Data Mining:

Concepts and Techniques

— Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei

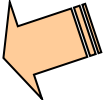University of Illinois at Urbana-Champaign

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	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

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 (**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**
    - Measured on a scale of **equal-sized units**
    - Values have order
        - E.g., *temperature in C° or F°, calendar dates*
    - No true zero-point
- **Ratio**
    - 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
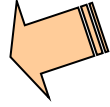
- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - 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

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

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

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

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

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

  - 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, 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

Symmetric vs. Skewed Data

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

symmetric

Mean
Median
Mode

Mode   Mean

Median

positively skewed

Mean   Mode

Median

negatively skewed
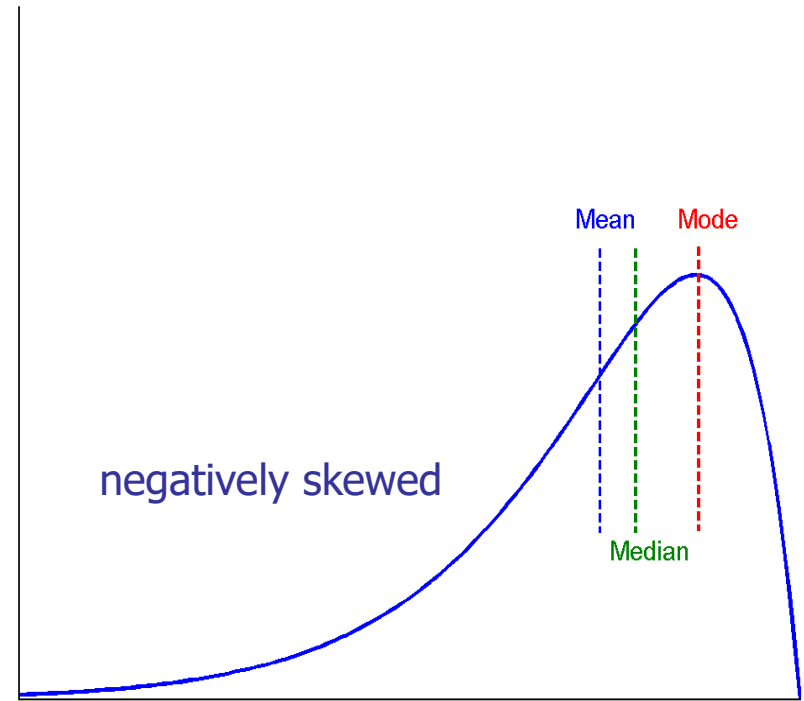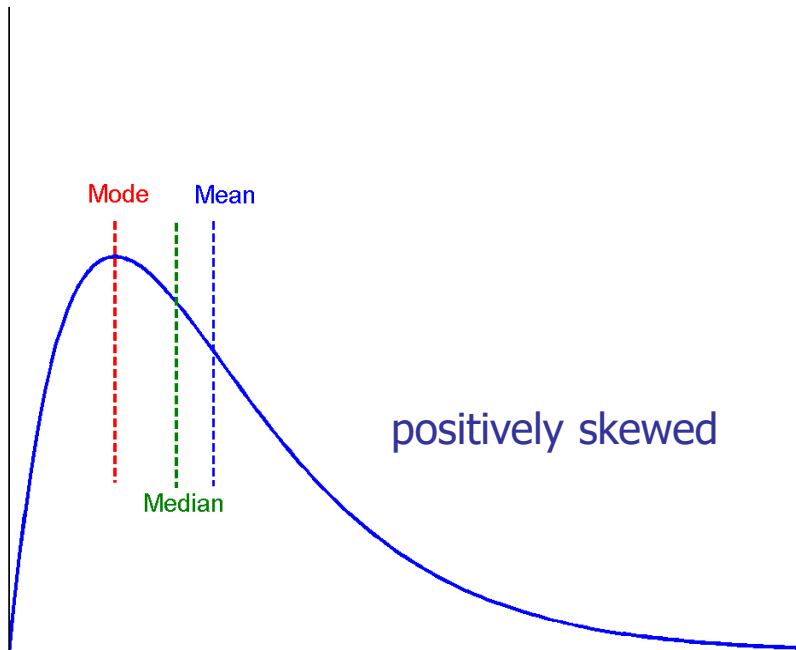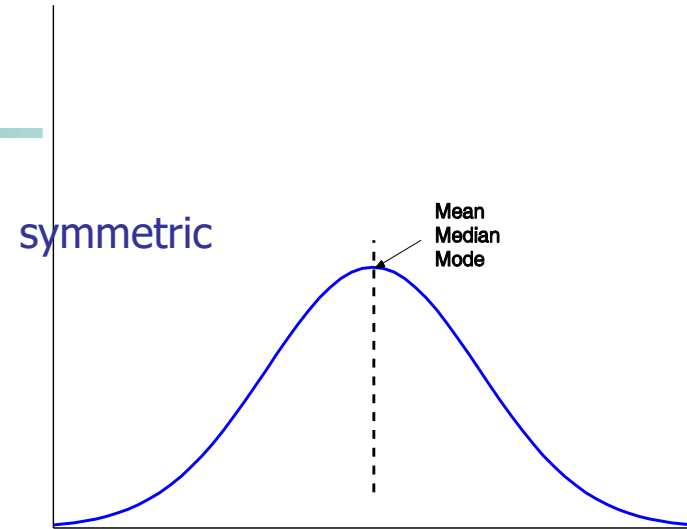
Measuring the Dispersion of Data

