Chapter 2: Getting to Know Your Data

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

Contents

Types of Data Sets

- **Tabular**
  - Data matrix / table
    - E.g.> a set of term-frequency vectors
  - Transaction data
- **Graph and network**
  - Social networks
  - World Wide Web
  - Molecular structures
- **Time-series (ordered)**
  - Video data: sequence of images
  - Temporal data: time-series of trajectories
  - Sequential data: transaction sequences
  - Genetic sequence data
- **Spatial, image, and multimodal:**
  - Spatial data: maps
  - Image data
  - Multimodal data (video + image + text + ….)

	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

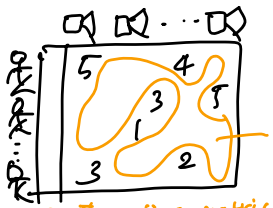
Data matrix

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

Transaction data

Characteristics of Data

- **Dimensionality**
  - # of features
  - Curse of dimensionality : 차원이 매우 클 경우 발생하는 문제 → difficult to measure the meaningful distance between data objects
- **Sparsity** density of the data is very small
  - Only a small portion of presence ex) rate matrix



→ The rating matrix is very sparse
And the data also becomes very difficult to analyze appropriatly
→ This is called data sparsity problem

- **Resolution**
  - Patterns depend on the scale
- **Distribution**
  - Centrality and dispersion

	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

Data Objects

- **Data sets are made up of data objects**
  - A **data object** represents a real-world entity
- **Examples:**
  - Sales database: customers, store items, sales
  - Medical database: patients, treatments
  - University database: students, professors, courses
- **Also called tuples, samples, examples, instances, data points, etc**
- **Data objects are typically described by features**
  - Database rows -> data objects; columns ->features

Features

- **Features (or dimensions, attributes, variables, etc):**
  - A measurable property or characteristics of each data object
  - *E.g., customer_ID, name, address, age, occupation, etc*
- **Types:**
  - Nominal
  - Binary
  - Numeric: quantitative
    - Ratio-scaled
    - Interval-scaled

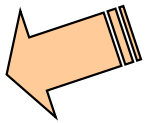
Feature Types

- **Nominal: categories, states, or "names of things"**
  - Has a finite number of values
  - e.g., Hair_color = {*black, blond, brown, grey, red, white, …*}
  - marital status, occupation, ID numbers, zip codes, ….
- **Binary**
  - Special case of a nominal feature with only 2 states (0 and 1)
  - Symmetric case and asymmetric case (will be mentioned later)
- **Ordinal**
  - Values have a meaningful order (ranking)
  - Magnitude between successive values is not known though
  - *E.g.> Size* = {*small, medium, large*}

Numeric Feature Types

- **Numeric (integer or real-valued)**
  - **Ratio-scaled**
    - Ratio is meaningful
    - Inherent **zero-point (0 means absence)**
    - We can speak of values as being an order of magnitude larger than the unit of measurement
      - 6kg is twice as high as 3kg
      - e.g., *temperature in Kelvin, length, counts, money, etc…*
  - **Interval-scaled**
    - Only difference is meaningful
    - Measured on a scale of **equal-sized units**
    - Values have order
      - e.g., *temperature in C°or F°*
    - No true zero-point

Contents



Basic Statistical Descriptions of Data

- **Motivation**
  - To better understand the data: central tendency, variation and spread
- **Data dispersion characteristics**
  - Median, max, min, quartiles, outliers, variance, etc.

