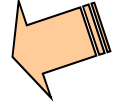


Chapter 2: Getting to Know Your Data

- 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
 - Text documents: term-frequency vector
 - Transaction 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

- Graph and network
 - Social or information networks
 - World Wide Web
 - 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

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



Important Characteristics of Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only a small portion of presence
- 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, objects*
- Data objects are described by **attributes**
- Database rows -> data objects; columns -> attributes



Attributes

- **Attribute (or dimensions, features, variables):**
 - Data field, representing a **characteristic** or a **feature** of a data object
 - *E.g., customer_ID, name, address*
- Types:
 - Nominal
 - Binary
 - Numeric: quantitative
 - Ratio-scaled
 - Times meaningful, zero means absence (ex: weight in kg)
 - Interval-scaled:
 - Only difference meaningful (ex: temperature in celcius)



Attribute Types

- **Nominal:** categories, states, or “names of things”
 - Has a finite number of values
 - *Hair_color* = {*black, blond, brown, grey, red, white*}
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - Special case of a 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)



Attribute Types

- **Ordinal**

- Values have a meaningful order (ranking)
- *Magnitude* between successive values is not known though
- $Size = \{small, medium, large\}$, grades, army rankings



Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval-scaled**
 - 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-scaled**
 - Inherent **zero-point (meaning 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, monetary quantities*



Discrete vs. Continuous Attributes

- **Discrete Attribute**

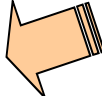
- Has only a *finite* or *countably infinite* set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

- **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, quartiles, outliers, variance, etc.

Measuring the Central Tendency

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

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

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 odd number of values, or average of the middle two values otherwise

- Estimated by interpolation (for *grouped data*):

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq}_l)}{\text{freq}_{\text{median}}} \right) * \text{width}$$

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

■ Example

- $n = 3194$, $n/2 = 1597$, $L_1 = 21$, $\text{freq}_{\text{median}} = 1500$
- Numerator = $1597 - (200 + 450 + 300) = 647$
- width = $(50 - 21) = 29$
- Median = $21 + (647/1500) * (50 - 21)$



