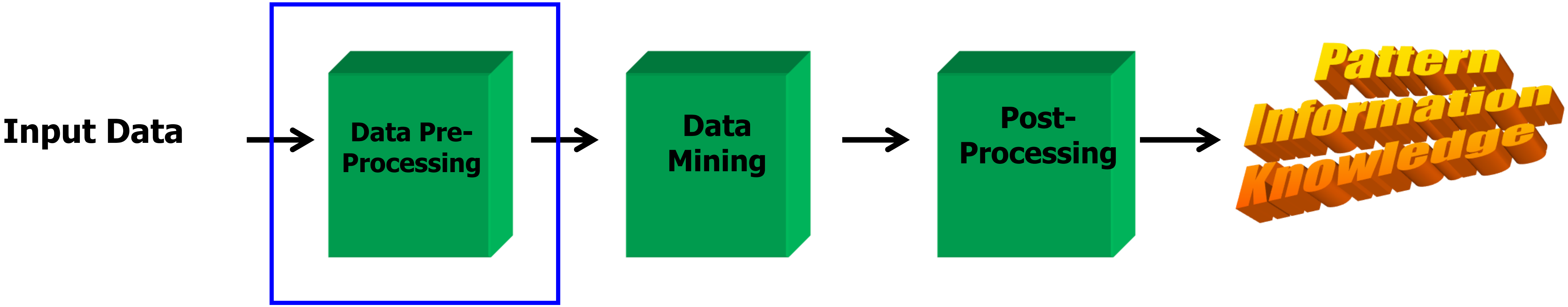
2-3 주차. 데이터 알아보기

과목명: 데이터사이언스

AI융합학부 박건우

Knowledge discovery process



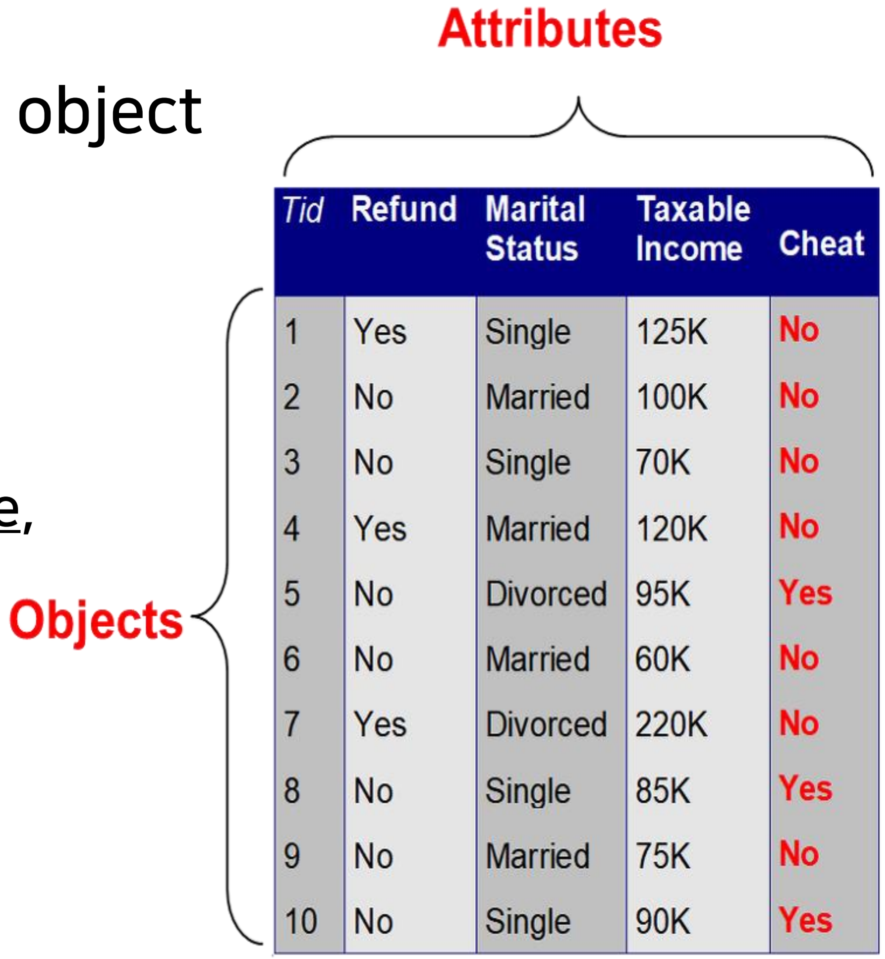
데이터를 이루는 구성요소들, 데이터 특징 파악을 위한 기초적인 분석 방법,
시각화 방법 등에 대해 다룹니다.

Contents

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

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
  - Example: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, or feature
- A collection of attributes describe an object
  - Data object is also known as record, data point, sample, entity, or instance
- Database rows -> data objects; columns -> attributes.


Attributes
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attributes (속성)

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
    - E.g., customer _ID, name, address
- Attribute types
    - Nominal
    - Binary
    - Ordinal
    - Numeric
        - Interval-scaled
        - Ratio-scaled

Attribute Types

- **Nominal (명목형)**: categories, states, or “names of things”
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- **Binary (이진)**
  - 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 the most important outcome (e.g., HIV positive)
- **Ordinal (순서형)**
  - If values have a meaningful order but magnitude between successive values is not known
  - Size = {small, medium, large}, grades, army rankings

Nominal, binary, and ordinal attributes indicate **quality** of a data object.

Numeric (수치형) Attribute

A numeric attribute indicates quantity (integer or real-valued)

- **Interval**
  - Measured on a scale of equal-sized units
  - Values have order. No true zero-point.
  - e.g., ℃, calendar dates
- **Ratio**
  - Inherent zero-point. e.g., temperature in Kelvin, length, counts
  - 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˚).