- 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 x IQR

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

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

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

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

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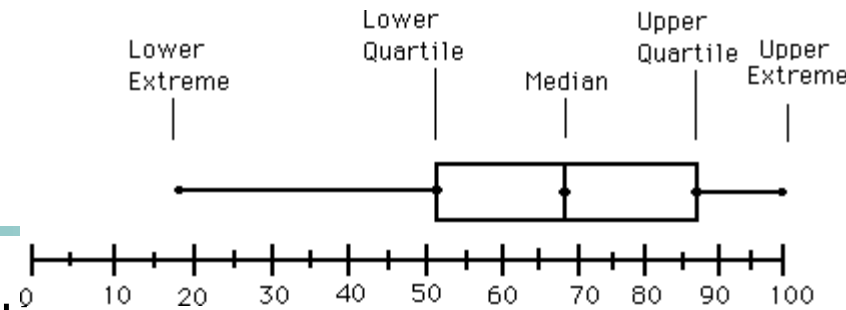
- **Five-number summary** of a distribution

  - 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

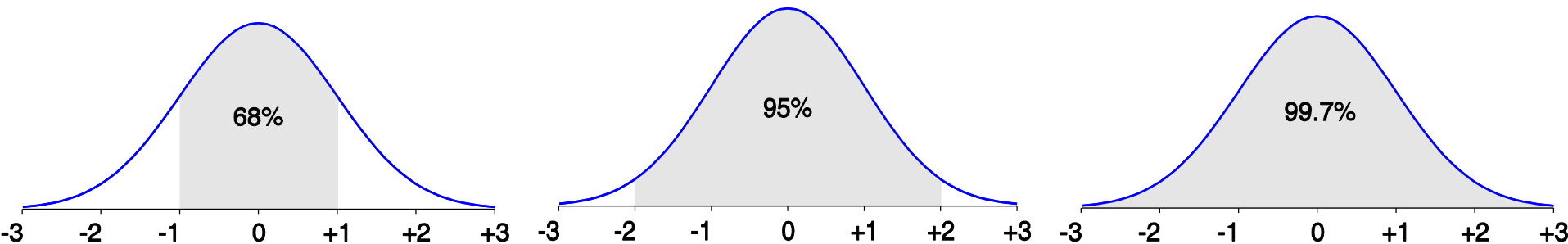 - Outliers: points beyond a specified outlier threshold, plotted individually
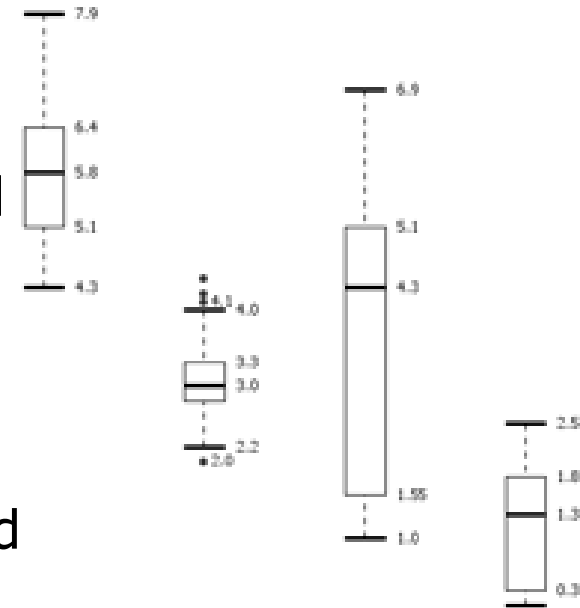
Visualization of Data Dispersion: 3-D Boxplots
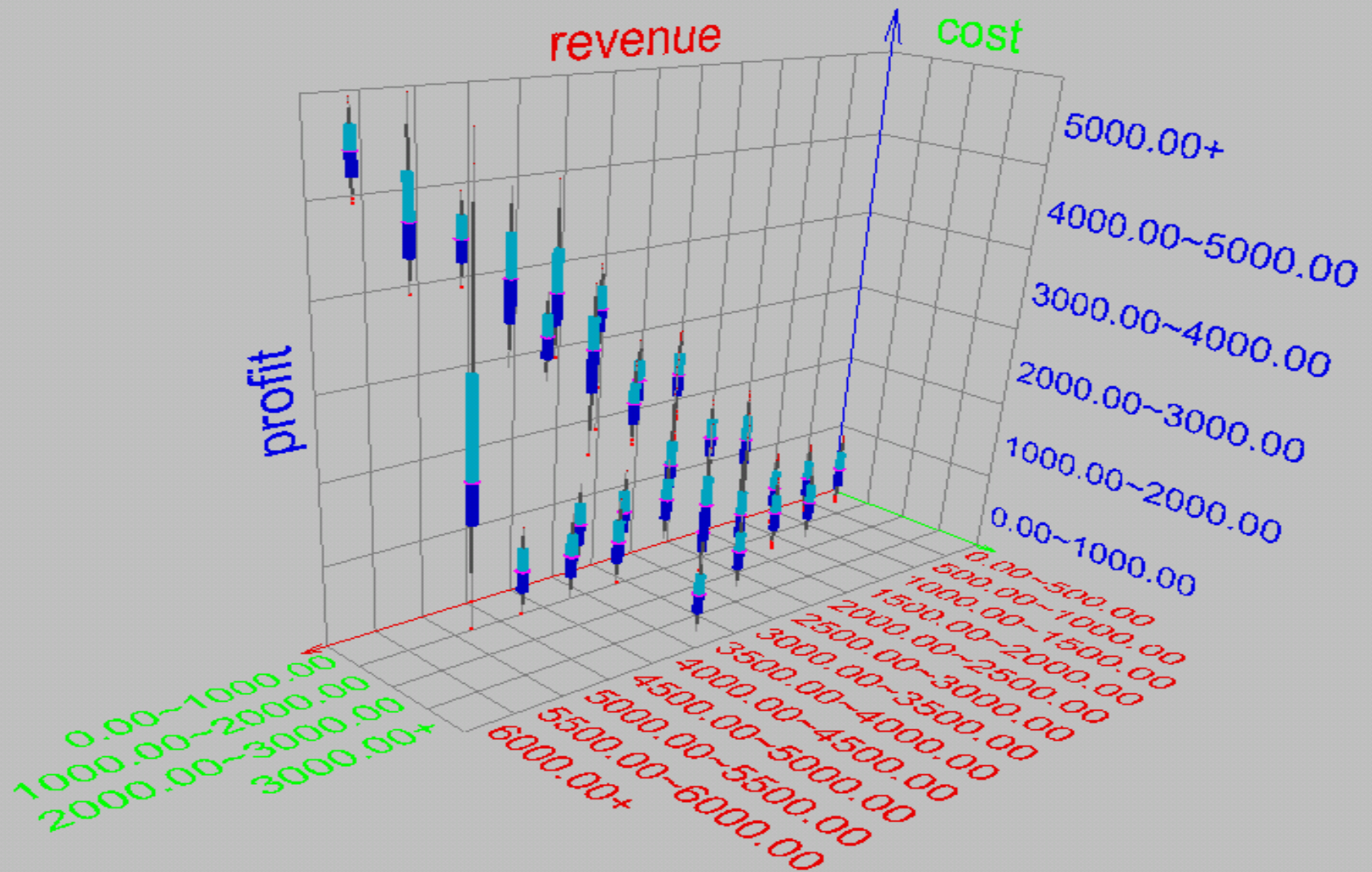
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
 - From μ–3σ to μ+3σ: contains about 99.7% of it