Measuring the Central Tendency

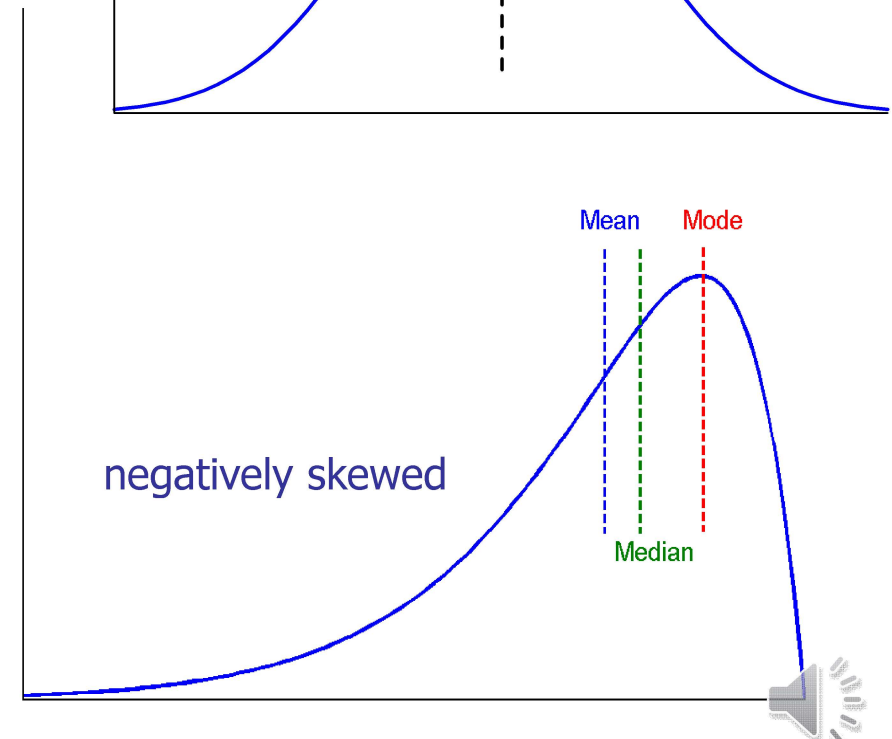
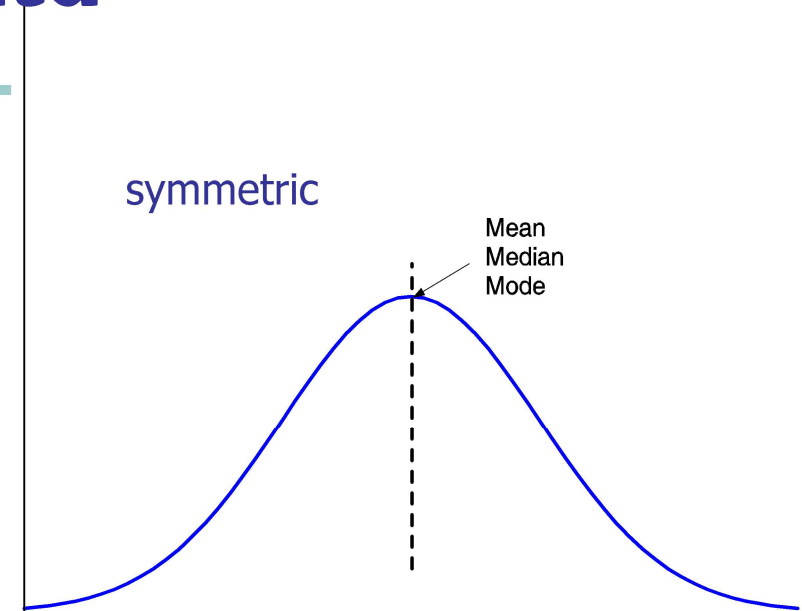
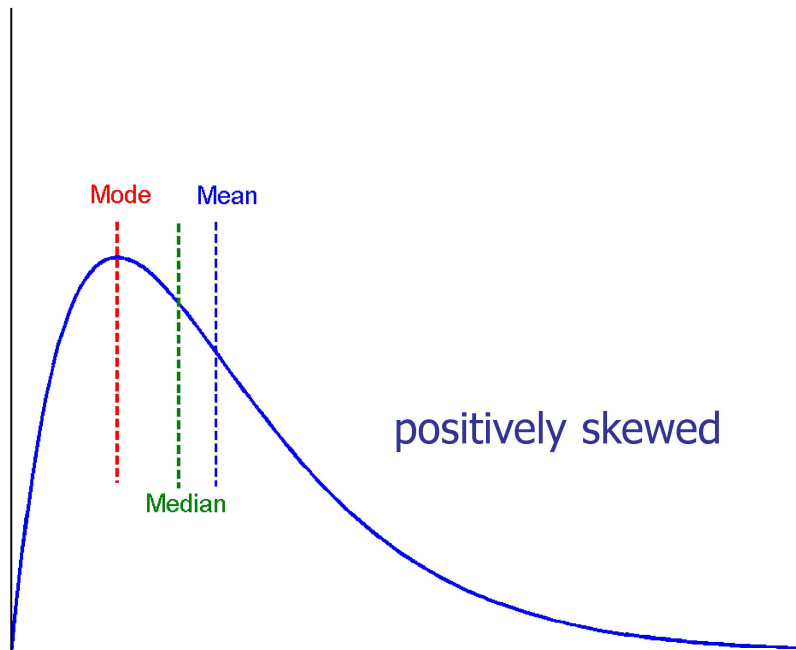
- Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

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

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):** $IQR = Q_3 - Q_1$
 - **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 $1.5 \times IQR$

