# 1 Basic Statistical Descriptions of Data

## 1.1 Measures of Central Tendency

### 1.1.1 Ungrouped Data

- **Mean:** The average value. For a set of $n$ values $\{x_1, x_2, \ldots, x_n\}$:

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

**Example:** For the data set $\{2, 3, 3, 5, 7, 10\}$:

$$\bar{x} = \frac{2+3+3+5+7+10}{6} = \frac{30}{6} = 5$$

- **Median:** The middle value in an ordered data set.
  - If $n$ is odd, the median is the value at position $(n+1)/2$.
  - If $n$ is even, it's the average of the two middle values.

**Example (Odd):** For $\{2, 3, \mathbf{4}, 6, 8\}$, the median is 4.
**Example (Even):** For $\{2, 3, \mathbf{3}, \mathbf{5}, 7, 10\}$, the median is $\frac{3+5}{2} = 4$.

- **Mode:** The most frequently occurring value.
**Example:** In the data set $\{2, 3, \mathbf{3}, 5, 7, 10\}$, the mode is 3.

### 1.1.2 Grouped Data

For data organized into frequency distribution tables.

- **Mean of Grouped Data:**

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

where $k$ is the number of classes, $f_i$ is the frequency of the $i$-th class, $m_i$ is the midpoint of the $i$-th class, and $n = \sum_{i=1}^{k} f_i$ is the total number of data points.

- **Median of Grouped Data:**

$$\text{Median} = L + \left( \frac{\frac{n}{2} - F}{f} \right) w$$

where:
  - $L$ = lower boundary of the median class.
  - $n$ = total frequency.
  - $F$ = cumulative frequency of the class preceding the median class.
  - $f$ = frequency of the median class.
  - $w$ = width of the median class.

- **Mode of Grouped Data:**

$$\text{Mode} = L + \left( \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) w$$

where:
  - $L$ = lower boundary of the modal class.
  - $f_m$ = frequency of the modal class.
  - $f_1$ = frequency of the class preceding the modal class.
  - $f_2$ = frequency of the class succeeding the modal class.
  - $w$ = width of the modal class.

| Score Range (Class) | Frequency ($f_i$) | Midpoint ($m_i$) | Cumulative Freq. |
|---|---|---|---|
| 41 - 50 | 4 | 45.5 | 4 |
| 51 - 60 | 7 | 55.5 | 11 |
| **61 - 70** | **15** | **65.5** | **26** |
| 71 - 80 | 12 | 75.5 | 38 |
| 81 - 90 | 8 | 85.5 | 46 |
| 91 - 100 | 4 | 95.5 | 50 |
| **Total** | $\boldsymbol{n = 50}$ | | |

**Example for Grouped Data:** Consider the following frequency distribution of exam scores for 50 students.

- **Mean Calculation:** $\sum f_i m_i = (4 \times 45.5) + \cdots + (4 \times 95.5) = 3525$

$$\bar{x} = \frac{3525}{50} = 70.5$$

- **Median Calculation:** The median position is $n/2 = 25$. This falls in the **61-70** class. $L = 60.5$, $n = 50$, $F = 11$, $f = 15$, $w = 10$.

$$\text{Median} = 60.5 + \left(\frac{25 - 11}{15}\right) \times 10 \approx 69.83$$

- **Mode Calculation:** The modal class (highest frequency) is **61-70**. $L = 60.5$, $f_m = 15$, $f_1 = 7$, $f_2 = 12$, $w = 10$.

$$\text{Mode} = 60.5 + \left(\frac{15 - 7}{(15 - 7) + (15 - 12)}\right) \times 10 \approx 67.77$$

## 1.2 Measures of Data Dispersion

- **Range:** Range $= \max(x) - \min(x)$
  **Example:** For $\{2, 3, 3, 5, 7, 10\}$, the range is $10 - 2 = 8$.
- **Interquartile Range (IQR):** $IQR = Q3 - Q1$
  **Example:** For the ordered data $\{2, 4, 5, 8, 10, 12, 15\}$,
  - Q1 is the median of the lower half $\{2, \mathbf{4}, 5\}$, so Q1 = 4.
  - Q2 (the overall median) is 8.
  - Q3 is the median of the upper half $\{10, \mathbf{12}, 15\}$, so Q3 = 12.
  - $IQR = 12 - 4 = 8$.
- **Variance ($\sigma^2$) and Standard Deviation ($\sigma$):** For a sample of size $n$:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad s = \sqrt{s^2}$$

  **Example:** For $\{2, 3, 3, 5, 7, 10\}$, the mean $\bar{x}$ is 5.

$$s^2 = \frac{(2-5)^2 + (3-5)^2 + (3-5)^2 + (5-5)^2 + (7-5)^2 + (10-5)^2}{6-1}$$

$$s^2 = \frac{(-3)^2 + (-2)^2 + (-2)^2 + 0^2 + 2^2 + 5^2}{5} = \frac{9 + 4 + 4 + 0 + 4 + 25}{5} = \frac{46}{5} = 9.2$$

  The standard deviation is $s = \sqrt{9.2} \approx 3.03$.