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary

- **Histogram**: x-axis are values, y-axis repres. 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

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars

- It shows what proportion of cases fall into each of several categories

- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width

- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent
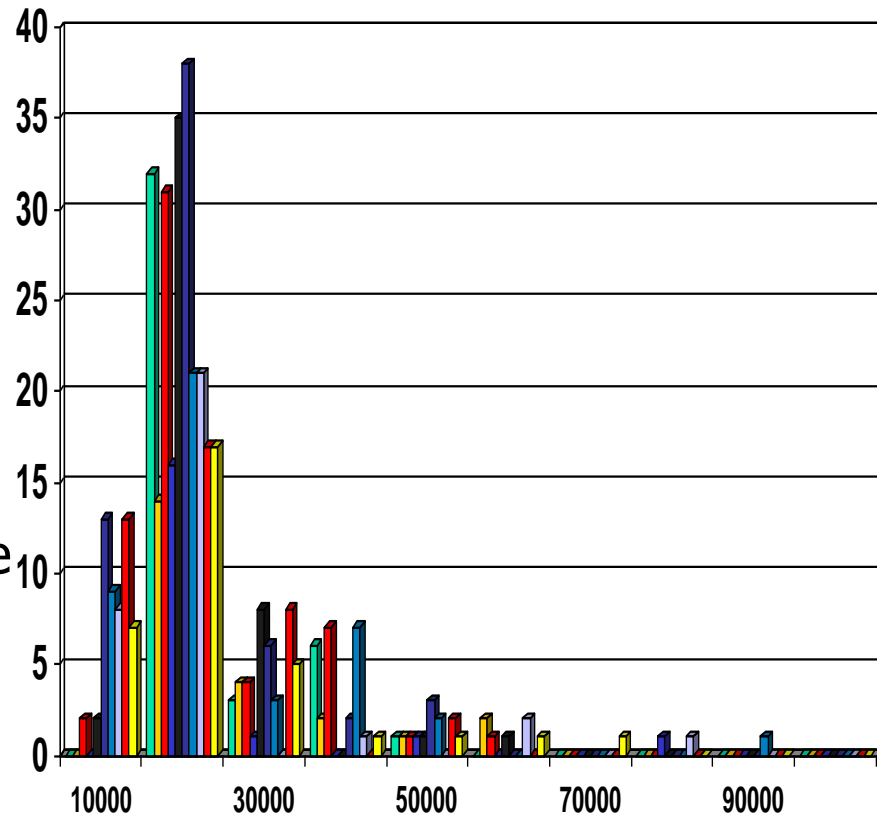
Histograms Often Tell More than Boxplots
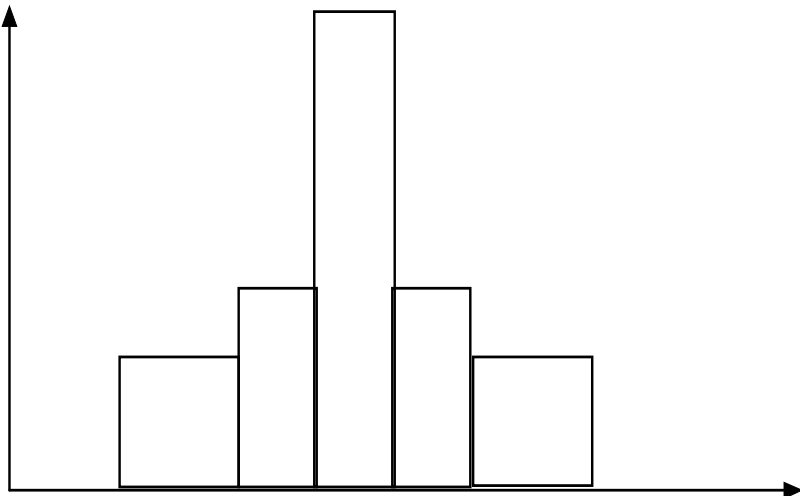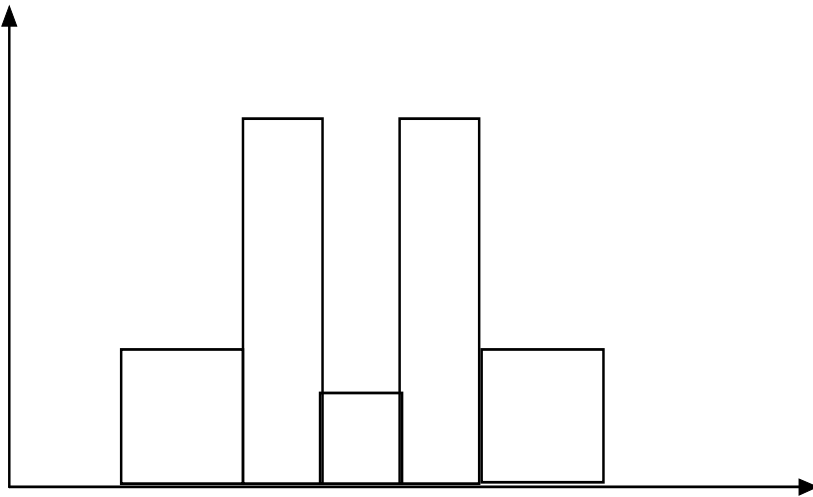
- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Quantile Plot