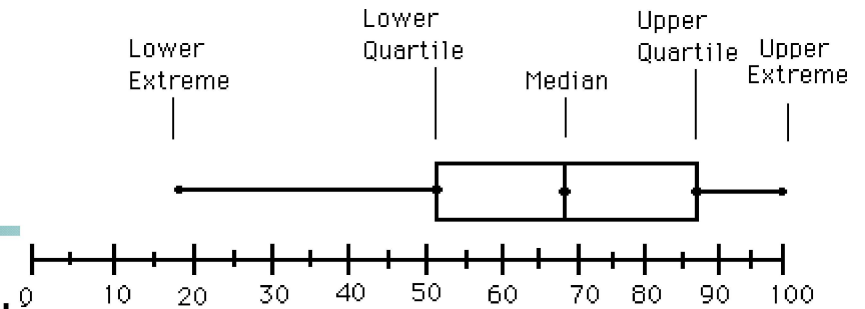
Measuring the Dispersion of Data

- 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

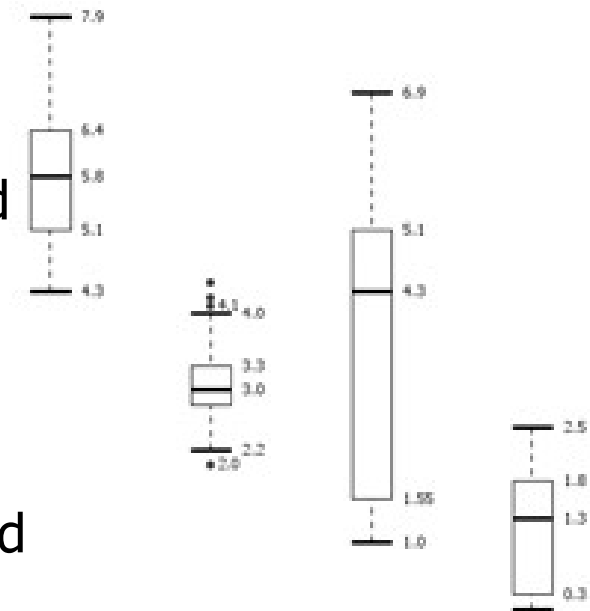


- **Five-number summary** of a distribution

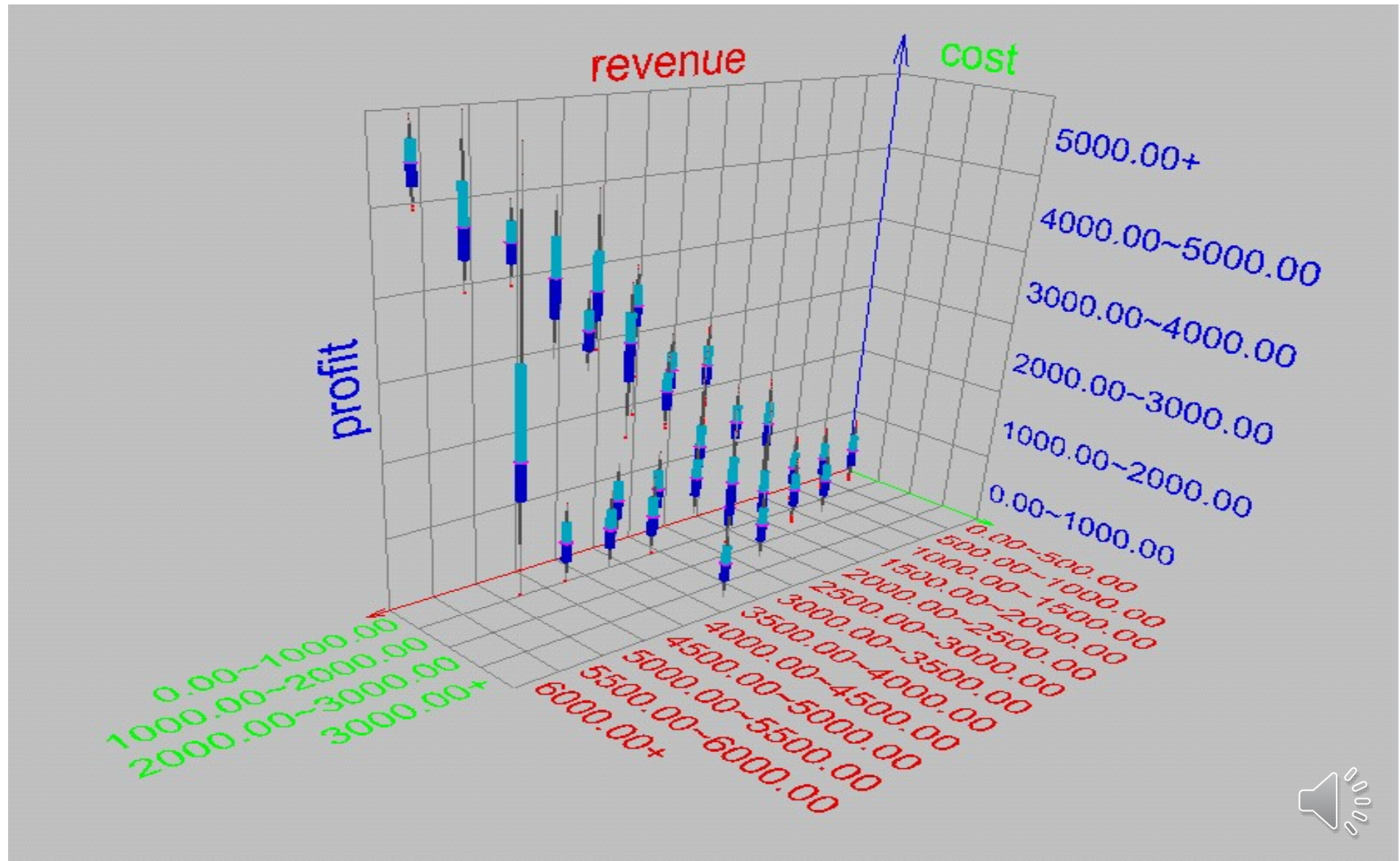
- Minimum, Q1, Median, Q3, Maximum