Discrete vs. Continuous Attributes

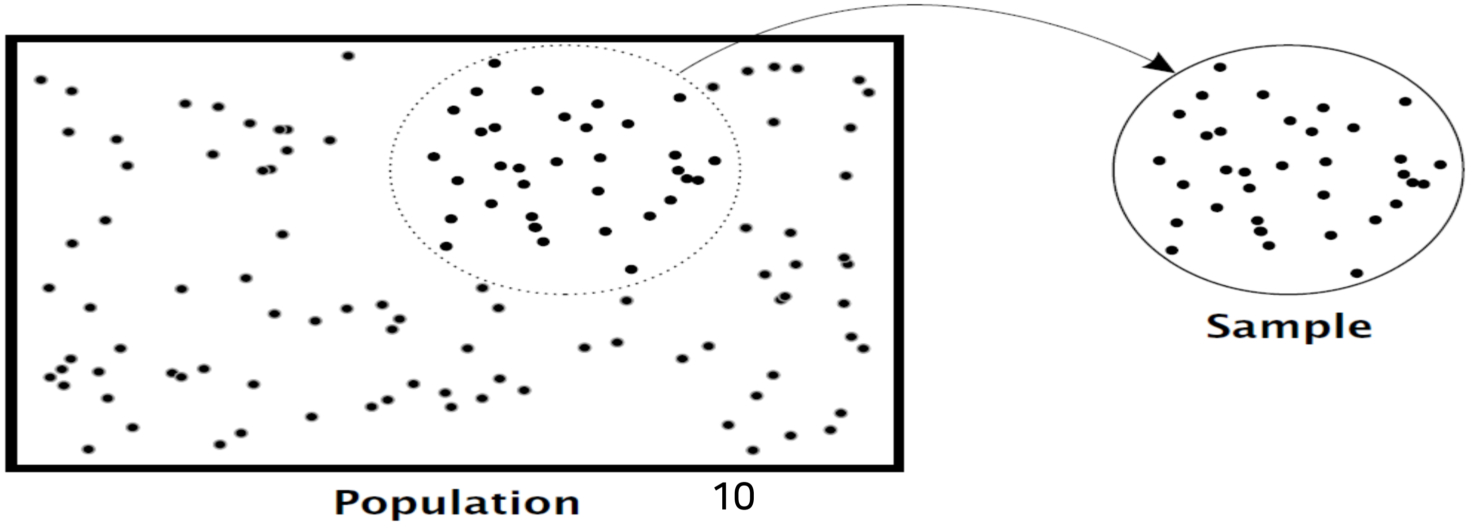
- **Discrete Attribute (이산, individually separate and distinct)**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute (연속형)**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables
  - The terms <u>numeric</u> attribute and <u>continuous</u> attribute are often used interchangeably

Contents

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

A quick recap: probability and statistics

- An **experiment** or **trial** (실험, 시행) is any procedure that can be infinitely repeated and has a well-defined set of possible outcomes, known as the sample space.
- A **random variable** (확률 변수) is a mapping from possible outcomes (e.g., {앞면, 뒷면}) to a measurable space (e.g., {0, 1}).
- An **observation** (관측) of a random variable is the value that is actually observed (what actually happened).
- The **population** (모집단) is the collection of all possible observations of a random variable.
- The **sample** (표본) contains the outcomes that are actually observed in an experiment.

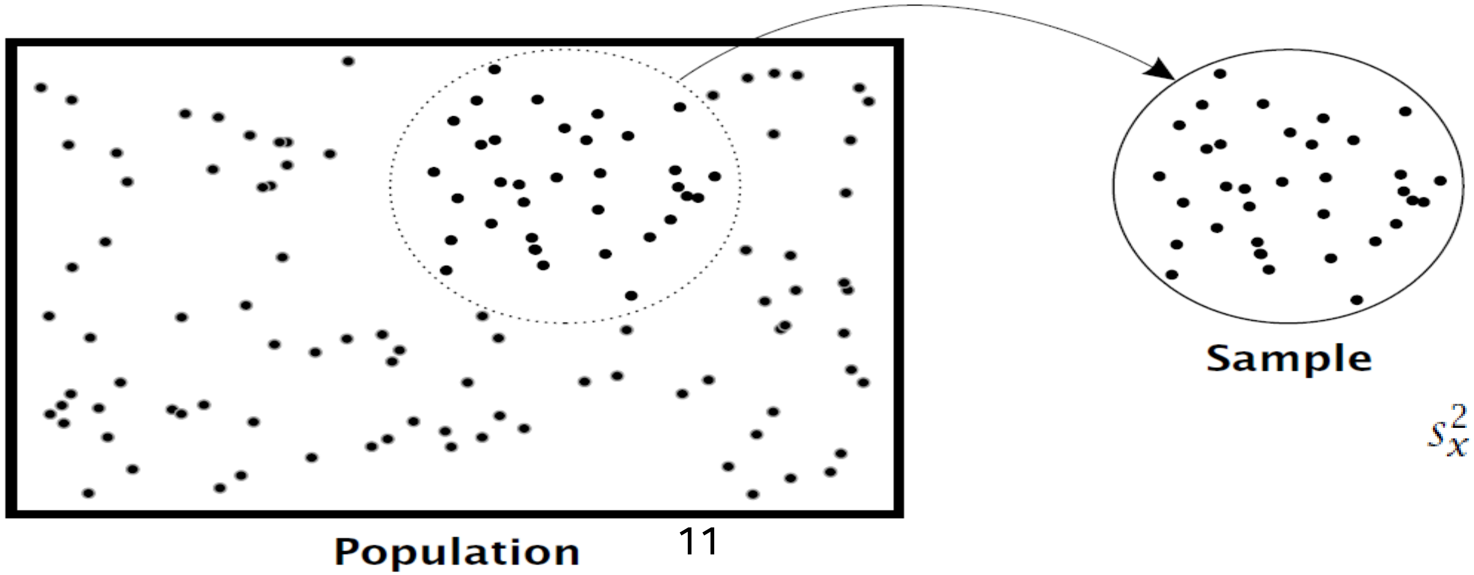


A quick recap: probability and statistics

- **Parameters (모수)**: values that describe the characteristics of the population, calculated from all the values in the population
  e.g., the mean and variance of the population
- In the most cases, we don't have an access to the entire population; Instead, we describe the characteristic of a sample by **sample statistics (표본 통계량)**

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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



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

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

A quick recap: probability and statistics

- Two purposes of sample statistics

  (1) *summarize* the characteristic of the sample

  (2) *estimate* population parameters

- In the usual KD process, you can use sample statistics for getting to know you data – **Descriptive Statistics** (기술 통계)

- According to the viewpoint of each analysis, you can use either population parameters or sample statistics.

Basic Statistical Descriptions of Data

- Motivation: To better understand your data!

- Measurement
  - Central Tendency (중심성): Mean, Median, and Mode
  - Dispersion of Data (산포도): Quartiles, Variance, Standard Deviation, Interquartile Range

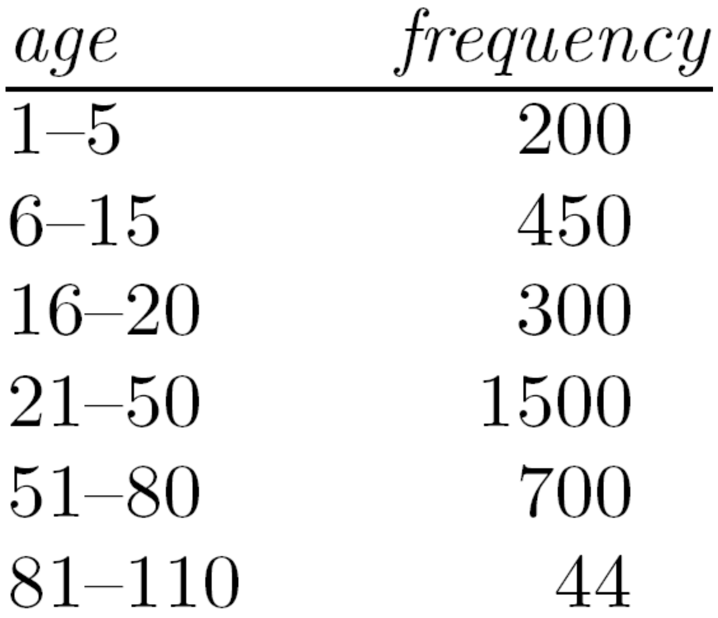
- You can use either population parameters or sample statistics.

