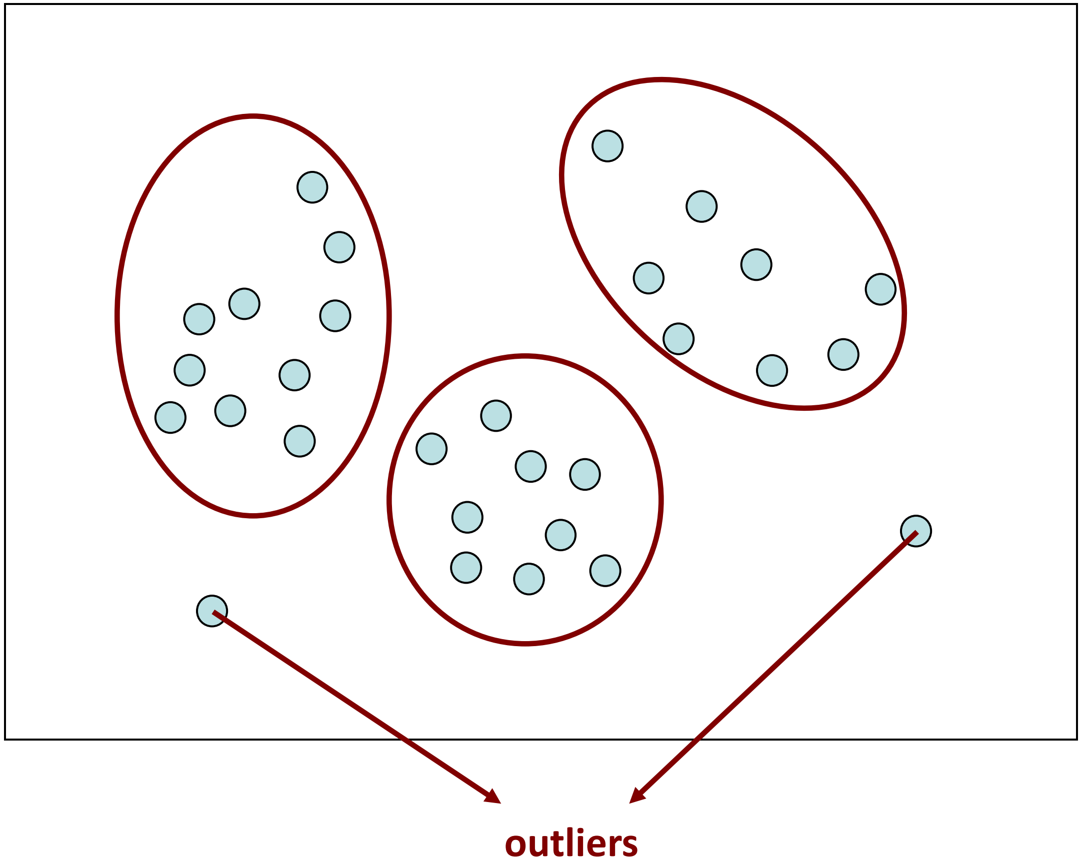


- **Multiple Regression**
  - Modeled as a linear function of a multi-dimensional feature vector, $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many non-linear functions can be transformed.

- **Log-Linear Model**
  - Approximates discrete multi-dimensional probability distributions.

Clustering

❑ Outlier Detection



Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. <u>Data Integration</u>
5. Data Transformation
6. Data Reduction

Data Integration

Definition

- Process to combine multiple data sources into coherent storage
- Process to provide uniform interface to multiple data sources

Process

- Data Modeling → Schema Matching → Data Extraction

Data Modeling

- Creating global schema (mediated schema)

Schema Matching

- Matching between two attributes of different sources
- The most critical step of data integration
- Schema-level matching / Instance-level matching

Instance-Level Matching

❑ Definition

- Detecting and resolving data value conflicts

❑ Entity Identification

- For the same real world entity, values from different sources might be different
- Possible reasons:
  1. different representations, e.g., Greg Hamerly = Gregory Hamerly
  2. different format, e.g., Sep 16, 2009 = 09/16/09
  3. different scale, e.g., meters ↔ inches

Schema-Level Matching

Definition

- Detecting and resolving attribute conflicts and redundant attributes

Object Identification

- The same attribute (or object) might have different names in different sources.

  e.g., transaction id = TID
- One attribute might be a "derived" attribute in another table.