# 2 Data Visualization

- **Boxplot (Box-and-Whisker Plot):** Displays the five-number summary.
  **Example:** For the data $\{2, 4, 5, 8, 10, 12, 15\}$, the five-number summary is:
  - Minimum: 2
  - Q1: 4
  - Median (Q2): 8
  - Q3: 12
  - Maximum: 15

  These five values define the structure of the boxplot.

# 3 Proximity for Binary Data

For binary vectors, we use a **contingency table** based on matching attributes.

| | | **Object y** | | |
|---|---|---|---|---|
| | | 1 | 0 | Total |
| **Object x** | 1 | $q$ | $r$ | $q+r$ |
| | 0 | $s$ | $t$ | $s+t$ |
| | Total | $q+s$ | $r+t$ | $n$ |

- $q$: number of attributes where $\mathbf{x} = 1, \mathbf{y} = 1$
- $t$: number of attributes where $\mathbf{x} = 0, \mathbf{y} = 0$

**Simple Matching Coefficient (SMC)**

- For **symmetric** variables (0 and 1 have equal weight, e.g., gender).

$$\text{SMC} = \frac{q+t}{q+r+s+t}$$

**Jaccard Coefficient**

- For **asymmetric** variables (0-0 matches are ignored, e.g., presence of a disease).

$$J = \frac{q}{q+r+s}$$

## 3.1 Similarity Measures for Symmetric Binary Attributes

### 3.1.1 Simple Matching Coefficient (SMC)

For symmetric binary attributes where both states are equally important:

$$SMC(i,j) = \frac{q+t}{q+r+s+t} = \frac{q+t}{p}$$

**Interpretation:** Proportion of attributes where both objects match.

### 3.1.2 Example: Symmetric Binary Attributes

Consider two patients with symptoms (1 = present, 0 = absent):

Table 1: Symptom Data for Two Patients

| Patient | Fever | Cough | Headache | Nausea | Fatigue | Total |
|---|---|---|---|---|---|---|
| $i$ | 1 | 0 | 1 | 0 | 1 | |
| $j$ | 1 | 1 | 1 | 0 | 0 | |

Contingency table:

- $q = 2$ (Fever, Headache - both have 1)
- $r = 1$ (Fatigue - i has 1, j has 0)
- $s = 1$ (Cough - i has 0, j has 1)
- $t = 1$ (Nausea - both have 0)
- $p = 5$

$$SMC(i, j) = \frac{2+1}{5} = \frac{3}{5} = 0.6$$

## 3.2 Similarity Measures for Asymmetric Binary Attributes

### 3.2.1 Jaccard Coefficient

For asymmetric binary attributes where 1-1 matches are more important:

$$J(i, j) = \frac{q}{q + r + s}$$

**Interpretation:** Proportion of positive matches among attributes where at least one object has 1.

### 3.2.2 Example: Asymmetric Binary Attributes

Consider two customers and products they purchased (1 = purchased, 0 = not purchased):

Table 2: Purchase Data for Two Customers

| Customer | Product A | Product B | Product C | Product D | Product E | Total |
|---|---|---|---|---|---|---|
| $i$ | 1 | 0 | 1 | 0 | 0 | |
| $j$ | 1 | 1 | 0 | 0 | 0 | |

Contingency table:

- $q = 1$ (Product A - both purchased)
- $r = 1$ (Product C - i purchased, j didn't)
- $s = 1$ (Product B - i didn't purchase, j did)
- $t = 2$ (Products D, E - neither purchased)

$$J(i, j) = \frac{1}{1+1+1} = \frac{1}{3} \approx 0.333$$

# 4 Measuring Data Similarity and Dissimilarity

For data objects $i = (x_{i1}, \ldots, x_{ip})$ and $j = (x_{j1}, \ldots, x_{jp})$. Let's use two data points, $i = (2, 2)$ and $j = (5, 6)$, for the following examples.

## 4.1 Distance Measures for Numeric Data

- **Euclidean Distance (L2 norm):**

$$d(i, j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

**Example:**

$$d(i, j) = \sqrt{(2-5)^2 + (2-6)^2} = \sqrt{(-3)^2 + (-4)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

- **Manhattan Distance (L1 norm):**

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

**Example:**

$$d(i,j) = |2-5| + |2-6| = |-3| + |-4| = 3 + 4 = 7$$

- **Minkowski Distance:** A generalization where $h$ is a positive integer.

$$d(i,j) = \left( \sum_{k=1}^{p} |x_{ik} - x_{jk}|^h \right)^{1/h}$$

**Example:** For $h = 3$:

$$d(i,j) = \left(|2-5|^3 + |2-6|^3\right)^{1/3} = \left(3^3 + 4^3\right)^{1/3} = (27+64)^{1/3} = \sqrt[3]{91} \approx 4.498$$

## 4.2 Cosine Similarity

Measures the cosine of the angle between two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$.

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}||\ ||\mathbf{y}||} = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2}\sqrt{\sum_{i=1}^{p} y_i^2}}$$

**Example:** Let vector $\mathbf{x} = (3, 4)$ and vector $\mathbf{y} = (2, 1)$.

- Dot product: $\mathbf{x} \cdot \mathbf{y} = (3 \times 2) + (4 \times 1) = 6 + 4 = 10$.
- Magnitudes: $||\mathbf{x}|| = \sqrt{3^2 + 4^2} = \sqrt{9+16} = \sqrt{25} = 5$.
- $||\mathbf{y}|| = \sqrt{2^2 + 1^2} = \sqrt{4+1} = \sqrt{5}$.
- Cosine Similarity:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{10}{5\sqrt{5}} = \frac{2}{\sqrt{5}} \approx 0.894$$

Since the value is close to 1, the vectors point in a similar direction.

