CSE 435: CHAPTER - 2

Md. Ataullha

Formula:

$$\text{Median} = L_1 + \left(\frac{\frac{N}{2} - F}{f_{\text{median}}} \right) \times \text{width}$$

Where:

- L_1 = Lower boundary of the median class
- N = Total number of data points (sum of frequencies)
- F = Cumulative frequency before the median class
- f_{median} = Frequency of the median class
- width = Class width (difference between the upper and lower boundaries of the class)

Class Interval	Frequency
10-20	5
20-30	8
30-40	12
40-50	10
50-60	5

Steps to Calculate Median:

1. **Find N**: The total number of data points is the sum of the frequencies.

$$N = 5 + 8 + 12 + 10 + 5 = 40$$

2. **Find $\frac{N}{2}$**: Half of the total number of data points.

$$\frac{N}{2} = \frac{40}{2} = 20$$

3. **Determine the median class**: The median class is the one where the cumulative frequency first exceeds $\frac{N}{2}$. We'll calculate the cumulative frequencies and find the class that includes the 20th data point.

Class Interval	Frequency	Cumulative Frequency
10-20	5	5
20-30	8	13
30-40	12	25

The **median class** is the **30-40** class, because the cumulative frequency of 25 exceeds 20 (i.e., it includes the 20th data point).

4. **Find** L_1: The lower boundary of the median class is **30**.
5. **Find** F: The cumulative frequency before the median class is **13** (for the class 20-30).
6. **Find** f_{median}: The frequency of the median class is **12**.
7. **Find the width**: The width of each class is **10** (since the class intervals are 10 units wide).

Now, plug the values into the formula:

$$\text{Median} = 30 + \left(\frac{20 - 13}{12}\right) \times 10$$

$$\text{Median} = 30 + \left(\frac{7}{12}\right) \times 10$$

$$\text{Median} = 30 + 5.83$$

$$\text{Median} = 35.83$$

So, the **median** is approximately **35.83**.

Five-Number Summary & Boxplot Visualization

The **Five-Number Summary** is a statistical summary that describes a dataset's distribution using the following values:

1. **Minimum (absolute)**: The smallest value in the dataset.
2. **Q1 (25th percentile)**: The value below which 25% of the data fall.
3. **Median (Q2, 50th percentile)**: The middle value that divides the dataset into two equal halves.
4. **Q3 (75th percentile)**: The value below which 75% of the data fall.
5. **Maximum (absolute)**: The largest value in the dataset.

Boxplot Visualization

A **Boxplot** visualizes the five-number summary with the following adjustments:

- **Whiskers** represent the range between the **minimum** and **maximum** within a certain limit (determined by the 1.5 × IQR rule).
- **Outliers** are any data points that fall outside the whisker range and are plotted separately.

Key Differences:

- The **Five-Number Summary** uses absolute min and max values from the dataset.
- A **Boxplot** adjusts the min and max based on the 1.5 × IQR rule. The whiskers end at the largest and smallest values within the whisker limit, and any points beyond are considered outliers.

Example:

For the dataset **[2, 4, 6, 8, 12, 14, 18, 2000]**:

- **Statistical Five-Number Summary**: (2, 5, 10, 16, 2000)
- **Boxplot**: (2, 5, 10, 16, 18) with 2000 as an outlier.

Conclusion:

- **Boxplots** do not always plot the absolute min and max values if outliers exist.
- They show the whisker range as min and max, excluding outliers.

For a simpler way to calculate the **correlation coefficient**, we can use the following formula for **Pearson's correlation coefficient**:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual data points in the x and y datasets, respectively.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y datasets, respectively.
- $\sum$ represents the summation.

Simplified Steps:

1. **Calculate the means** of x and y:

$$\bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}$$

2. **Calculate the deviations** of each x_i and y_i from their respective means ($x_i - \bar{x}$ and $y_i - \bar{y}$).
3. **Compute the sum of the products of deviations** $(x_i - \bar{x})(y_i - \bar{y})$.
4. **Compute the sum of squared deviations** for both x and y.
5. **Substitute** these sums into the correlation formula to find r.