- 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
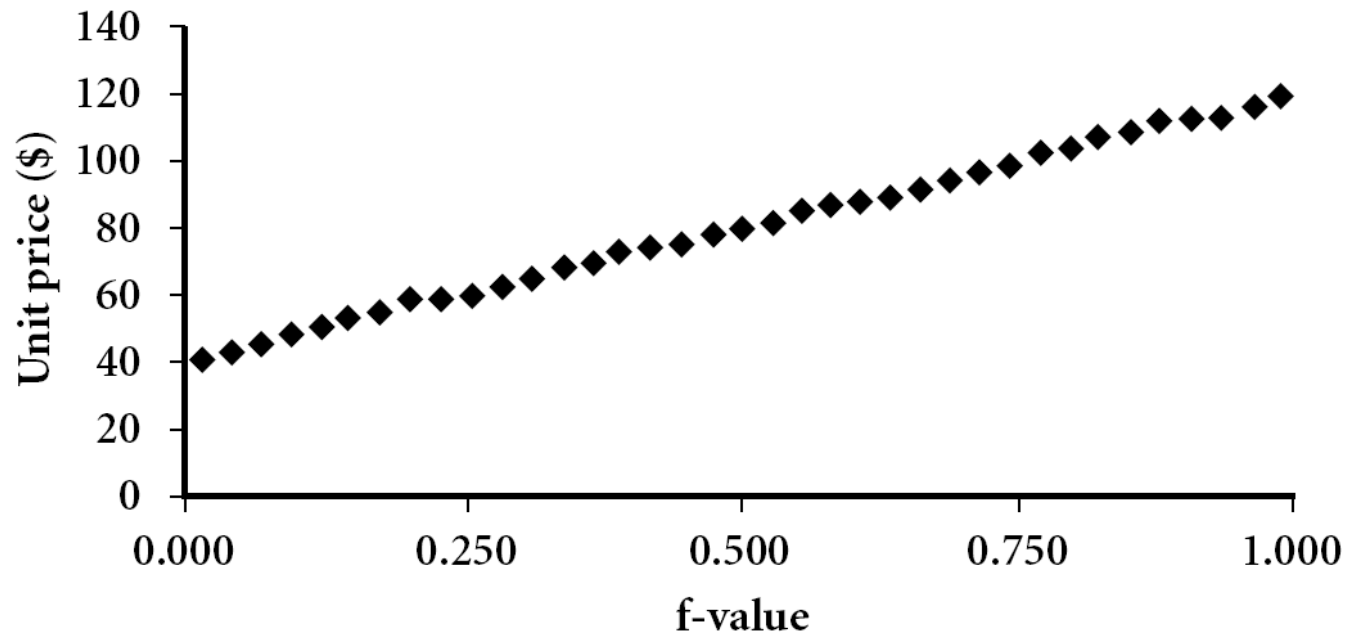
Quantile-Quantile (Q-Q) Plot