- **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

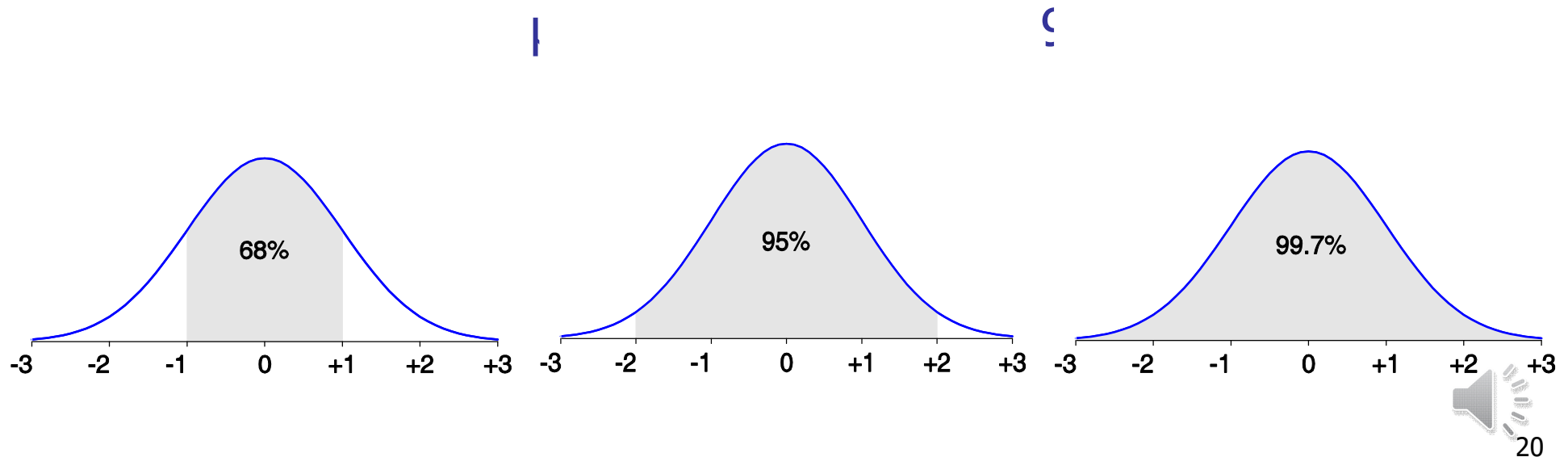


Visualization of Data Dispersion: 3-D Boxplots



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



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