

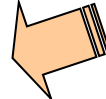
Data Mining:

Concepts and Techniques

— Chapter 2 —

Slides Curtesy of Textbook

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

Entity / Rel

Types of Data Sets

Student
netid | age | ph.

Struct data / Tabular

million

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	n	wi	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

personality

Student
10

Netflix

web

Viewer < Page

100k

10^6

leverage
address

Zoom in/out
Drill & / Roll u.

	m_1	m_2	m_3	...
John	X		X	

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples: *Entity*
 - 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
 - Ordinal
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled
- Handwritten notes:*
{ rep. distinct obj
sufficient
features

by name 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** *→ Ranking* *L - m ≠ m - s*
 - 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*

Handwritten notes illustrating Interval and Ratio scales:

Interval scale examples: 90°F, 45°F, 30°F. Brackets above 90°F and 45°F indicate a difference of 45. A bracket above 15 and 30°F indicates a difference of 15.

Ratio scale example: $90^\circ\text{F} = 2 \text{ times } 45^\circ$

Discrete vs. Continuous Attributes

■ Discrete Attribute

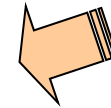
- Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

■ 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

Trend

- To better understand the data.
- Central tendency– the center, the representative.
- Dispersion– variation, spread.

Measuring the Central Tendency

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

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

Weighted arithmetic mean:

Trimmed mean: chopping extreme values

Median:

Middle value if odd number of values, or average of the middle two values otherwise

Estimated by interpolation (for grouped data):

Mode

Value that occurs most frequently in the data

Unimodal, bimodal, trimodal

Empirical formula:

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

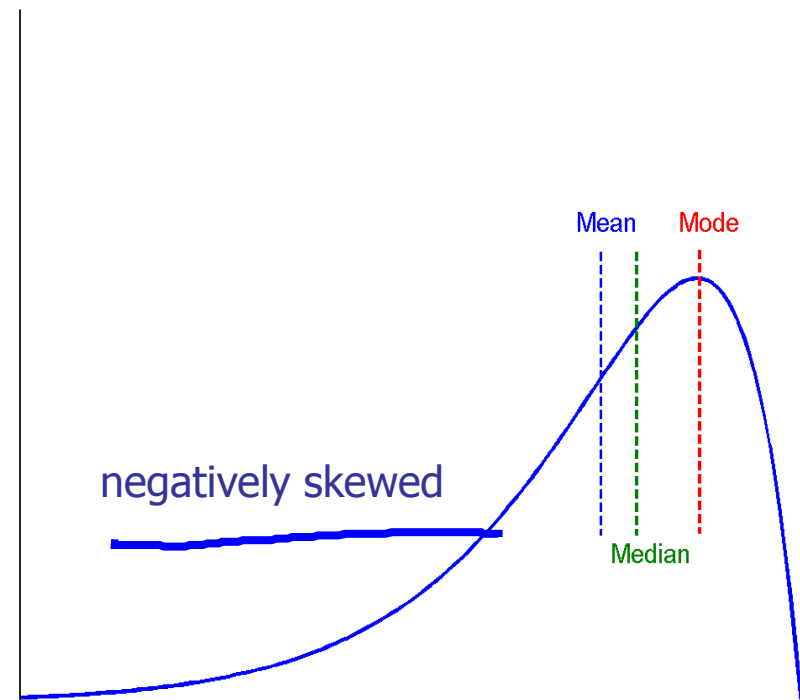
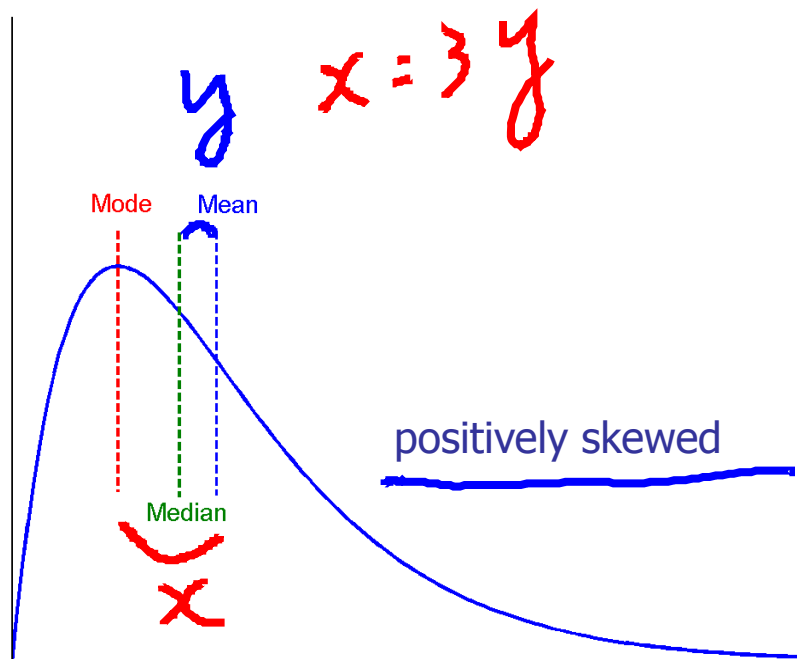
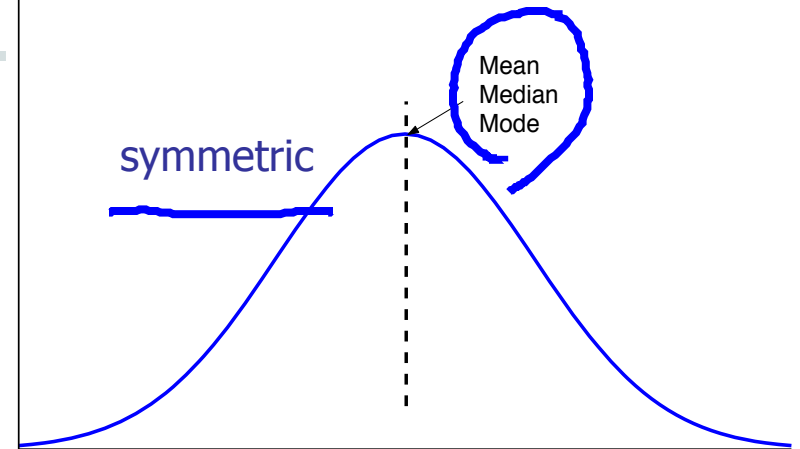
Sample $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ Pop $\mu = \frac{\sum x}{N}$

Grading $\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$ sum weight

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

■ 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

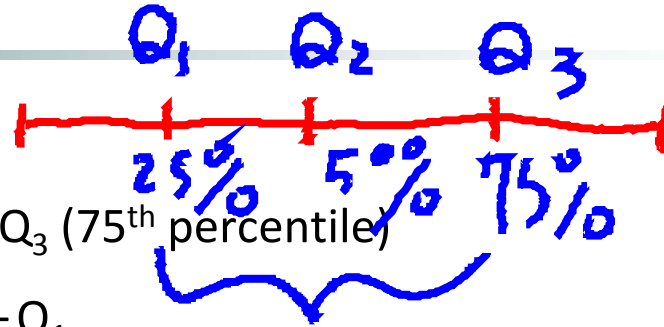
■ **Outlier:** usually, a value higher/lower than $1.5 \times IQR$