- 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.
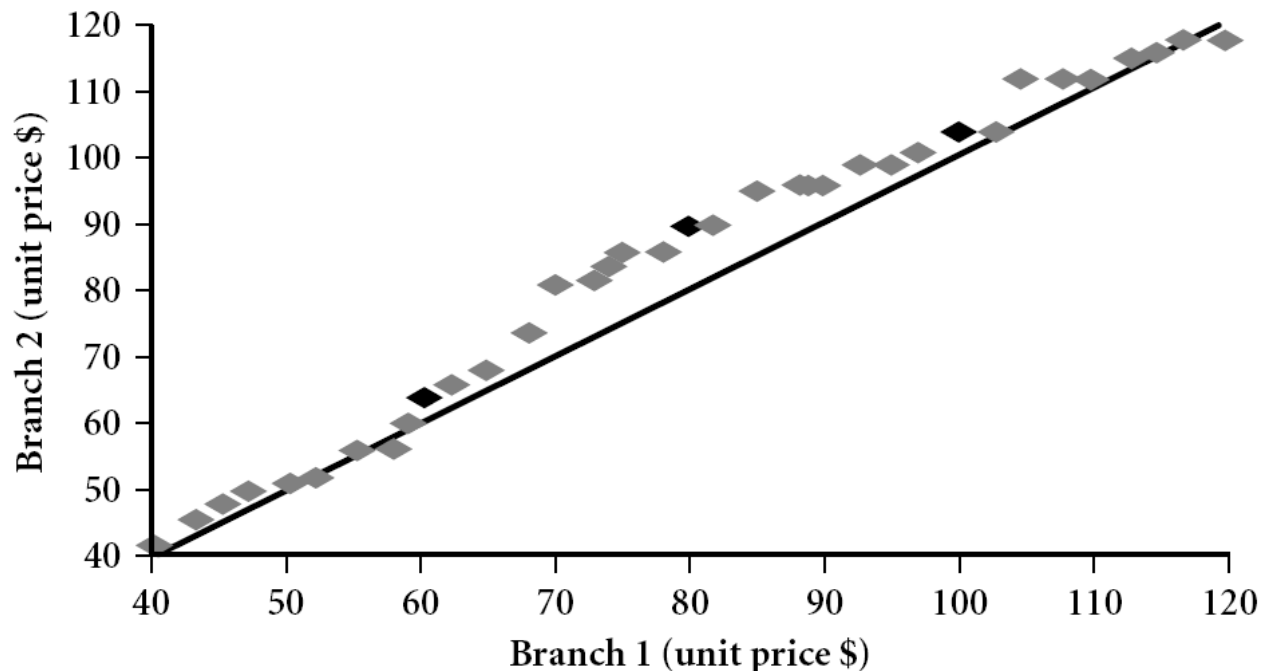