Measuring the Central Tendency

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

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

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

- Weighted arithmetic mean:

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

- Trimmed mean:
  - Taking mean after chopping extreme values

Measuring the Central Tendency

Median:

- Middle value if there are odd number of values, or average of the middle two values otherwise
- Simple median requires sorting, not good at a dynamic situation
- Solution: estimation via **interpolation** (for grouped data):

approximation method

⇒ 새 데이터 추가/삭제 될때마다 Sorting 할 필요 없어짐

$$median = L_1 + (\frac{n/2 - (\sum freq_l)}{freq_{median}}) * width$$

L1: start point of the range

- Example
  - n = 3194, n/2 = 1597, $freq_{median}$ = 1500 , **L1 = 21**
  - Numerator = 1597 – (200+450+300) = 647
  - **width = (50-21) = 29**
  - Median = **21**+(647/1500)***29**

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

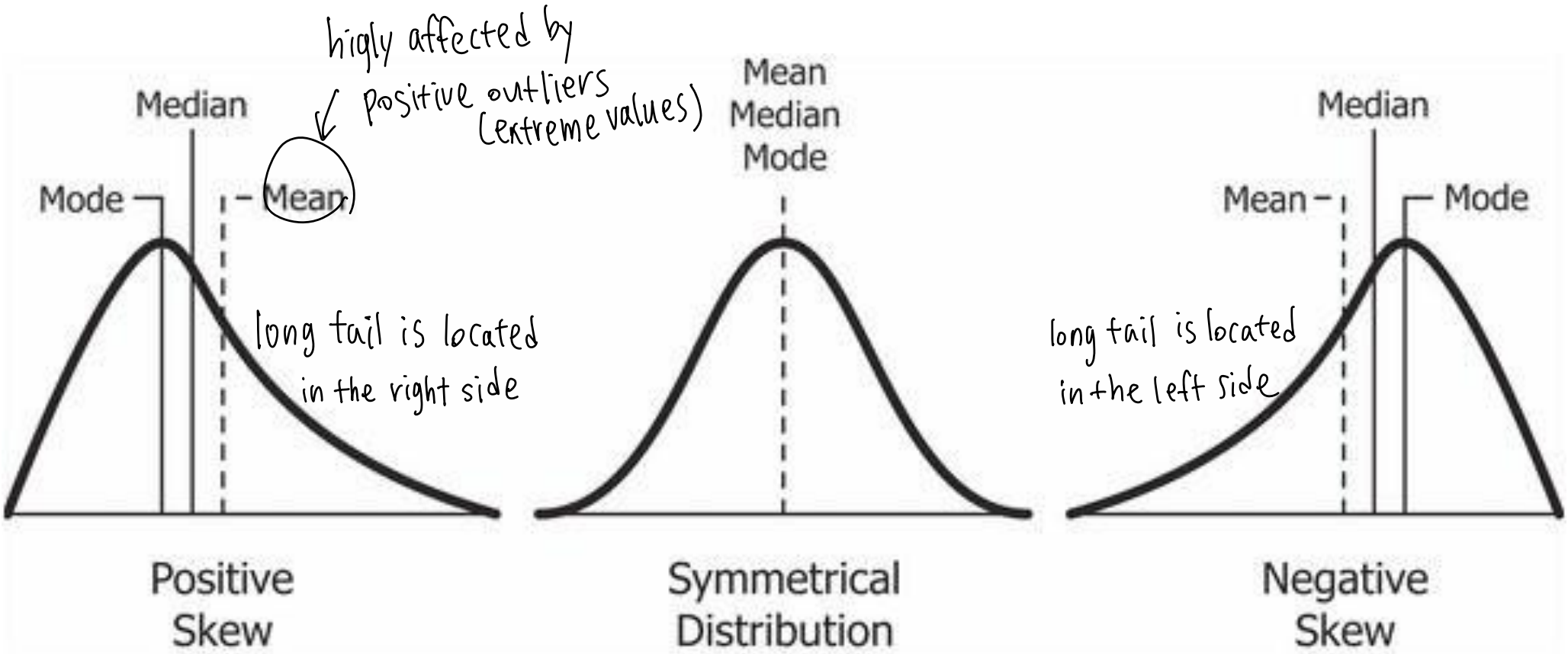
Median is located in this group: 21 ~ 50. Within this range, we approximate the exact position of median.