Measuring the Central Tendency (중심성)

- **Mean**

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

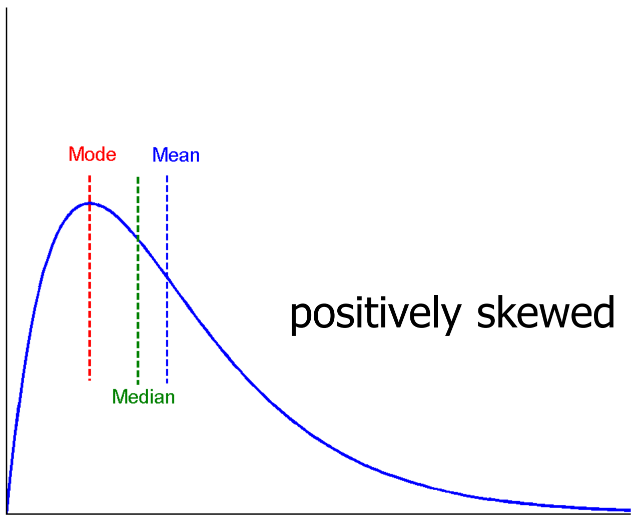
- **Median**
  - Middle value if odd number of values
  - Otherwise, average of the middle two values
- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

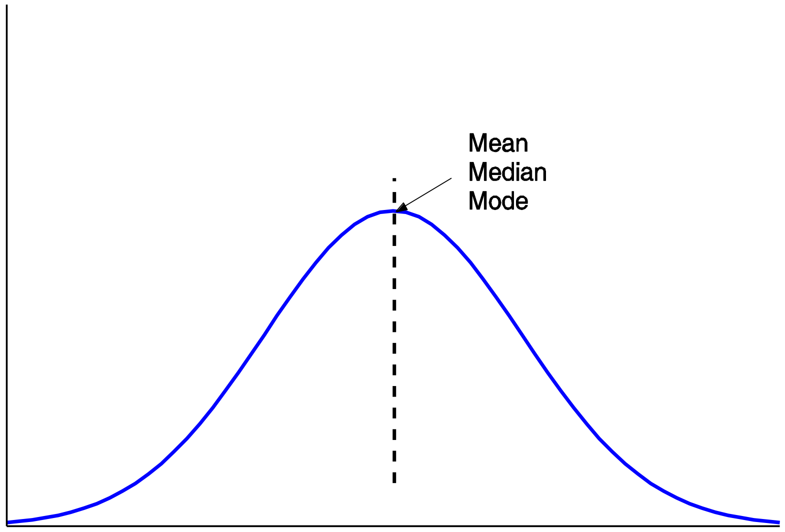


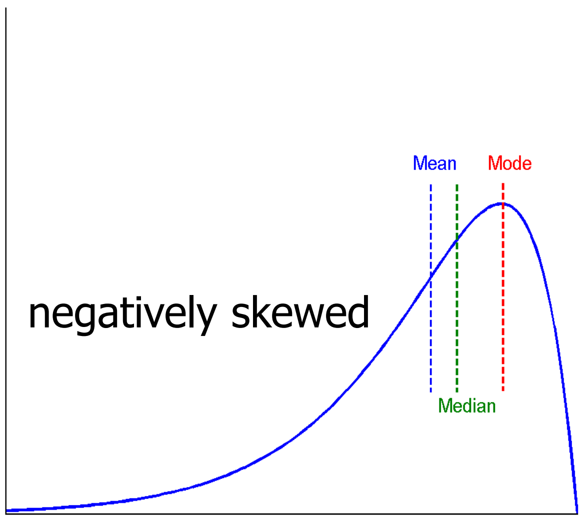
age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Symmetric vs. Skewed Data

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







Measuring the Dispersion of Data (산포도)

- 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)

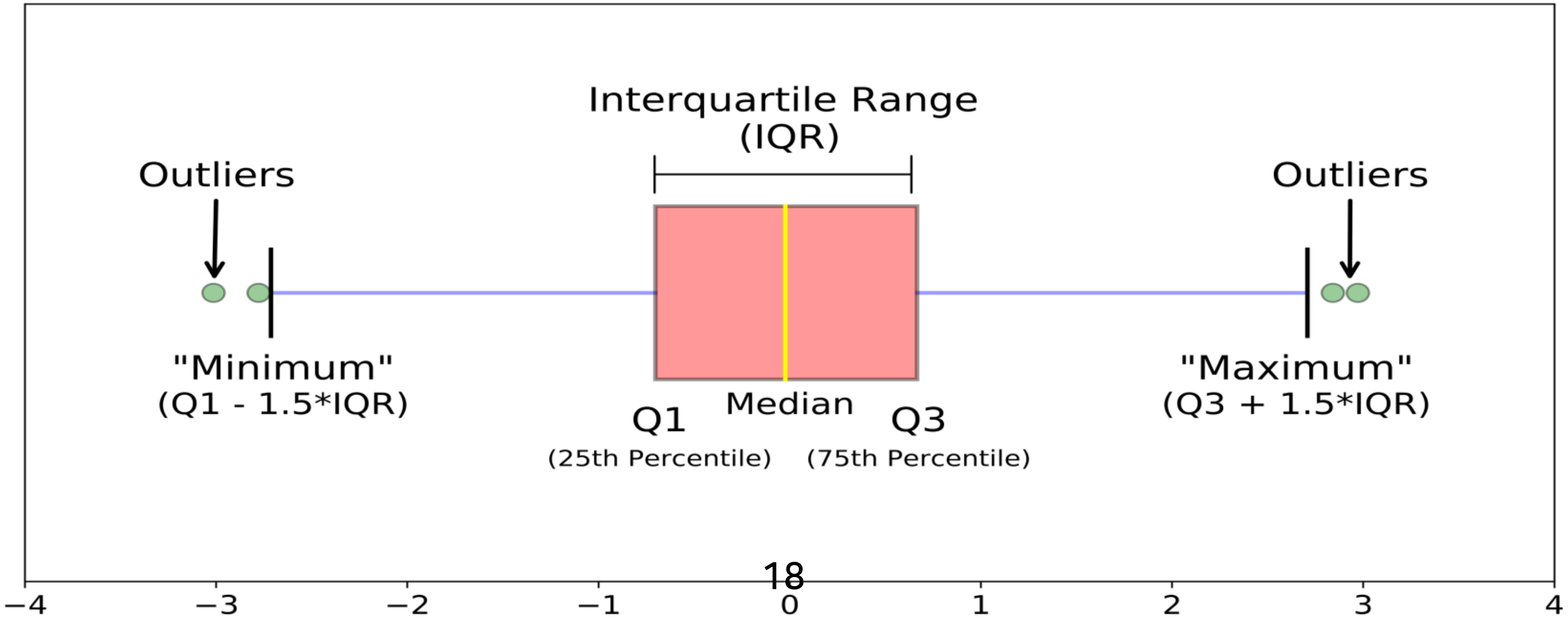
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Measuring the Dispersion of Data (산포도)

- **Quartiles (사분위수)**:
  Q1 (lower quartile, 25th percentile), Q3 (higher quartile, 75th percentile)
- **Inter-quartile range (IQR)**: Q3 – Q1
- **Five number summary**: min, Q1, median, Q3, max
- **Outlier (이상치)**:
  a value higher than Q3 + 1.5 x IQR or lower than Q1 – 1.5 x IQR

Boxplot Analysis

- Boxplot: **five-number summary** of a distribution
  - 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



Source:
KDnuggets

Contents

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

Why Data Visualization?