  e.g., Age = Birthday

Attribute Redundancy Analysis

- Can be analyzed by correlation / variation measures

  e.g., χ^2 test, Pearson coefficient, *t*-test, *F*-test

Pearson Coefficient

- **Pearson Coefficient**
  - Evaluates correlation between two samples.
  - Given two samples X = {$x_1, x_2, \ldots, x_n$} and Y = {$y_1, y_2, \ldots, y_n$},

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x \sigma_y}$$

co-variance between *X* and *Y*

individual variance (standard deviation) of *X* and *Y*

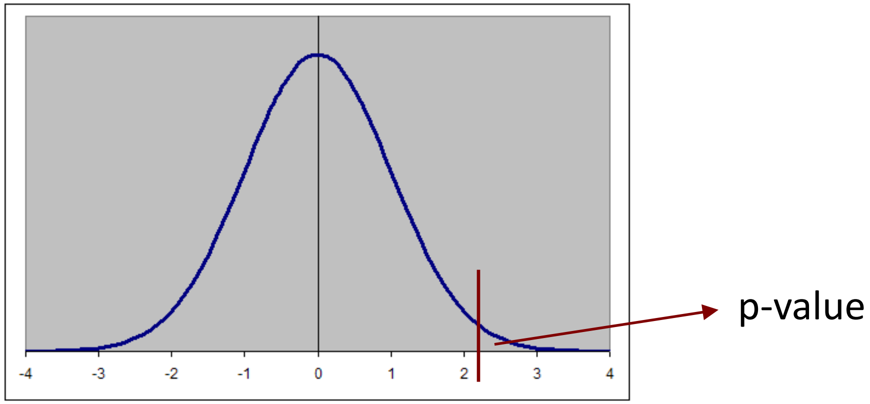
  - If r > 0, X and Y are positively correlated.
  - If r = 0, X and Y are independent.
  - If r < 0, X and Y are negatively correlated.

t-Test and *F*-Test

- ***t*-Test (*t*-statistics)**
  - Independent two-sample *t*-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1{}^2 / n_1 + s_2{}^2 / n_2}}$$

  - Evaluates statistical variance between two samples.



- **ANOVA (Analysis of Variance) / *F*-test (*F*-statistics)**
  - Evaluates statistical variance among three or more samples

Chi-Square Test

χ^2 Test (χ^2 Statistic)

- Evaluates whether an observed distribution in a sample differs from a theoretical distribution (i.e., hypothesis).
- $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ where E_i is an expected frequency and O_i is an observed frequency
- The larger χ^2, the more likely the variables are related (positively or negatively).

Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum (col.)	300	1200	1500

Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. Data Integration
5. <u>Data Transformation</u>
6. Data Reduction

Data Transformation

- **Definition**
  - Process that maps an entire set of values of a given attribute into a new set of values
- **Purpose**
  - To remove noise from data
  - To change scales
- **Methods**
  - Smoothing (including binning and regression)
  - Normalization

General Normalization Methods

- **Min-Max Normalization**
  - Maps the values in the range [min, max] into a new range [min’, max’]

$$\frac{v' - min'}{max' - min'} = \frac{v - min}{max - min}$$

- **z-score Normalization**
  - Transforms the values of an attribute A based on its mean and standard deviation

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- **Decimal Scaling**
  - Moves decimal point of values $v' = \frac{v}{10^j}$ where j is the maximal digit

Quantile Normalization (1)

Motivation

- In a Q-Q plot, if two distributions are the same, then the plot should be a straight line.
- Can be extended to n dimensions

Description

- $q_k = (q_{k1}, \ldots, q_{kn})$: a vector of the kth quantile for all n dimensions

$$proj_d q_k = \left(\frac{1}{n}\sum_{i=1}^{n} q_{ki}, \ldots, \frac{1}{n}\sum_{i=1}^{n} q_{ki}\right)$$

Algorithm

- Sort each column (dimension) of X to give X'
- Assign the means across rows of X' into each element of the row
- Rearrange each column of X' to the same order of X

Quantile Normalization (2)

- **Advantages**
  - Efficient in high dimensional data (popularly used for biological data pre-processing)

- **Disadvantages**
  - In practice, each dimension may have different distribution

- **References**
  - Bolstad, B.M., et al. , “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”, Bioinformatics, Vol.19 (2003)