## 4.3 3D Data Examples

Let's extend our analysis to three-dimensional data points. Consider two data points in 3D space: $i = (1, 3, 2)$ and $j = (4, 7, 5)$.

- **Euclidean Distance (L2 norm) in 3D:**

$$d(i,j) = \sqrt{(1-4)^2 + (3-7)^2 + (2-5)^2} = \sqrt{(-3)^2 + (-4)^2 + (-3)^2} = \sqrt{9+16+9} = \sqrt{34} \approx 5.831$$

- **Manhattan Distance (L1 norm) in 3D:**

$$d(i,j) = |1-4| + |3-7| + |2-5| = |-3| + |-4| + |-3| = 3 + 4 + 3 = 10$$

- **Minkowski Distance in 3D:** For $h = 3$:

$$d(i,j) = \left(|1-4|^3 + |3-7|^3 + |2-5|^3\right)^{1/3} = \left(3^3 + 4^3 + 3^3\right)^{1/3} = (27+64+27)^{1/3} = \sqrt[3]{118} \approx 4.905$$

- **Cosine Similarity in 3D:** Let vector $\mathbf{x} = (1, 3, 2)$ and vector $\mathbf{y} = (4, 7, 5)$.
  - Dot product: $\mathbf{x} \cdot \mathbf{y} = (1 \times 4) + (3 \times 7) + (2 \times 5) = 4 + 21 + 10 = 35$
  - Magnitudes: $||\mathbf{x}|| = \sqrt{1^2 + 3^2 + 2^2} = \sqrt{1+9+4} = \sqrt{14} \approx 3.742$
  - $||\mathbf{y}|| = \sqrt{4^2 + 7^2 + 5^2} = \sqrt{16+49+25} = \sqrt{90} \approx 9.487$
  - Cosine Similarity:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{35}{3.742 \times 9.487} \approx \frac{35}{35.5} \approx 0.986$$

This very high value (close to 1) indicates the vectors point in almost the same direction in 3D space.

## 4.4 Comparison of Measures

The 3D examples demonstrate how these distance measures scale with additional dimensions:

- Euclidean distance considers the straight-line distance in multidimensional space
- Manhattan distance sums absolute differences along each dimension
- Minkowski distance provides a flexible framework that can emphasize larger differences when $h > 1$
- Cosine similarity remains robust to the magnitude of vectors, focusing only on their directional relationship

## 4.5 Cosine Similarity Formula

For two vectors $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity is defined as:

$$\text{cosine}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

## 4.6 Text Vectorization

To apply cosine similarity to sentences:

1. Create a vocabulary from all unique words in both sentences
2. Represent each sentence as a vector in this vocabulary space
3. Use TF (Term Frequency) or TF-IDF weights

## 4.7 Example 1: Simple Sentence Similarity

## 4.8 Given Sentences

- Sentence 1: "I love machine learning"
- Sentence 2: "I love deep learning"

## 4.9 Step 1: Create Vocabulary

Unique words: {I, love, machine, deep, learning}
Vocabulary size: 5 dimensions

## 4.10 Step 2: Vector Representation

Using binary representation (1 if word present, 0 otherwise):

Table 3: Binary Vector Representation

| | I | love | machine | deep | learning |
|---|---|---|---|---|---|
| Sentence 1 ($\mathbf{A}$) | 1 | 1 | 1 | 0 | 1 |
| Sentence 2 ($\mathbf{B}$) | 1 | 1 | 0 | 1 | 1 |

$\mathbf{A} = [1, 1, 1, 0, 1]$, $\mathbf{B} = [1, 1, 0, 1, 1]$

## 4.11 Step 3: Calculate Dot Product

$$\mathbf{A} \cdot \mathbf{B} = (1 \times 1) + (1 \times 1) + (1 \times 0) + (0 \times 1) + (1 \times 1) = 1 + 1 + 0 + 0 + 1 = 3$$

## 4.12 Step 4: Calculate Magnitudes

$$\|\mathbf{A}\| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 1^2} = \sqrt{1 + 1 + 1 + 0 + 1} = \sqrt{4} = 2$$
$$\|\mathbf{B}\| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{1 + 1 + 0 + 1 + 1} = \sqrt{4} = 2$$

## 4.13 Step 5: Compute Cosine Similarity

$$\text{cosine}(\mathbf{A}, \mathbf{B}) = \frac{3}{2 \times 2} = \frac{3}{4} = 0.75$$

## 4.14 Example 2: Using Term Frequency (TF)

## 4.15 Given Sentences

- Sentence 1: "artificial intelligence is the future of technology"
- Sentence 2: "machine learning is part of artificial intelligence"

## 4.16 Step 1: Create Vocabulary

Unique words: {artificial, intelligence, is, the, future, of, technology, machine, learning, part}

## 4.17 Step 2: Binary Vector Representation

Table 4: Binary Vector Representation

| | artificial | intelligence | is | the | future | of | technology | machine | learning | part |
|---|---|---|---|---|---|---|---|---|---|---|
| S1 (**A**) | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| S2 (**B**) | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |

$\mathbf{A} = [1, 1, 1, 1, 1, 1, 1, 0, 0, 0]$, $\mathbf{B} = [1, 1, 1, 0, 0, 1, 0, 1, 1, 1]$

## 4.18 Step 3: Calculate Dot Product

$$\begin{aligned}\mathbf{A} \cdot \mathbf{B} =& (1 \times 1) + (1 \times 1) + (1 \times 1) + (1 \times 0) + (1 \times 0) + \\ & (1 \times 1) + (1 \times 0) + (0 \times 1) + (0 \times 1) + (0 \times 1) \\ =& 1 + 1 + 1 + 0 + 0 + 1 + 0 + 0 + 0 + 0 = 4\end{aligned}$$

## 4.19 Step 4: Calculate Magnitudes

$$\|\mathbf{A}\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2} = \sqrt{7} \approx 2.646$$
$$\|\mathbf{B}\| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2 + 1^2} = \sqrt{7} \approx 2.646$$

## 4.20 Step 5: Compute Cosine Similarity

$$\text{cosine}(\mathbf{A}, \mathbf{B}) = \frac{4}{2.646 \times 2.646} = \frac{4}{7} \approx 0.571$$

### 4.20.1 TF-IDF Weighting

For better results, use TF-IDF weights:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

where $\text{IDF}(t) = \log \frac{N}{n_t}$, with $N$ being total documents and $n_t$ documents containing term $t$.

# 5 Given Data

The age data in ascending order: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Total number of data points: $n = 27$

## 5.1 Binning Methods

## 5.2 Equal-Frequency (Equal-Depth) Binning

Since we have 27 data points, we can create 3 bins with 9 elements each:

- **Bin 1:** 13, 15, 16, 16, 19, 20, 20, 21, 22
- **Bin 2:** 22, 25, 25, 25, 25, 30, 33, 33, 35
- **Bin 3:** 35, 35, 35, 36, 40, 45, 46, 52, 70

## 5.3 Smoothing by Bin Means

Table 5: Smoothing by Bin Means

| Bin | Original Data | Smoothed Data |
|---|---|---|
| Bin 1 | 13, 15, 16, 16, 19, 20, 20, 21, 22 | 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0 |
| Bin 2 | 22, 25, 25, 25, 25, 30, 33, 33, 35 | 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1 |
| Bin 3 | 35, 35, 35, 36, 40, 45, 46, 52, 70 | 43.8, 43.8, 43.8, 43.8, 43.8, 43.8, 43.8, 43.8, 43.8 |

**Calculations:**

- Bin 1 mean: $\frac{13+15+16+16+19+20+20+21+22}{9} = \frac{162}{9} = 18.0$
- Bin 2 mean: $\frac{22+25+25+25+25+30+33+33+35}{9} = \frac{253}{9} = 28.1$
- Bin 3 mean: $\frac{35+35+35+36+40+45+46+52+70}{9} = \frac{394}{9} = 43.8$

## 5.4 Smoothing by Bin Boundaries

Table 6: Smoothing by Bin Boundaries

| Bin | Original Data | Smoothed Data |
|---|---|---|
| Bin 1 | 13, 15, 16, 16, 19, 20, 20, 21, 22 | 13, 13, 13, 13, 22, 22, 22, 22, 22 |
| Bin 2 | 22, 25, 25, 25, 25, 30, 33, 33, 35 | 22, 22, 22, 22, 35, 35, 35, 35, 35 |
| Bin 3 | 35, 35, 35, 36, 40, 45, 46, 52, 70 | 35, 35, 35, 35, 70, 70, 70, 70, 70 |

**Boundary Rules:**

- For each value, replace with closest boundary (min or max of the bin)
- Bin 1 boundaries: min = 13, max = 22
- Bin 2 boundaries: min = 22, max = 35
- Bin 3 boundaries: min = 35, max = 70

## 5.5 Smoothing by Bin Median

**Calculations:**

- Bin 1 median: 5th element (19) from sorted data: 13, 15, 16, 16, **19**, 20, 20, 21, 22
- Bin 2 median: 5th element (25) from sorted data: 22, 25, 25, 25, **25**, 30, 33, 33, 35
- Bin 3 median: 5th element (40) from sorted data: 35, 35, 35, 36, **40**, 45, 46, 52, 70

Table 7: Smoothing by Bin Median

| Bin | Original Data | Smoothed Data |
|---|---|---|
| Bin 1 | 13, 15, 16, 16, 19, 20, 20, 21, 22 | 19, 19, 19, 19, 19, 19, 19, 19, 19 |
| Bin 2 | 22, 25, 25, 25, 25, 30, 33, 33, 35 | 25, 25, 25, 25, 25, 25, 25, 25, 25 |
| Bin 3 | 35, 35, 35, 36, 40, 45, 46, 52, 70 | 40, 40, 40, 40, 40, 40, 40, 40, 40 |

### 5.6 Comparison of Methods

Table 8: Comparison of Smoothing Methods

| Method | Characteristics |
|---|---|
| Bin Means | Replaces all values in bin with mean, preserves some statistical properties but loses individual variations |
| Bin Boundaries | Replaces values with closest boundary, preserves extreme values but creates more discrete distribution |
| Bin Median | Replaces all values with median, robust to outliers and preserves central tendency |

All three smoothing methods reduce the noise in the data by grouping values into bins and replacing them with representative values. The choice of method depends on the specific data mining task and the importance of preserving certain data characteristics.