Example: Let's calculate the correlation for this simple dataset:

x	y
2	8
4	12
6	14
8	18
10	20

Step 1: Calculate the means

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

$$\bar{y} = \frac{8 + 12 + 14 + 18 + 20}{5} = \frac{72}{5} = 14.4$$

Step 2: Calculate the deviations from the mean and their products

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
2	8	-4	-6.4	25.6	16	40.96
4	12	-2	-2.4	4.8	4	5.76
6	14	0	-0.4	0	0	0.16
8	18	2	3.6	7.2	4	12.96
10	20	4	5.6	22.4	16	31.36

Step 3: Sum of products and sum of squares

- **Sum of products of deviations**:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 25.6 + 4.8 + 0 + 7.2 + 22.4 = 60$$

- **Sum of squared deviations for x**:

$$\sum (x_i - \bar{x})^2 = 16 + 4 + 0 + 4 + 16 = 40$$

- **Sum of squared deviations for y**:

$$\sum (y_i - \bar{y})^2 = 40.96 + 5.76 + 0.16 + 12.96 + 31.36 = 91.2$$

Step 4: Substitute into the correlation formula

$$r = \frac{60}{\sqrt{40 \times 91.2}} = \frac{60}{\sqrt{3648}} = \frac{60}{60.34} \approx 0.995$$

Result:

The **correlation coefficient** r is approximately **0.995**, indicating a very strong positive linear relationship between x and y.

Feature	Histogram	Bar Chart
Data type	Continuous (numbers with ranges, e.g. height, age)	Categorical (distinct groups, e.g. fruits, countries)
X-axis	Intervals (bins of values)	Categories (labels)
Y-axis	Frequency (how many fall in each bin)	Count or value for each category
Bars	Touch each other (no gaps)	Separated by gaps
Purpose	Shows distribution of data	Compares categories
Example	Number of students in score ranges (60–70, 70–80...)	Number of students liking each subject (Math, Science, Art...)

`2, 5, 9, 3, 4, 8, 3`

1. **Order the data:**
   `2, 3, 3, 4, 5, 8, 9`
2. **Minimum:**
   The smallest number → **2**
3. **Maximum:**
   The largest number → **9**
4. **Median (Q2):**
   Since there are 7 numbers (odd count), the middle one is the 4th value: **4**
5. **Q1 (25th percentile):**
   Look at the lower half (below the median): `2, 3, 3`
   The middle of these is **3**
6. **Q3 (75th percentile):**
   Look at the upper half (above the median): `5, 8, 9`
   The middle of these is **8**, but note that when using percentile interpolation, many methods average between positions—giving **6.5** instead.

Theoretical Result

So depending on method:

- **Minimum:** 2
- **Q1:** 3 (discrete method) or 3.0 (percentile method)
- **Median:** 4
- **Q3:** 8 (discrete) or 6.5 (percentile interpolation method, which matches the boxplot)
- **Maximum:** 9

1. Proximity

- **Proximity** refers to how similar or dissimilar two objects are.
- It can be measured in two ways:
  - **Similarity**: higher value means objects are more alike.
  - **Dissimilarity**: higher value means objects are less alike.

Matrix representation:

- Rows and columns represent objects.
- Entries contain similarity or dissimilarity values.

Formula link:

$$\text{sim}(i, j) = 1 - d(i, j)$$

2. Binary Attributes

- Binary attributes take values **0/1** (e.g., Male/Female, Yes/No).
- **Symmetric Binary Attribute**: both values are equally important.
  Example: Gender.
- **Asymmetric Binary Attribute**: only one state (often "1") carries importance.
  Example: Medical test result (disease presence = 1).