Overview

1. General Data Characteristics
2. Descriptive Data Summarization
3. Data Cleaning
4. Data Integration
5. Data Transformation
6. **Data Reduction**

Data Reduction

- **Definition**
  - Process to obtain a reduced representation of a data set, which is much smaller in volume but produces almost the same analytical results

- **Problems**
  - Data mining algorithms take a very long time to run on the complete data sets
  - Data analysis methods are complex, inaccurate in the high dimensional data

- **Methods**
  - Dimensionality reduction
  - Numerosity reduction

Dimensionality Reduction

- **Curse of Dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Possible combinations of subspaces will grow exponentially
  - Density and similarity between data values becomes less meaningful

- **Purpose**
  - To avoid the curse of dimensionality
  - To eliminate irrelevant features and reduce noise
  - To reduce time and space required in data mining
  - To allow easier visualization

- **Methods**
  - Feature extraction
  - Feature selection

Feature Extraction

- **Process**
  1) Combining a multitude of correlated features
  2) Creating a new dimensional feature space for the combined features

- **Example**
  - Principal component analysis (PCA)
    - Find the eigenvectors of the covariance matrix
    - Define a new space with the eigenvectors
  - Wavelet transformation

- **Problem**
  - New dimensional spaces might not be meaningful in the domain of data sets

Feature Selection

- **Methods**
  - Eliminating redundant features or irrelevant features
  - Selecting significant (informative) features

- **Example**
  - Redundant features:

    e.g., purchase price of a product and the amount of sales tax paid
  - Irrelevant features

    e.g., parent's name is irrelevant for selecting student scholarship candidates
  - Informative features

    e.g., student's name, student's GPA, parent's income are informative for selecting student scholarship candidates

Heuristic Search for Feature Selection

- **Problem of Feature Selection**
  - If d features, how many possible combinations of the features?

    → 2^d

- **Typical Heuristic Methods**
  - Step-wise feature selection: Repeatedly pick the best feature
  - Step-wise feature elimination: Repeatedly remove the worst feature
  - Best combined feature selection and elimination
  - Optimal branch and bound

Numerosity Reduction

- **Purpose**
  - To reduce data volume by choosing alternative, smaller forms of data representation

- **Parametric Methods**
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data.
  - e.g., Regression

- **Non-parametric Methods**
  - Do not assume models, and use data values.
  - e.g., Discretization, Clustering, Conceptual Hierarchy Generation

Discretization

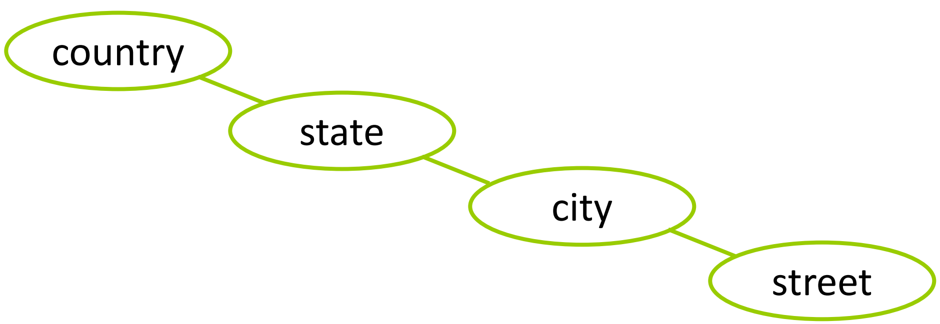
- **Methods**
  - Dividing the range of continuous data into intervals
  - Selecting significant (frequent) data
- **Strategy**
  - Supervised vs. Unsupervised
  - Splitting (top-down) vs. Merging (bottom-up)
- **Examples**
  - Binning: top-down, unsupervised
  - Sampling: top-down, supervised
  - Entropy-based Discretization: top-down, supervised

Conceptual Hierarchy Generation

- **Ordering Attributes**
  - Partial/total ordering of attributes at the schema level
  - e.g., street < city < state < country

- **Hierarchy Generation**
  - A hierarchy for a set of values by explicit data grouping
  - e.g., {Dallas, Waco, Austin} < Texas

- **Automatic Method**
  - Based on the number of distinct values per attribute



15 distinct values

365 distinct values

3567 distinct values

674,339 distinct values

Questions?

- Lecture Slides on the Course Website, “https://ads.yonsei.ac.kr/faculty/data_mining”