## 6 Normalization Methods

### 6.1 Given Data

The original data: 200, 300, 400, 600, 1000

### 6.2 Min-Max Normalization

Given: new_min $= 0$, new_max $= 1$

Formula:

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max} - \text{new\_min}) + \text{new\_min}$$

Where:

- $\min_A = 200$
- $\max_A = 1000$
- new_max $-$ new_min $= 1 - 0 = 1$

Calculations:

$$200' = \frac{200 - 200}{1000 - 200} \times 1 + 0 = \frac{0}{800} = 0.000$$
$$300' = \frac{300 - 200}{1000 - 200} \times 1 + 0 = \frac{100}{800} = 0.125$$
$$400' = \frac{400 - 200}{1000 - 200} \times 1 + 0 = \frac{200}{800} = 0.250$$
$$600' = \frac{600 - 200}{1000 - 200} \times 1 + 0 = \frac{400}{800} = 0.500$$
$$1000' = \frac{1000 - 200}{1000 - 200} \times 1 + 0 = \frac{800}{800} = 1.000$$

**Result:** 0.000, 0.125, 0.250, 0.500, 1.000

## 6.3 Z-Score Normalization

Formula:

$$v' = \frac{v - \mu}{\sigma}$$

Where:

- $\mu$ = mean = $\frac{200+300+400+600+1000}{5} = \frac{2500}{5} = 500$
- $\sigma$ = standard deviation

Calculate standard deviation:

$$\begin{aligned}
\sigma &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \\
&= \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}} \\
&= \sqrt{\frac{(-300)^2 + (-200)^2 + (-100)^2 + (100)^2 + (500)^2}{5}} \\
&= \sqrt{\frac{90000 + 40000 + 10000 + 10000 + 250000}{5}} \\
&= \sqrt{\frac{400000}{5}} = \sqrt{80000} = 282.843
\end{aligned}$$

Calculations:

$$\begin{aligned}
200' &= \frac{200 - 500}{282.843} = \frac{-300}{282.843} = -1.061 \\
300' &= \frac{300 - 500}{282.843} = \frac{-200}{282.843} = -0.707 \\
400' &= \frac{400 - 500}{282.843} = \frac{-100}{282.843} = -0.354 \\
600' &= \frac{600 - 500}{282.843} = \frac{100}{282.843} = 0.354 \\
1000' &= \frac{1000 - 500}{282.843} = \frac{500}{282.843} = 1.768
\end{aligned}$$

**Result:** -1.061, -0.707, -0.354, 0.354, 1.768

## 6.4 Z-Score Normalization using Mean Absolute Deviation

Formula:

$$v' = \frac{v - \mu}{MAD}$$

Where:

- $\mu = 500$ (same as above)
- $MAD$ = mean absolute deviation

Calculate MAD:

$$\begin{aligned}
MAD &= \frac{\sum_{i=1}^{n} |x_i - \mu|}{n} \\
&= \frac{|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|}{5} \\
&= \frac{300 + 200 + 100 + 100 + 500}{5} = \frac{1200}{5} = 240
\end{aligned}$$

Calculations:

$$200' = \frac{200 - 500}{240} = \frac{-300}{240} = -1.250$$
$$300' = \frac{300 - 500}{240} = \frac{-200}{240} = -0.833$$
$$400' = \frac{400 - 500}{240} = \frac{-100}{240} = -0.417$$
$$600' = \frac{600 - 500}{240} = \frac{100}{240} = 0.417$$
$$1000' = \frac{1000 - 500}{240} = \frac{500}{240} = 2.083$$

**Result:** -1.250, -0.833, -0.417, 0.417, 2.083

## 6.5 Normalization by Decimal Scaling

Formula:

$$v' = \frac{v}{10^j}$$

Where $j$ is the smallest integer such that $\max(|v'|) < 1$
Find $j$:

- Maximum absolute value in data: 1000
- We need $10^j > 1000$
- $10^3 = 1000$ (not sufficient since we need $< 1$)
- $10^4 = 10000 > 1000$ (sufficient)

So $j = 4$
Calculations:

$$200' = \frac{200}{10000} = 0.0200$$
$$300' = \frac{300}{10000} = 0.0300$$
$$400' = \frac{400}{10000} = 0.0400$$
$$600' = \frac{600}{10000} = 0.0600$$
$$1000' = \frac{1000}{10000} = 0.1000$$

**Result:** 0.0200, 0.0300, 0.0400, 0.0600, 0.1000

## 6.6 Summary of Results

Table 9: Comparison of Normalization Methods

| **Method** | **200** | **300** | **400** | **600** | **1000** |
|---|---|---|---|---|---|
| Original Data | 200 | 300 | 400 | 600 | 1000 |
| Min-Max [0,1] | 0.000 | 0.125 | 0.250 | 0.500 | 1.000 |
| Z-Score | -1.061 | -0.707 | -0.354 | 0.354 | 1.768 |
| Z-Score (MAD) | -1.250 | -0.833 | -0.417 | 0.417 | 2.083 |
| Decimal Scaling | 0.0200 | 0.0300 | 0.0400 | 0.0600 | 0.1000 |

### 6.7 Characteristics of Each Method

- **Min-Max:** Preserves relationships among original data values, bounded between 0 and 1
- **Z-Score:** Results in mean = 0 and standard deviation = 1, good for outlier detection
- **Z-Score (MAD):** More robust to outliers than standard z-score, uses mean absolute deviation
- **Decimal Scaling:** Simple method that preserves data relationships by moving decimal point

## 7 Discretization Methods

### 7.1 Equal-Width Binning

Divides the range of values into $k$ intervals of equal width.