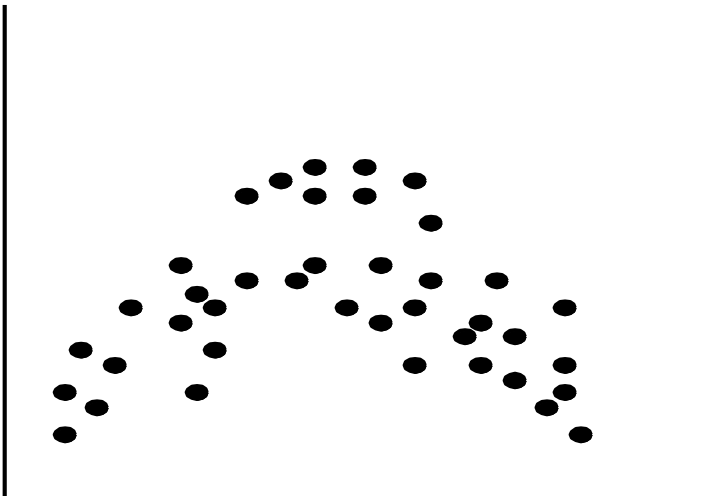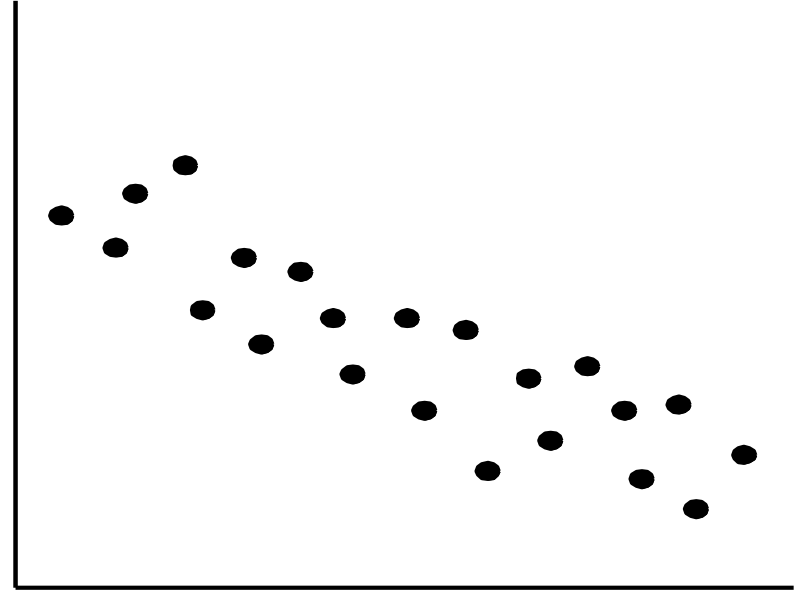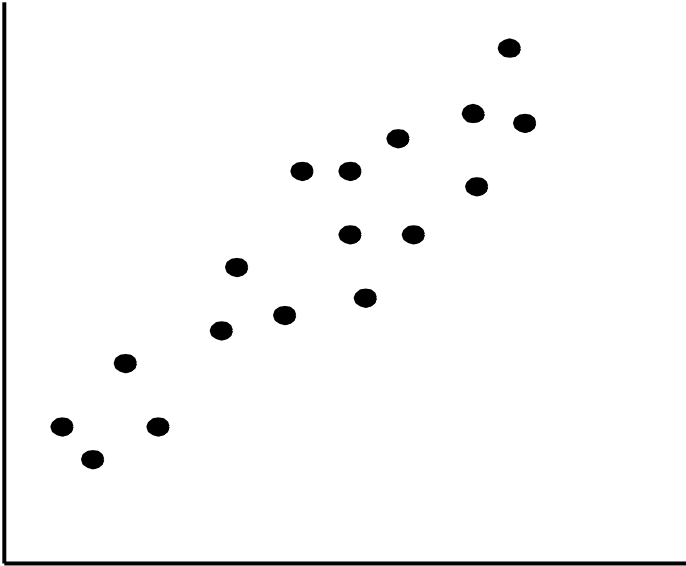
Scatter plot