Measuring the Central Tendency

- **Mode**
  - Value that occurs most frequently in the data
    - Usually defined on discrete features
  - Unimodal (1), bimodal (2), trimodal (3)

Symmetric vs. Skewed Data

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



Measuring the Dispersion of Data

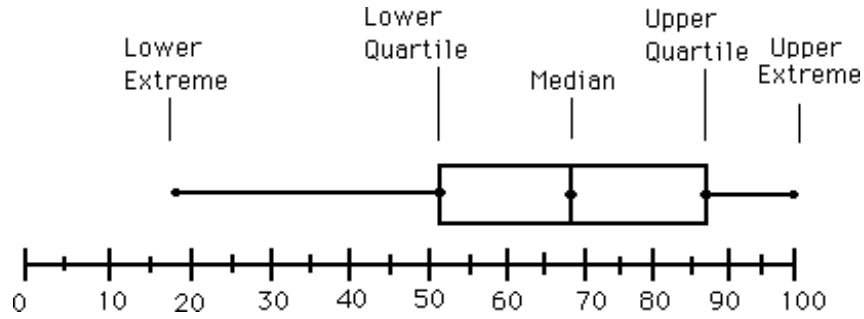
how the data points are distributed over the all range

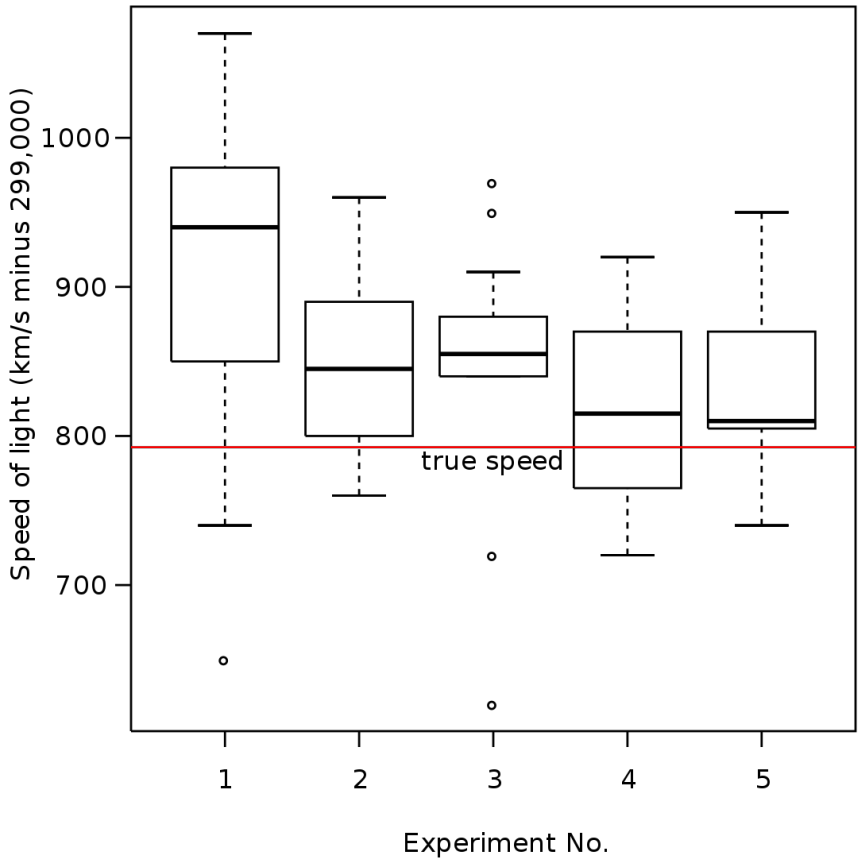
Quartiles, outliers and boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range (IQR)**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: visualization of the above five numbers
  - Each end of the box is Q_3 and Q_1 ; median is marked; add whiskers to express min & max; and plot outliers individually
- **Outlier**: usually, a value higher/lower than 1.5 x IQR

Measuring the Dispersion of Data

- **Box plot: 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 **Q1** and **Q3**.
  - => 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