#### 7.1.1 Algorithm

1. Find minimum and maximum values: $\min_A$, $\max_A$
2. Calculate width: $w = \frac{\max_A - \min_A}{k}$
3. Create intervals: $[\min_A + (i-1)w, \min_A + iw)$ for $i = 1, 2, \ldots, k$

#### 7.1.2 Example

Given data: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

$k = 3$, min = 13, max = 70, $w = \frac{70-13}{3} = 19$

Table 10: Equal-Width Binning

| Bin | Range | Values |
|---|---|---|
| 1 | [13, 32) | 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30 |
| 2 | [32, 51) | 33, 33, 35, 35, 35, 35, 36, 40, 45, 46 |
| 3 | [51, 70] | 52, 70 |

### 7.2 Equal-Frequency Binning

Divides data into $k$ intervals, each containing approximately the same number of observations.

#### 7.2.1 Algorithm

1. Sort data in ascending order
2. Calculate number of observations per bin: $n_{\text{bin}} = \frac{n}{k}$
3. Create intervals containing $n_{\text{bin}}$ observations each

#### 7.2.2 Example

Same data with $k = 3$, $n = 27$, $n_{\text{bin}} = 9$

Table 11: Equal-Frequency Binning

| Bin | Range | Values |
|---|---|---|
| 1 | [13, 22] | 13, 15, 16, 16, 19, 20, 20, 21, 22 |
| 2 | [22, 35] | 22, 25, 25, 25, 25, 30, 33, 33, 35 |
| 3 | [35, 70] | 35, 35, 35, 35, 36, 40, 45, 46, 52, 70 |

# 8 Discretization with Smoothing

## 8.1 Bin Means Smoothing

Replace all values in a bin with the mean of the bin.

### 8.1.1 Example

Using equal-frequency bins from previous example:

Table 12: Bin Means Smoothing

| Bin | Original Values | Smoothed Values |
|---|---|---|
| 1 | 13, 15, 16, 16, 19, 20, 20, 21, 22 | 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0, 18.0 |
| 2 | 22, 25, 25, 25, 25, 30, 33, 33, 35 | 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1, 28.1 |
| 3 | 35, 35, 35, 35, 36, 40, 45, 46, 52, 70 | 42.9, 42.9, 42.9, 42.9, 42.9, 42.9, 42.9, 42.9, 42.9, 42.9 |

## 8.2 Bin Boundaries Smoothing

Replace each value with the closest boundary value (min or max of the bin).

### 8.2.1 Example

Table 13: Bin Boundaries Smoothing

| Bin | Original Values | Smoothed Values |
|---|---|---|
| 1 | 13, 15, 16, 16, 19, 20, 20, 21, 22 | 13, 13, 13, 13, 22, 22, 22, 22, 22 |
| 2 | 22, 25, 25, 25, 25, 30, 33, 33, 35 | 22, 22, 22, 22, 35, 35, 35, 35, 35 |
| 3 | 35, 35, 35, 35, 36, 40, 45, 46, 52, 70 | 35, 35, 35, 35, 70, 70, 70, 70, 70, 70 |

## 8.3 Comparison of Methods

Table 14: Comparison of Discretization Methods

| Method | Advantages | Disadvantages |
|---|---|---|
| Equal-Width | Simple, fast | Sensitive to outliers, uneven distribution |
| Equal-Frequency | Handles outliers better, even distribution | May put same values in different bins |
| ChiMerge | Supervised, considers class information | Computationally expensive |
| Entropy-Based | Maximizes information gain, supervised | Complex, may overfit |

## 8.4 Practical Considerations

## 8.5 Choosing the Number of Bins

- **Sturges' Rule:** $k = 1 + \log_2 n$
- **Square-root Choice:** $k = \sqrt{n}$
- **Rice Rule:** $k = 2n^{1/3}$

For our example with $n = 27$:

- Sturges: $k = 1 + \log_2 27 \approx 1 + 4.75 = 5.75 \to 6$
- Square-root: $k = \sqrt{27} \approx 5.19 \to 5$
- Rice: $k = 2 \times 27^{1/3} \approx 2 \times 3 = 6$

### 8.6 Handling Categorical Labels

After discretization, bins can be labeled:

- Numeric: 1, 2, 3, ...
- Descriptive: Low, Medium, High
- Range-based: 13-22, 23-35, 36-70

Discretization is a crucial preprocessing step in data mining that enables the use of categorical algorithms on continuous data. The choice of discretization method depends on the specific dataset, the presence of class labels, and the requirements of the subsequent analysis.

## 9 Introduction to Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms correlated variables into a set of uncorrelated variables called principal components. The first principal component accounts for the largest possible variance in the data, and each succeeding component accounts for the highest remaining variance under the constraint of being orthogonal to previous components.