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

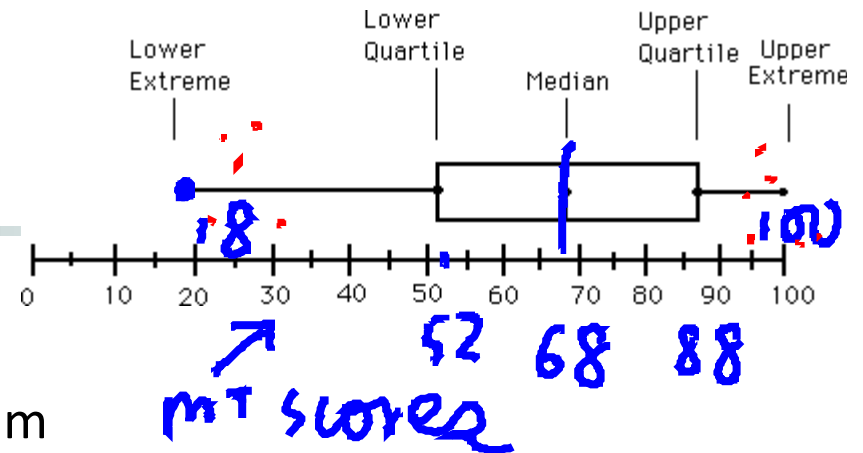
■ **Variance:** (algebraic, scalable computation)