“Data visualization (데이터 시각화) is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.” (IBM)

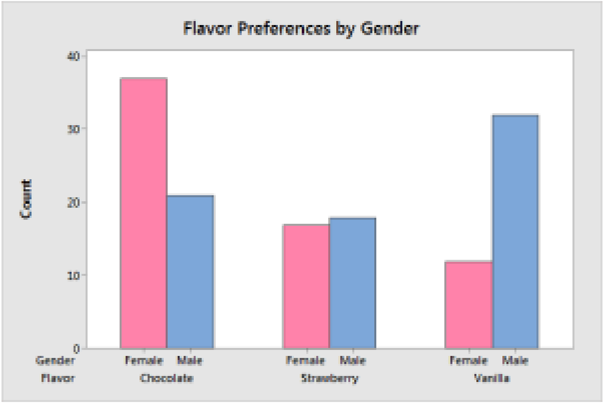
- **Gain insight** into an information space by mapping data onto graphical primitives
- **Provide qualitative overview** of large data sets
- **Search** for patterns, trends, structure, irregularities, relationships
- **Help find interesting regions and suitable parameters** for further quantitative analysis

Graphic Displays of Basic Statistical Descriptions

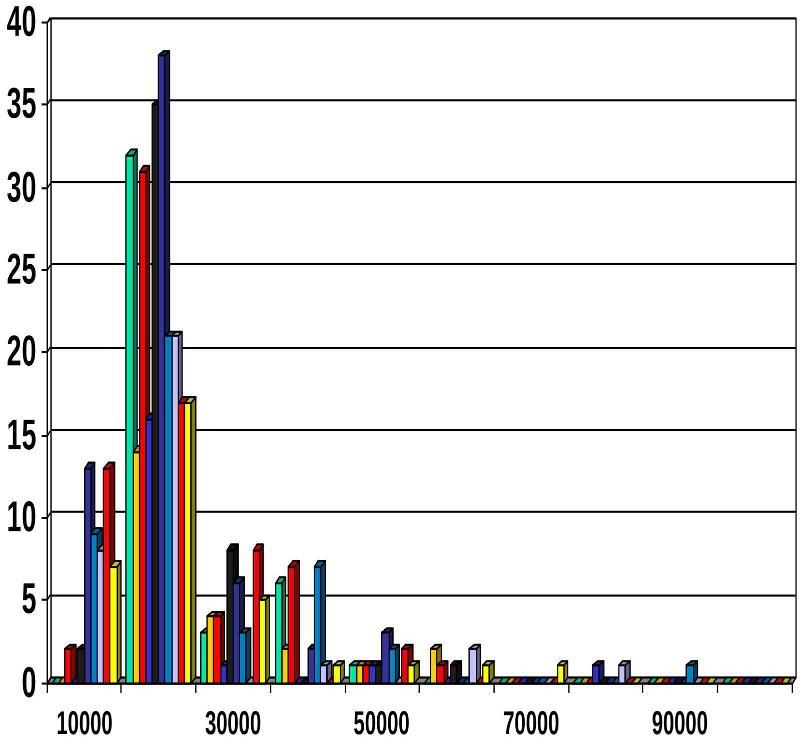
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represents frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points

Histogram

참고: bar chart (or barplot)

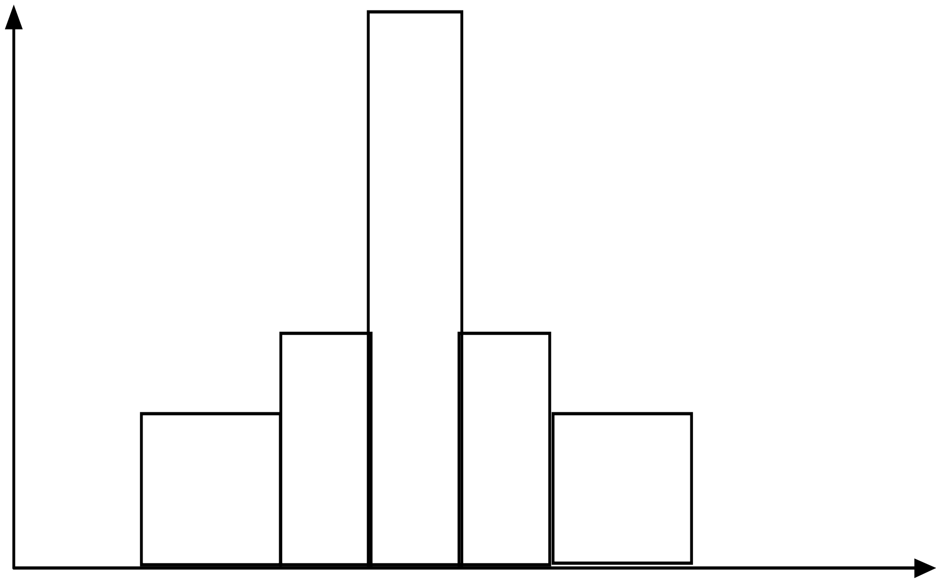


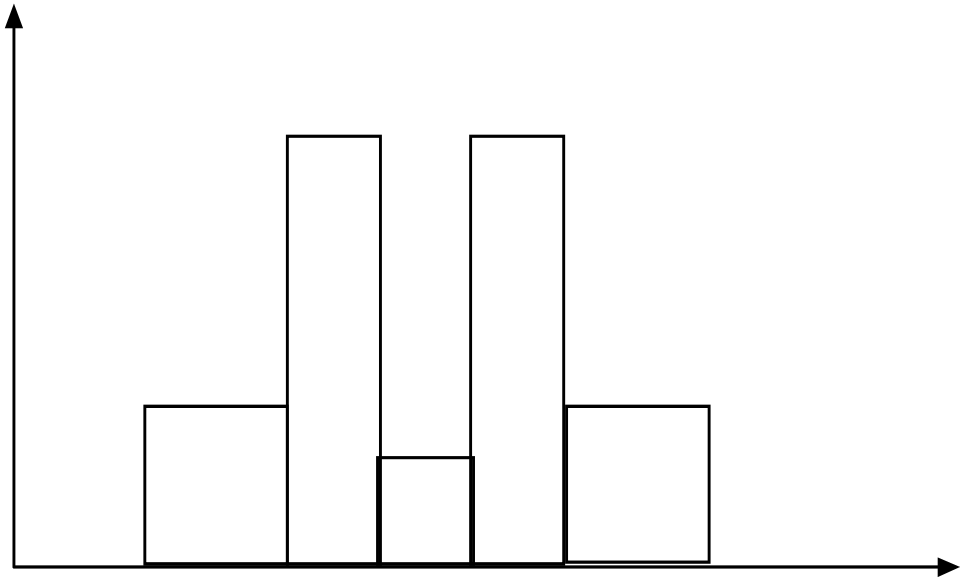
- Graph display of frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