### 9.1 Mathematical Foundations

### 9.2 Covariance Matrix

Given a data matrix $X$ with $n$ observations and $p$ variables:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

The covariance matrix $\Sigma$ is calculated as:

$$\Sigma = \frac{1}{n-1} X^T X \quad \text{(for mean-centered data)}$$

Or explicitly:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

where $\sigma_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$

### 9.3 Eigenvalue Decomposition

PCA involves finding the eigenvalues and eigenvectors of the covariance matrix:

$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

where:

- $\lambda_i$ are eigenvalues (variances of principal components)
- $\mathbf{v}_i$ are eigenvectors (principal component directions)
- Eigenvalues are ordered: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

### 9.4 Principal Components

The principal components are linear combinations of original variables:

$$PC_i = X\mathbf{v}_i$$

The proportion of variance explained by the $i$-th principal component is:

$$\text{Proportion}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

## 10 Step-by-Step PCA Example

### 10.1 Given Data

Consider the following 2D dataset with 5 observations:

$$X = \begin{bmatrix} 2 & 3 \\ 4 & 5 \\ 6 & 7 \\ 8 & 9 \\ 10 & 11 \end{bmatrix}$$

### 10.2 Step 1: Mean Centering

Calculate means: $\bar{x}_1 = 6$, $\bar{x}_2 = 7$

Mean-centered data:

$$X_c = \begin{bmatrix} 2-6 & 3-7 \\ 4-6 & 5-7 \\ 6-6 & 7-7 \\ 8-6 & 9-7 \\ 10-6 & 11-7 \end{bmatrix} = \begin{bmatrix} -4 & -4 \\ -2 & -2 \\ 0 & 0 \\ 2 & 2 \\ 4 & 4 \end{bmatrix}$$

### 10.3 Step 2: Covariance Matrix

$$\Sigma = \frac{1}{n-1} X_c^T X_c = \frac{1}{4} \begin{bmatrix} -4 & -2 & 0 & 2 & 4 \\ -4 & -2 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} -4 & -4 \\ -2 & -2 \\ 0 & 0 \\ 2 & 2 \\ 4 & 4 \end{bmatrix}$$

$$\Sigma = \frac{1}{4} \begin{bmatrix} 40 & 40 \\ 40 & 40 \end{bmatrix} = \begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix}$$

### 10.4 Step 3: Eigenvalue Decomposition

Solve $|\Sigma - \lambda I| = 0$:

$$\begin{vmatrix} 10-\lambda & 10 \\ 10 & 10-\lambda \end{vmatrix} = 0$$

$$(10-\lambda)^2 - 100 = 0$$

$$\lambda^2 - 20\lambda = 0$$

$$\lambda(\lambda - 20) = 0$$

Eigenvalues: $\lambda_1 = 20$, $\lambda_2 = 0$

## 10.5 Step 4: Eigenvectors

For $\lambda_1 = 20$:

$$\begin{bmatrix} 10-20 & 10 \\ 10 & 10-20 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -10 & 10 \\ 10 & -10 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$v_{11} = v_{12}$, so $\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

For $\lambda_2 = 0$:

$$\begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$v_{21} = -v_{22}$, so $\mathbf{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

## 10.6 Step 5: Principal Components

First principal component:

$$PC_1 = X_c \mathbf{v}_1 = \begin{bmatrix} -4 & -4 \\ -2 & -2 \\ 0 & 0 \\ 2 & 2 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -5.657 \\ -2.828 \\ 0 \\ 2.828 \\ 5.657 \end{bmatrix}$$

Second principal component:

$$PC_2 = X_c \mathbf{v}_2 = \begin{bmatrix} -4 & -4 \\ -2 & -2 \\ 0 & 0 \\ 2 & 2 \\ 4 & 4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

## 10.7 Step 6: Variance Explained

Total variance: $\lambda_1 + \lambda_2 = 20 + 0 = 20$

Proportion by PC1: $\frac{20}{20} = 100\%$

Proportion by PC2: $\frac{0}{20} = 0\%$

## 10.8 General PCA Algorithm

1. **Standardize the data:** Center and scale variables

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

2. **Compute covariance matrix:** $\Sigma = \frac{1}{n-1} X^T X$
3. **Calculate eigenvalues and eigenvectors:** Solve $\Sigma \mathbf{v} = \lambda \mathbf{v}$
4. **Sort components:** Order by decreasing eigenvalues
5. **Select components:** Choose $k$ components that explain sufficient variance
6. **Transform data:** $Y = XV_k$ where $V_k$ contains first $k$ eigenvectors

### 10.9 Properties of PCA

### 10.10 Optimality Conditions

- Principal components are uncorrelated
- First PC has maximum variance
- PCs are orthogonal to each other
- Total variance is preserved: $\sum \lambda_i = \text{trace}(\Sigma)$

### 10.11 Dimensionality Reduction

The cumulative proportion of variance explained by first $k$ components:

$$R_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Common criteria for choosing $k$:

- Kaiser criterion: $\lambda_i > 1$ (for standardized data)
- Cumulative variance $> 70 - 90\%$
- Scree plot elbow

### 10.12 Applications

- Data visualization
- Noise reduction
- Feature extraction
- Collinearity removal
- Data compression

### 10.13 Limitations

- Linear assumptions
- Sensitivity to scaling
- Interpretation challenges
- Outlier sensitivity

## 11 Pearson's Correlation Coefficient

### 11.1 Definition

Pearson's product-moment correlation coefficient measures the linear relationship between two variables $X$ and $Y$. The formula is:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Alternatively, it can be written as:

$$r_{XY} = \frac{\text{cov}(X, Y)}{s_X s_Y}$$

where $s_X$ and $s_Y$ are the sample standard deviations.