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

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



Boxplot Analysis

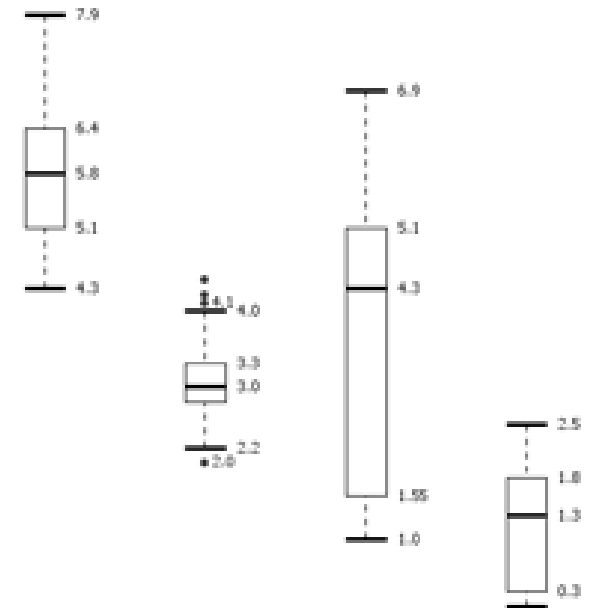


- **Five-number summary** of a distribution

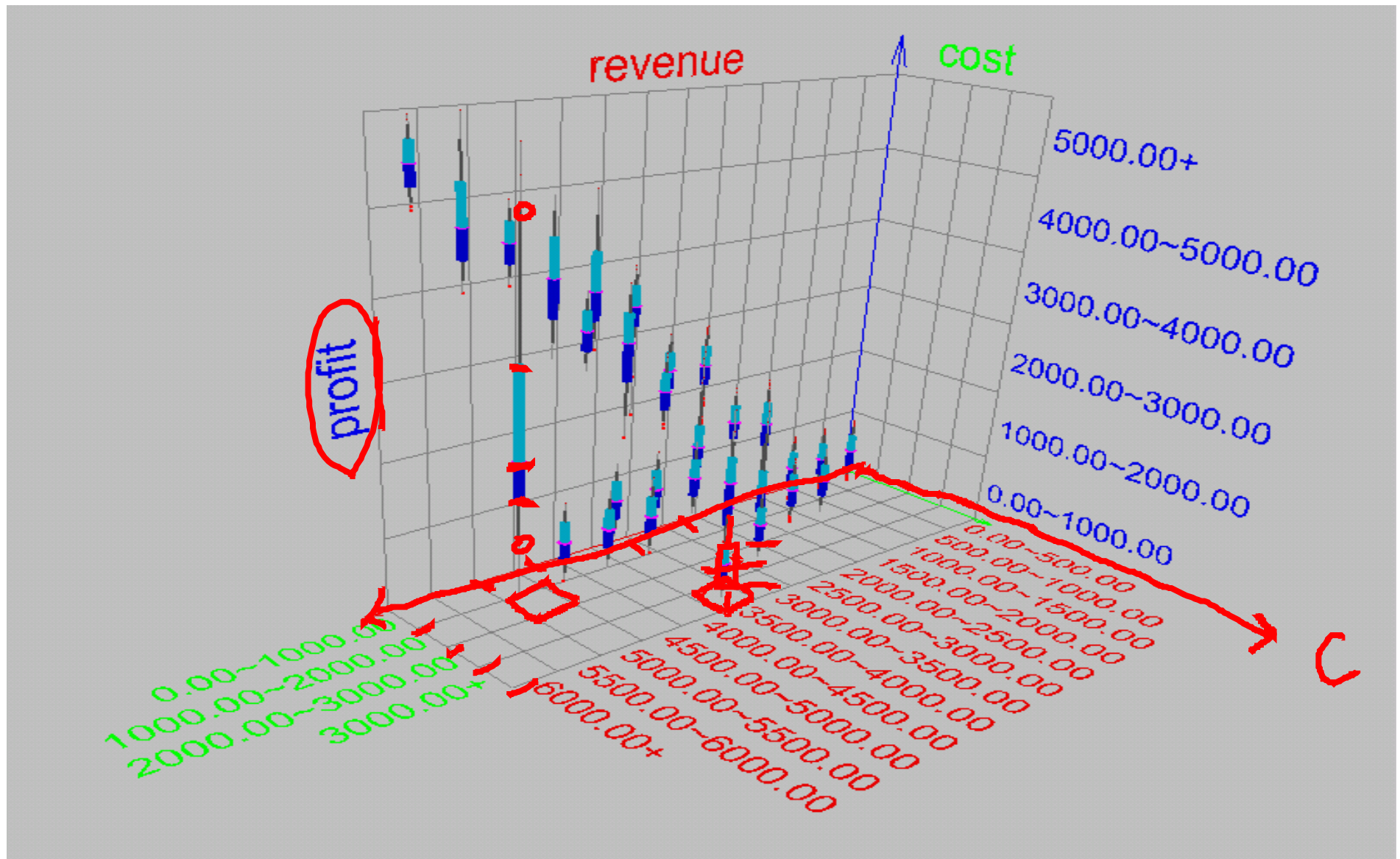
- Minimum, Q1, Median, Q3, Maximum

- **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

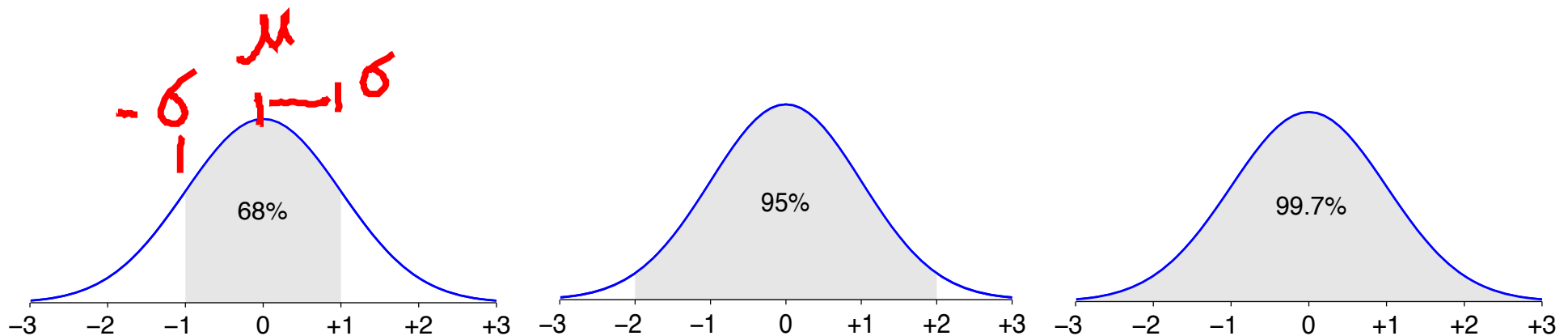


Visualization of Data Dispersion: 3-D Boxplots



Properties of Normal Distribution Curve

- The normal (distribution) curve
 - From $\mu - \sigma$ 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



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