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





An alternative: Violin Plot

- combines the strengths of histogram and boxplots

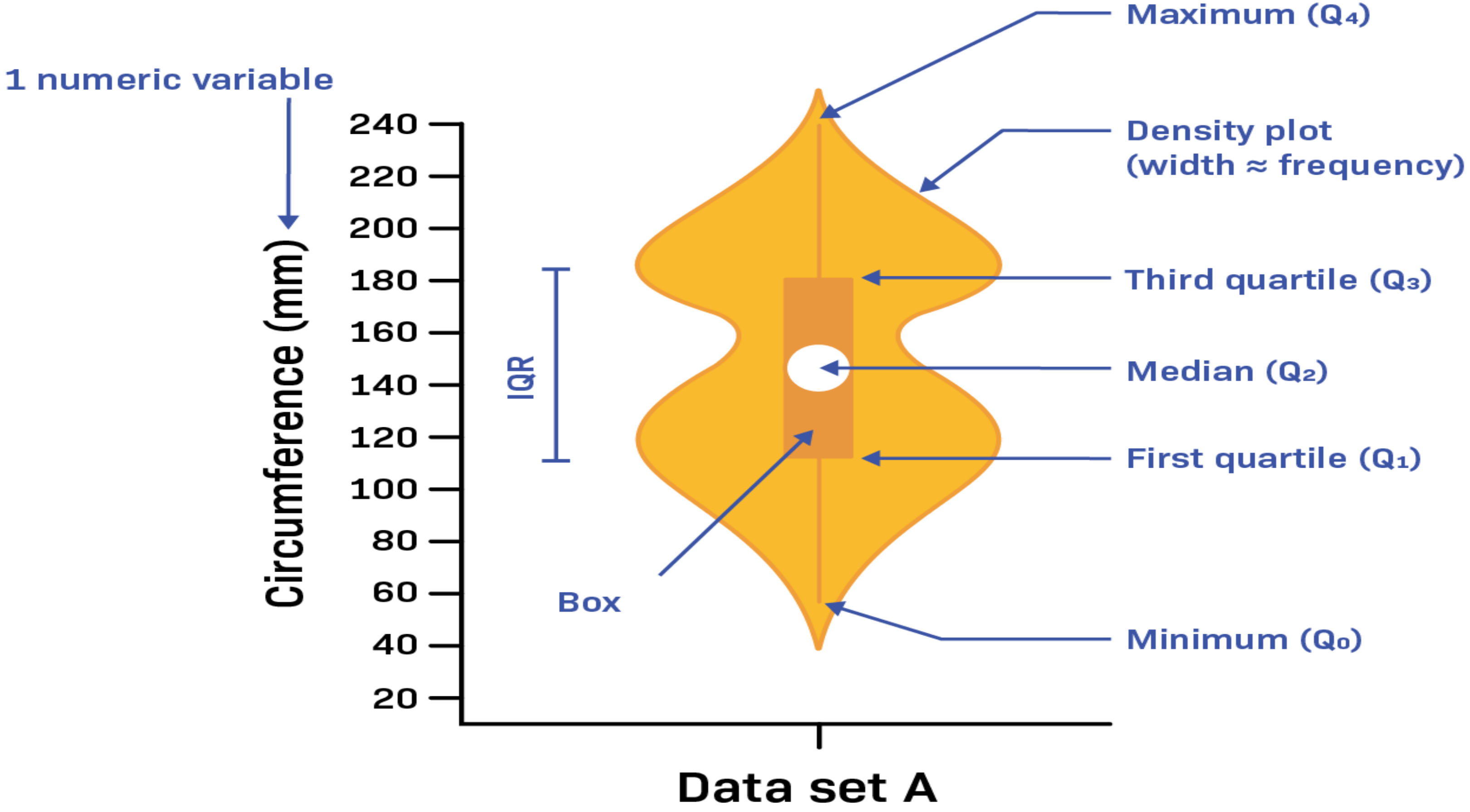
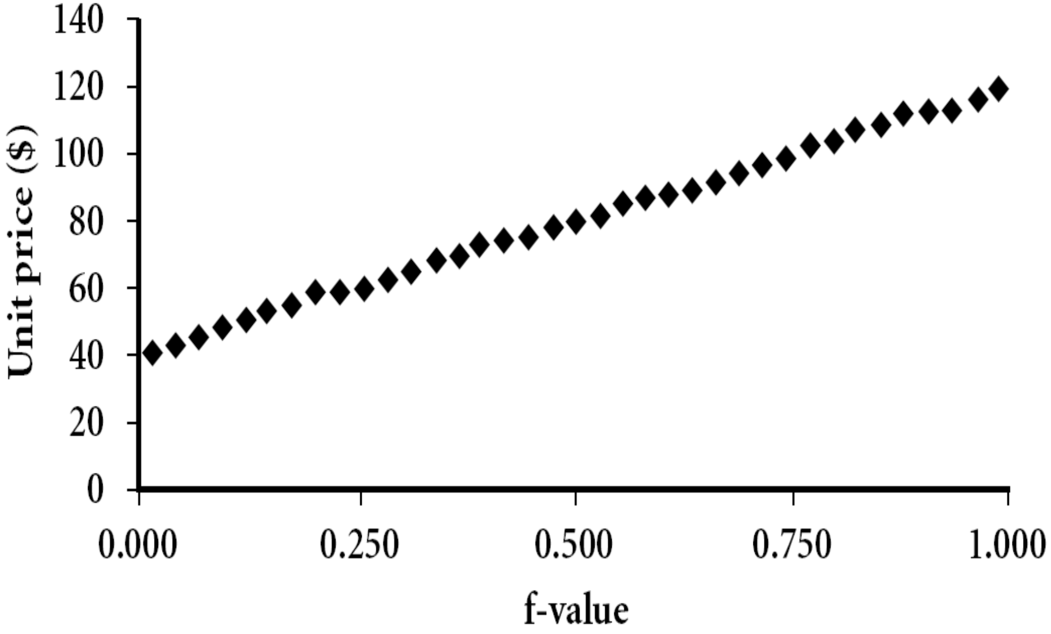