## 11.2 Interpretation

- $r = +1$: Perfect positive correlation
- $0 < r < +1$: Positive correlation
- $r = 0$: No linear correlation
- $-1 < r < 0$: Negative correlation
- $r = -1$: Perfect negative correlation

## 11.3 Covariance

## 11.4 Definition

Covariance measures how two variables change together:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

## 11.5 Interpretation

- $\text{cov}(X, Y) > 0$: Positive relationship
- $\text{cov}(X, Y) = 0$: No linear relationship
- $\text{cov}(X, Y) < 0$: Negative relationship

## 11.6 Example Calculation

## 11.7 Given Data

Let's consider two attributes $X$ and $Y$ with the following data:

Table 15: Example Dataset

| Observation | $X$ | $Y$ |
|---|---|---|
| 1 | 2 | 5 |
| 2 | 4 | 7 |
| 3 | 6 | 10 |
| 4 | 8 | 12 |
| 5 | 10 | 15 |

## 11.8 Step 1: Calculate Means

$$\bar{x} = \frac{2 + 4 + 6 + 8 + 10}{5} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{5 + 7 + 10 + 12 + 15}{5} = \frac{49}{5} = 9.8$$

## 11.9 Step 2: Calculate Deviations and Products

## 11.10 Step 3: Calculate Covariance

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{50.0}{4} = 12.5$$

Table 16: Calculation Table

| $i$ | $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ |
|---|---|---|---|---|---|
| 1 | 2 | 5 | -4 | -4.8 | 19.2 |
| 2 | 4 | 7 | -2 | -2.8 | 5.6 |
| 3 | 6 | 10 | 0 | 0.2 | 0.0 |
| 4 | 8 | 12 | 2 | 2.2 | 4.4 |
| 5 | 10 | 15 | 4 | 5.2 | 20.8 |
| Sum | | | | | 50.0 |

## 11.11 Step 4: Calculate Standard Deviations

$$s_X = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{4}}$$
$$= \sqrt{\frac{16 + 4 + 0 + 4 + 16}{4}} = \sqrt{\frac{40}{4}} = \sqrt{10} \approx 3.162$$

$$s_Y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(-4.8)^2 + (-2.8)^2 + 0.2^2 + 2.2^2 + 5.2^2}{4}}$$
$$= \sqrt{\frac{23.04 + 7.84 + 0.04 + 4.84 + 27.04}{4}} = \sqrt{\frac{62.8}{4}} = \sqrt{15.7} \approx 3.962$$

## 11.12 Step 5: Calculate Pearson's Correlation Coefficient

$$r_{XY} = \frac{\text{cov}(X, Y)}{s_X s_Y} = \frac{12.5}{3.162 \times 3.962} = \frac{12.5}{12.53} \approx 0.998$$

## 11.13 Interpretation of Results

## 11.14 Covariance Interpretation

$$\text{cov}(X, Y) = 12.5 > 0$$

The positive covariance indicates that $X$ and $Y$ tend to increase together.

## 11.15 Correlation Interpretation

$$r_{XY} \approx 0.998$$

This value is very close to $+1$, indicating a very strong positive linear relationship between $X$ and $Y$.

## 11.16 Mathematical Properties

## 11.17 Properties of Pearson's Correlation

- Symmetric: $r_{XY} = r_{YX}$
- Range: $-1 \leq r_{XY} \leq 1$
- Scale invariant: $r_{aX+b,cY+d} = \text{sign}(a)\text{sign}(c)r_{XY}$
- Measures linear relationship only

## 11.18 Properties of Covariance

- Symmetric: $\text{cov}(X, Y) = \text{cov}(Y, X)$
- $\text{cov}(X, X) = \text{var}(X)$
- Bilinear: $\text{cov}(aX + b, cY + d) = ac \cdot \text{cov}(X, Y)$
- Dependent on scales of $X$ and $Y$

# 12 Relationship between Covariance and Correlation

The correlation coefficient is essentially a normalized version of covariance:

$$r_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X,Y)}{s_X s_Y}$$

While covariance indicates the direction of the relationship, correlation measures both the direction and strength of the linear relationship.

## 12.1 Significance Testing

## 12.2 Hypothesis Test for Correlation

To test if the correlation is statistically significant:

- $H_0 : \rho = 0$ (No correlation)
- $H_1 : \rho \neq 0$ (Significant correlation)

Test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

which follows a t-distribution with $n-2$ degrees of freedom.

## 12.3 Limitations

- Only measures linear relationships
- Sensitive to outliers
- Does not imply causation
- Can be misleading with non-linear relationships

Prepared By:
**Md. Atikuzzaman**
Lecturer
Department of Computer Science and Engineering
Green University of Bangladesh
Email: atik@cse.green.edu.bd