MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY (MUST), MIRPUR
DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DATA MINING

BCS-3605

Lecture 06

Dr Yasir Mehmood

(Assistant Professor)

Agenda of Today's Lecture

- ***Types of Attributes***
- ***Types of Attributes: Qualitative***
- ***Types of Attributes: Quantitative***
- ***Basic Statistical Descriptions of Data***
- ***Measuring the Central Tendency***
- ***Symmetric vs. Skewed Data***

Types of Attributes

1. Qualitative
    - Nominal
    - Binary
    - Ordinal

2. Quantitative
    - Numeric

Types of Attributes: Qualitative

- **Nominal:** relating to names of things, category, state, code, without any meaningful order and qualitative, also called categorical
  - Hair color = {*black*, *blond*, *brown*, *grey*, *red*, *white*}
  - Marital status ={*single*, *married*, *divorced*, *widows* }
  - Occupation ={*teacher*, *dentist*, *programmer*, *farmer*}
  - ID numbers, zip codes
  - No mean and median, only mode (the most occurring value)
- **Binary:** Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally valuable without any weight
    - 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)

Types of Attributes: Qualitative

- **Ordinal:** Values have a meaningful order (ranking) but magnitude between successive values is not known
    - Size = {*small, medium, large*},
    - Grade =*{A+, A, B+, B}*
    - Professional ranking =*{Assistant, Associate, Professor}*
    - Survey rating = {*0: very dissatisfied,1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied*}
- Subjective assessment of quantities
- No mean, only median and mode can be found

Types of Attributes: Quantitative

- **Numeric**: quantitative, represented in integer or real values
  - **Interval-scaled**: Measured on a scale of equal-sized units, values have order, can compare and quantify the difference
    - E.g., temperature of 20◦C is five degrees higher than a temperature of 15◦C calendar dates, the years 2002 and 2010 are eight years apart
  - Mean, median, mode
  - No true zero-point, neither 0◦C nor 0◦F indicates “no temperature, Similarly, there is no true zero-point for calendar dates
  - **Ratio-scaled:** a value is a multiple (or ratio) of another value, values have order, can compare and quantify the difference, inherent zero-point,
    - *count as years of experience, length and width of a house, monetary quantities*
  - mean, median, and mode.

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of numeric values, may or may not be integer
    - E.g., hair_color, smoker, medical_test, drink_size
    - E.g, customer_ID, zip codes, profession, or the set of words in a collection of documents
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - e.g., temperature, height, or weight
  - Measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Basic Statistical Descriptions of Data

- **Motivation**

  Used to identify properties of the data and highlight which data values should be treated as noise or outliers

- **Statistical descriptions includes:**
  - Central tendency
    - Where most of value falls? location of middle or center of a data distribution. *mean*, *median*, *mode*, *midrange*
  - Dispersion of the data
    - How are the data spread out? *range*, *quartiles*, *five-number summary* and *boxplots*, *variance* and *standard deviation*. Identifies outliers.
  - Graphic displays
    - *quantile plots*, *quantile–quantile plots*, *histograms*, and *scatter plots*.

Measuring the Central Tendency

- Mean (arithmetic)

Let $x_1, x_2, \ldots, x_N$ be the set of N observed values or *observations* for numeric attribute X.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58.$$

Measuring the Central Tendency: Cont’d

- Weighted mean: each value of x_i is associated with w_i
  - Reflects significance, importance of respective value

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values

Measuring the Central Tendency: Cont'd

- Trimmed mean: chopping extreme values (say 2%)

- Median: Middle value if odd number of values, or average of the middle two sorted numeric values. For ordinal values two middlemost values.

  e.g 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. the average of two middle values are

$$\frac{52+56}{2} = \frac{108}{2} = 54.$$

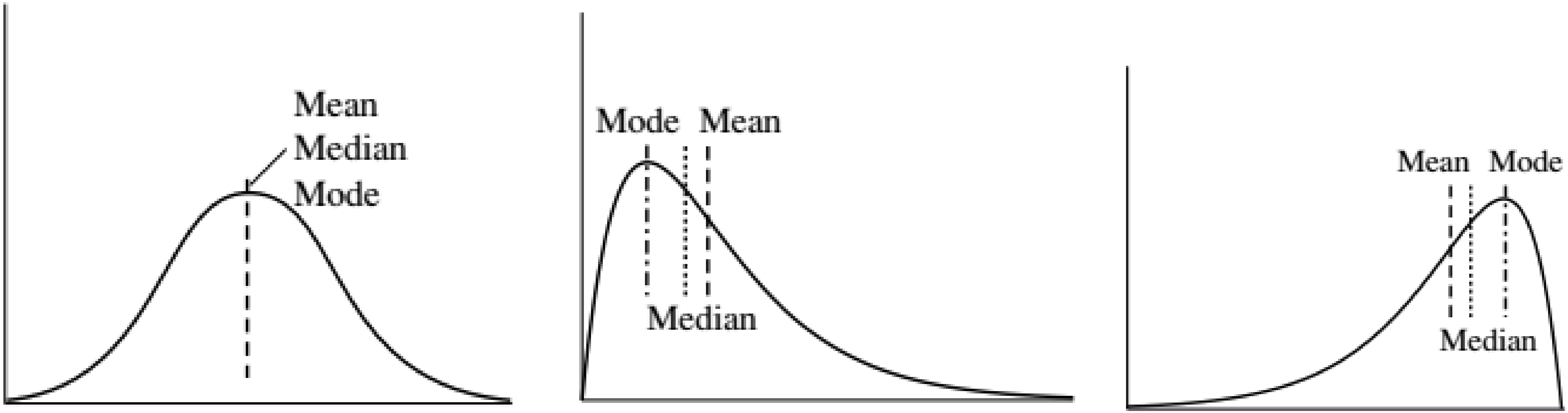
Measuring the Central Tendency: Cont'd

- Mode
  - Value that occurs most frequently in the data
  - Both for qualitative and quantitative attributes.
  - Unimodal, bimodal, trimodal, multimodal
  - For unimodal skewed data: $mean - mode = 3 \times (mean - median)$

- Midrange: average of largest and smallest values

Symmetric vs. Skewed Data

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



(a) Symmetric data

(b) Positively skewed data

(c) Negatively skewed data

Reference

Data Mining Concepts and Techniques Third Edition

2.1 Data Objects and Attribute Types

2.2 Basic Statistical Descriptions of Data

THANKS