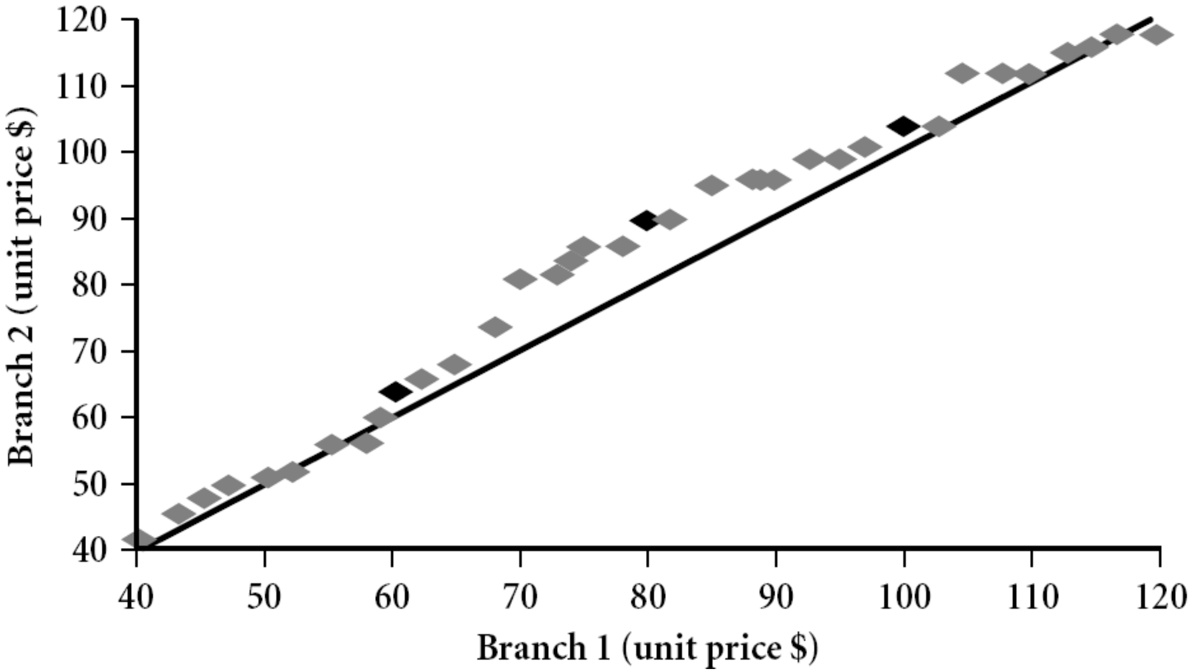
Image source: LabXchange

Quantile Plot

- Displays all of the data, allowing the user to assess both the overall behavior and unusual occurrences
- Plots **quantile** information
- Definitions
  - q-quantile: values that partition a finite set of values into q subsets of equal sizes ($q = 4$; $\mathrm{Q}_1, \mathrm{Q}_2, \mathrm{Q}_3$ widely known as qua**r**tile)
  - 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



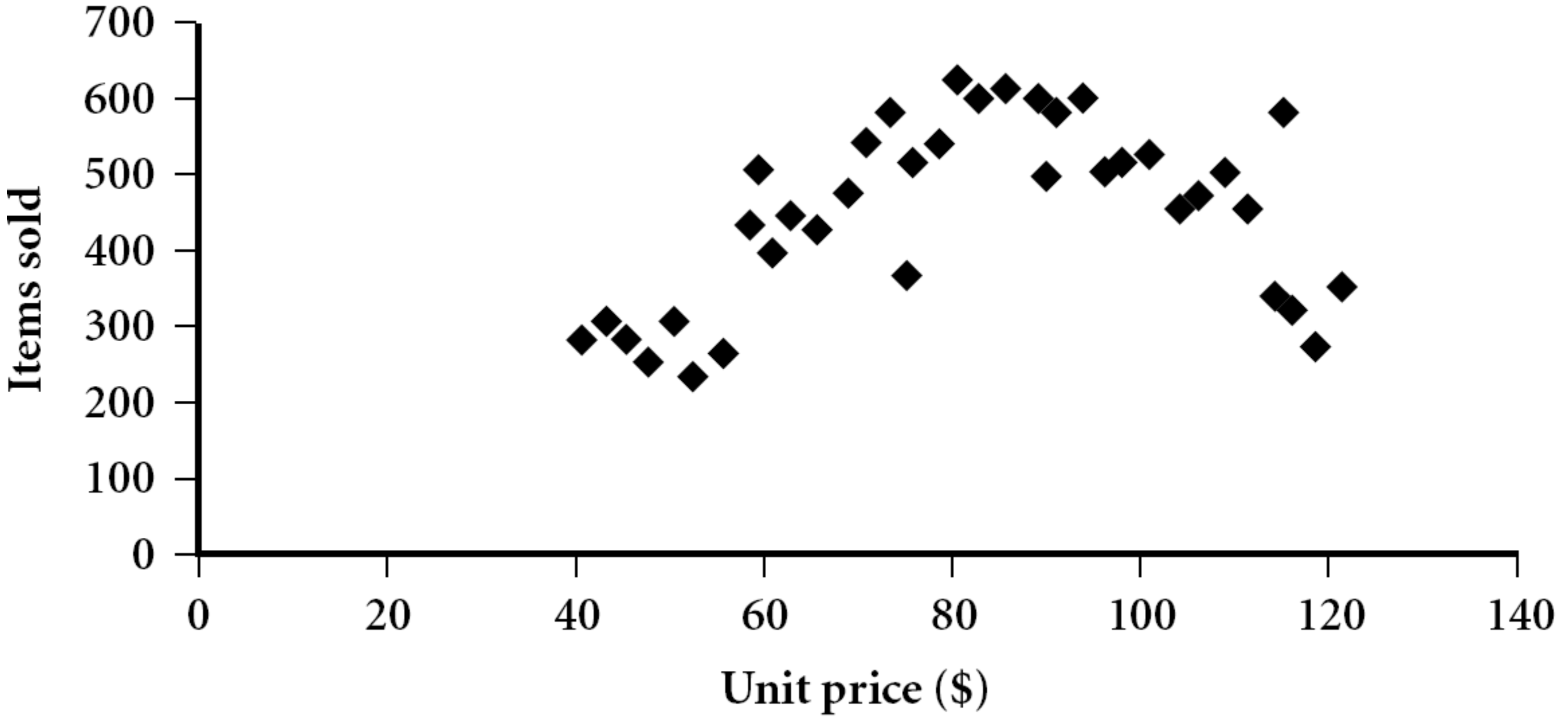
Quantile-Quantile (Q-Q) Plot



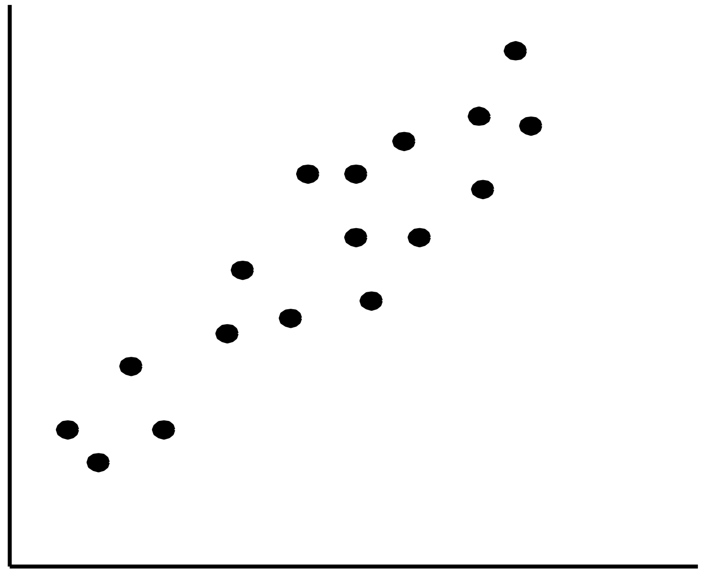
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Commonly used for regression analysis