Dissimilarity for Asymmetric Binary:

$$d(x, y) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

- M_{10}: number of attributes where $x = 1, y = 0$
- M_{01}: number of attributes where $x = 0, y = 1$
- M_{11}: number of attributes where both are 1

Similarity:

$$\text{sim}(x, y) = 1 - d(x, y)$$

- **SMC (Simple Matching Coefficient)** also used for symmetric case:

$$\text{SMC} = \frac{M_{11} + M_{00}}{M_{11} + M_{00} + M_{01} + M_{10}}$$

3. Nominal Attributes

- Nominal = categorical values (e.g., color: red, green, blue).
- Dissimilarity formula:

$$d(i, j) = \frac{p - m}{p}$$

- p: total number of attributes.
- m: number of matches.

4. Ordinal Attributes

- Attributes with a meaningful order, but differences are not exact (e.g., low, medium, high).
- Steps:
  1. Assign ranks: High = 1, Medium = 2, Low = 3.
  2. Normalize:

$$z = \frac{x-1}{N-1}$$

     where N = number of ranks.
- Example dissimilarity values:
  - High vs High = 0
  - High vs Low = 1
  - High vs Medium = 0.5

5. Numerical Attributes

Proximity for numerical data can be measured by distance functions:

1. **Euclidean Distance**:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

2. **Manhattan Distance**:

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

3. **Supremum (Chebyshev) Distance**:

$$d(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|)$$

- Also called **L∞ norm**.

6. Example: Supermarket Products

- Suppose 1000 products in supermarket.
- C1 = {sugar, coffee, tea, rice, egg}
- C2 = {sugar, coffee, bread, biscuit}

Dissimilarity:

$$d(C1, C2) = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

= $\frac{5}{1000} = 0.005$

7. Quantiles & Spread Measures

- **Range** = Max – Min
- **Variance**:

$$\sigma^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$$

- **Standard deviation** = $\sqrt{\text{Variance}}$
- **IQR (Interquartile Range)** = Q3 – Q1
- **Quantiles**: divide data into equal-sized intervals.
  - Quartiles → 4 parts
  - Percentiles → 100 parts

Central Point

A data object → **entity**

Attribute → **data field / feature**

Univariate

Multivariate

Nominal, binary, ordinal or numeric

Categorical

- ✕ Average / Mean
- ✕ Median
- ✓ Mode

Binary

- Symmetric (equivalent)
- Asymmetric (not equal, positive or negative)

Objective measures

Ordinal

- Meaningful order and ranking

Interval-scaled attributes

→ No true zero point exists

Ratio-scaled attributes

→ We can find multiply (money)

Mean, median, mode, midrange

→ *Dispersion of data*

Range, quartiles, IQR

Boxplots / five-number summary

Variance, standard deviation

Formulae:

Mean – Mode = 3 × (Mean – Median)

Midrange = (Min + Max) / 2

Central Point

A data object → **entity**
Attribute → **data field / feature**

Univariate
Multivariate

Nominal, binary, ordinal or numeric

Categorical

- ✕ Average / Mean
- ✕ Median
- ✓ Mode

Binary

- Symmetric (equivalent)
- Asymmetric (not equal, positive or negative)

Objective measures

Ordinal

- Meaningful order and ranking

Interval-scaled attributes

→ No true zero point exists

Ratio-scaled attributes

→ We can find multiply (money)

Mean, median, mode, midrange

→ *Dispersion of data*

Range, quartiles, IQR
Boxplots / five-number summary
Variance, standard deviation

Formulae:

Mean – Mode = 3 × (Mean – Median)
Midrange = (Min + Max) / 2

- **Interval-scaled** data lacks a true zero (like temperature in Celsius).
- **Ratio-scaled** data has a true zero and allows meaningful ratios (like money, weight, height).

  Then it lists statistical measures — mean, median, mode, range, IQR, variance, and standard deviation — plus two classic relationships:

1. **Empirical relationship** among mean, median, and mode.
2. **Midrange** as the midpoint between the smallest and largest values.