Measuring the Dispersion of Data

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

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

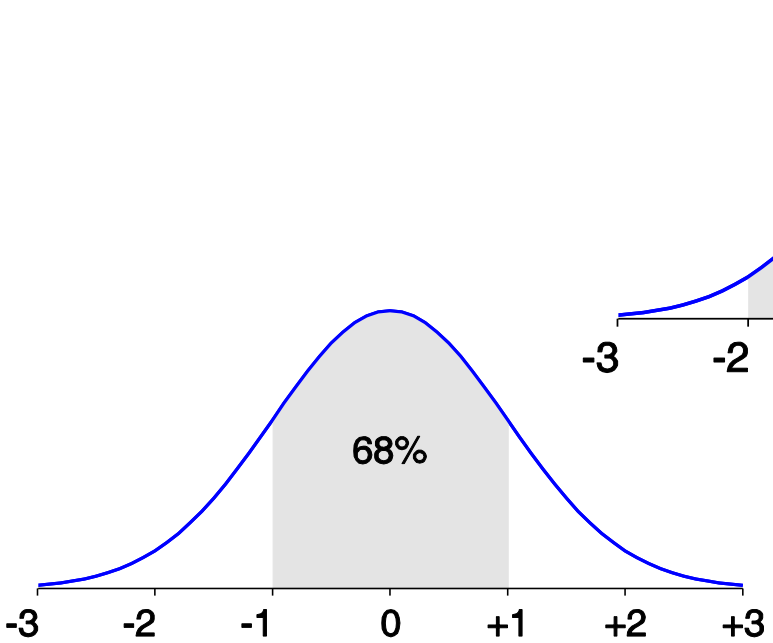
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

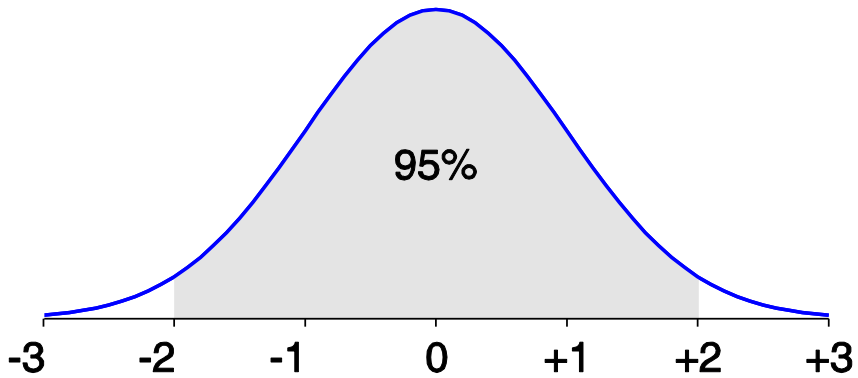
  - **Standard deviation** s (or, σ) is the square root of variance s^2 (or, σ^2)

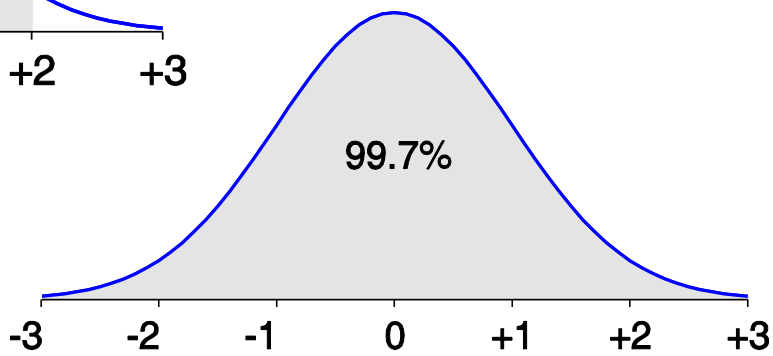
Measuring the Dispersion of Data

Using the normal distribution property

- 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







Thank You

Data
Intelligence
Lab