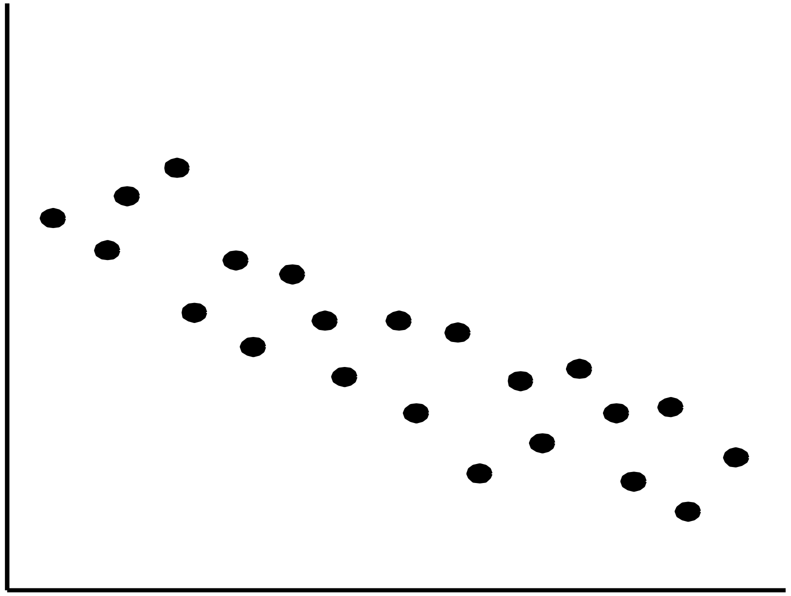
Scatter plot (산점도)

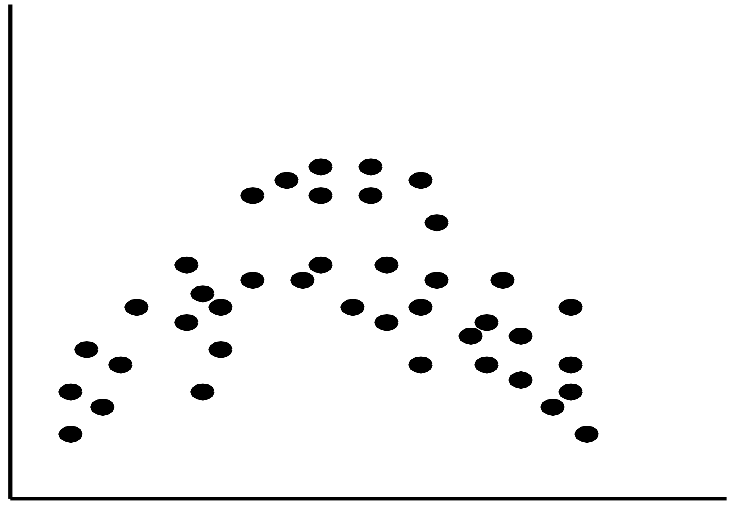
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points