- 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
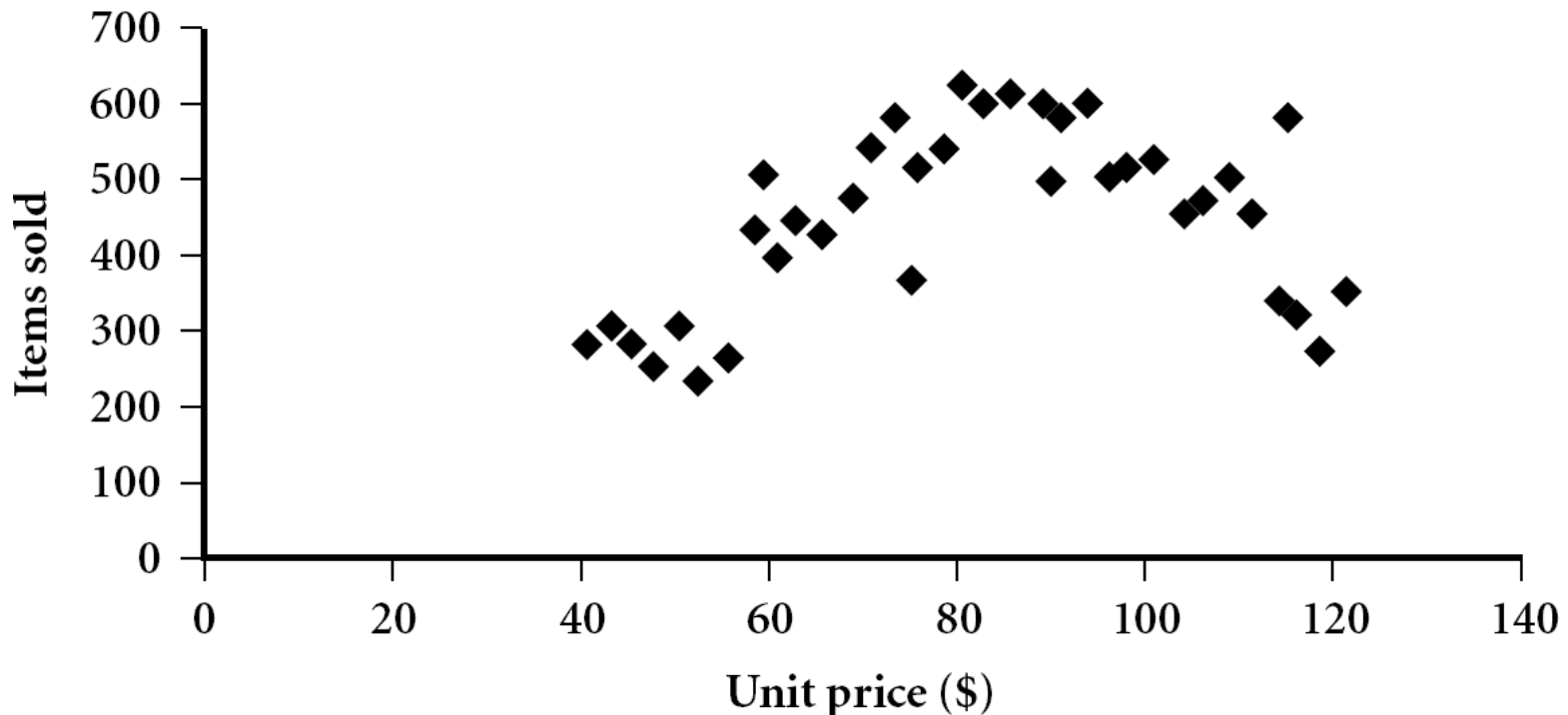
Positively and Negatively Correlated Data

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

Uncorrelated Data