Positively and Negatively Correlated Data



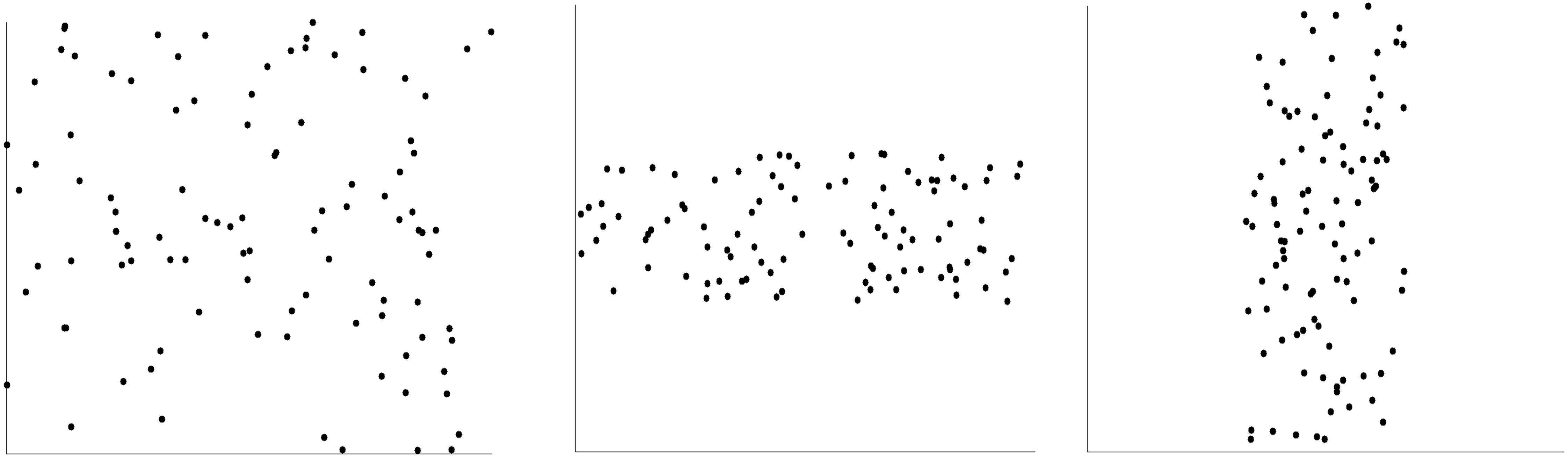




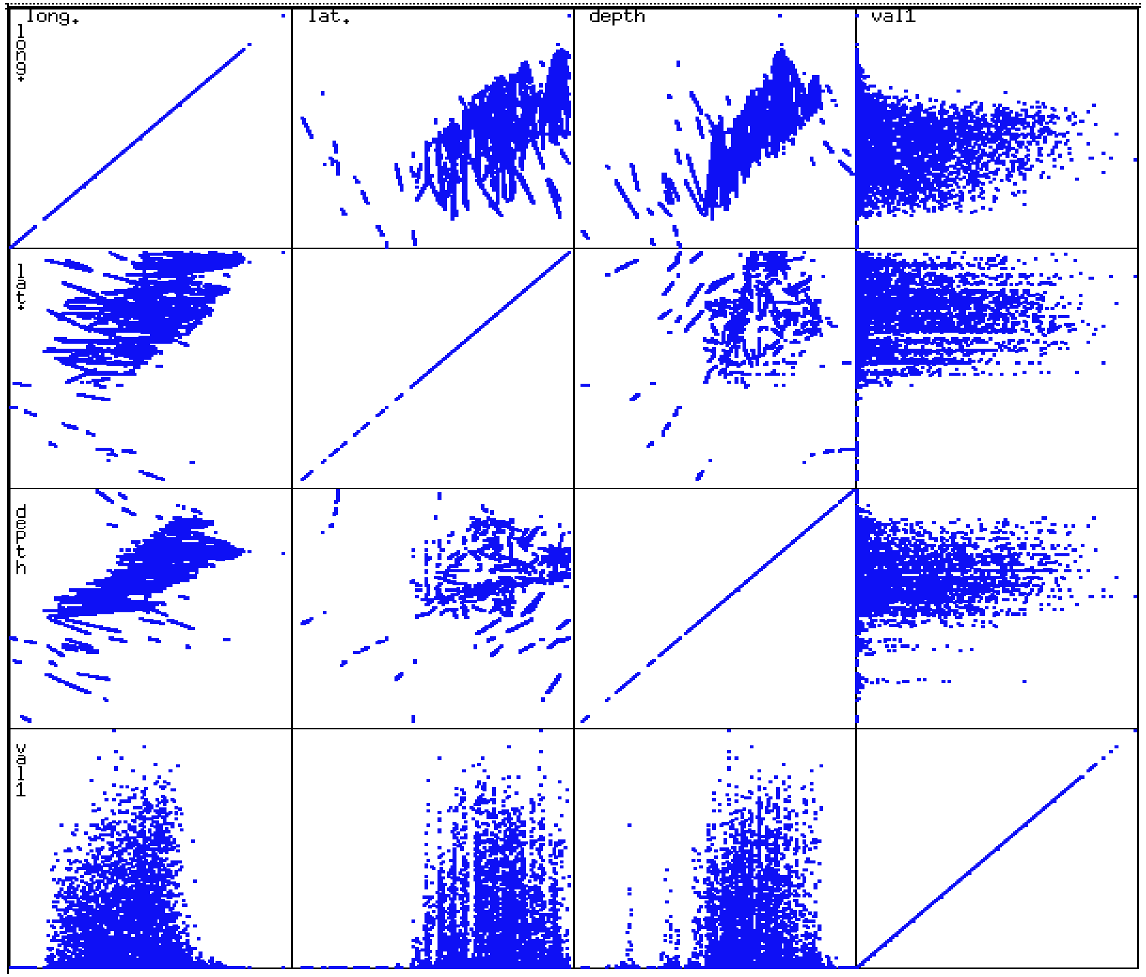
The left half fragment is positively correlated

The right half is negatively correlated

Uncorrelated Data

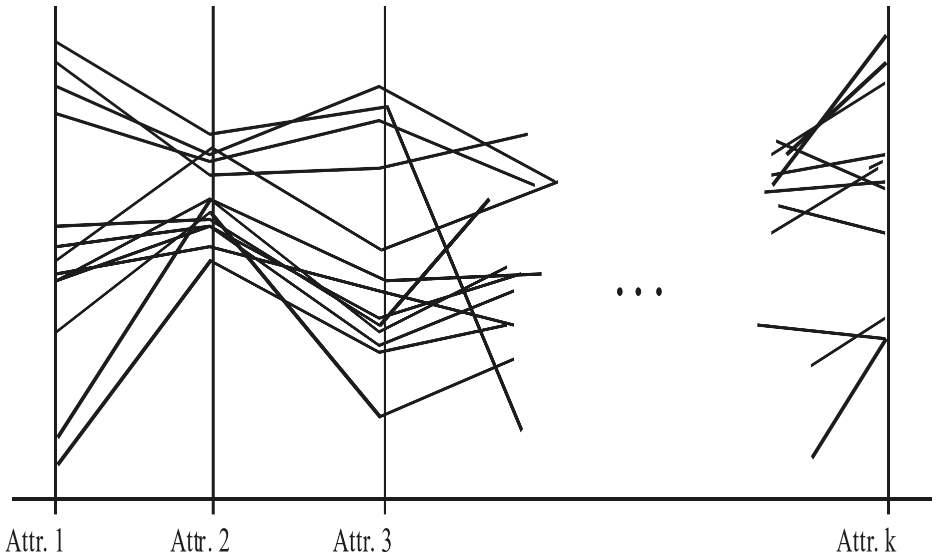


Scatterplot Matrices



Parallel Coordinates (평행 좌표)

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the $[min, max]$: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



Parallel Coordinates of a Data Set

It helps to figure out the trend of inter-variable relationships

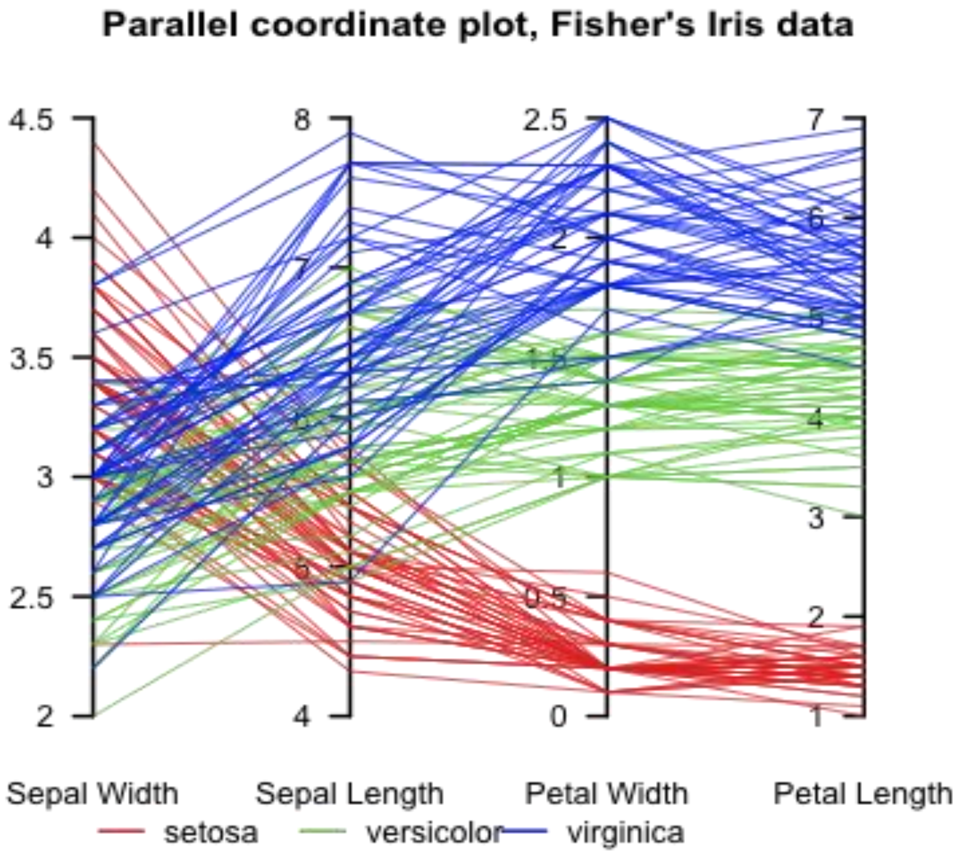


Image source: Wikipedia