Symmetric

- Normal distribution
- Mean = Mode

  (Graph shows a bell curve centered at the mean)

Asymmetric

- Mode < Median < Mean
- Positively skewed

  (Graph skews to the right)

Boxplot concepts

Min (excluding outlier)

Q1

Q2 (Median)

Q3

Max (excluding outlier)

$Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR$ → defines the range of non-outlier data

5 Points in a boxplot:

- Min
- Q1
- Q2
- Q3
- Max

Boxplot / Whisker Plot

Min, Q1, Q2, Q3, Max

$Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR$ → *value included*

Example dataset:

18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22

Steps:

1. **Sorting the data:**
   15, 18, 20, 22, 25, 29, 32, 34, 38, 41, 43, 46, 54, 76
2. Since *n = even*,
   Median = (n/2 + n/2 + 1) / 2 = (7th + 8th) / 2
   Median (Q2) = (32 + 34) / 2 = 33
3. **Lower half (below median):**
   15, 18, 20, 22, 25, 29, 32
   → Q1 = 22
4. **Upper half (above median):**
   34, 38, 41, 43, 46, 54, 76
   → Q3 = 43

IQR (Interquartile Range):

IQR = Q3 – Q1 = 43 – 22 = **21**

Outlier boundaries:

$Q1–1.5 \times IQR, Q3 + 1.5 \times IQR$

= [–9.5, 74.5]

Interpretation:

Any data point *below –9.5* or *above 74.5* is an **outlier**.

Here, **76** is an outlier.

Values:

- Min = 15
- Q1 = 22
- Median (Q2) = 33
- Q3 = 43
- Max = 54
- Outlier ≈ 76

Variance formula:

$\sigma^2 = (\Sigma (x_i - \bar{x})^2) / n$

Cosine Similarity

$A \cdot B = |A|\,|B| \cos\theta$

$\therefore \cos\theta = (A \cdot B) / (\sqrt{A^2} \times \sqrt{B^2})$

- $\cos\theta = 1$ → vectors are **similar**
- $\cos\theta = 0$ → vectors are **dissimilar**

Range:

= (Max – Min)

Variance (σ²):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Standard Deviation (std dev):

$$std\,dev = \sqrt{\sigma^2} = \sqrt{var}$$

IQR (Interquartile Range):

$$IQR = Q3 - Q1$$

Quantiles:

Quantiles are **points taken at regular intervals of data.**

- **k-th quantile** divides the data into *q* equal parts.
- **q = number of divisions, k = specific position.**

$$q\text{-quantile} = \frac{k}{q}$$

Examples:

- 100 percentiles → divide data into 100 equal parts.
- 4 quartiles → divide data into 4 equal parts.

Data:

2, 5, 9, 3, 4, 8, 3

Quartile calculation:

Sorted data → 2, 3, 3, 4, 5, 8, 9

Q1 = (3 + 3) / 2 = **3**

Q2 = **4** (median)

Q3 = (8 + 5) / 2 = **6.5**

Measuring Similarities vs Dissimilarities

- Use a **matrix** to represent distances or similarities.
- **Proximity** measures how close or far objects are from each other.
  - Can represent **similarity** or **dissimilarity** between:
    - **Object vs Object** (rows)
    - **Object vs Attribute** (columns)

Data types and comparison:

- **Nominal** – categorical values
- **Binary** – two possible states:
  - **Symmetric** (equal importance)
  - **Asymmetric** (opposite or not equal importance)
- **Ordinal** – ranked data
  Example: 5 – 2 = 3 (difference between ranks)

Proximity
→ Refers to the closeness between **two objects**
→ Can be expressed as **similarity** or **dissimilarity**

Dissimilarity Matrix Example:

	A	B	C	D
A	0	x	x	x
B	x	0	x	x
C	x	x	0	x
D	x	x	x	0

Where:

- $d(A, B) = d(B, A)$
- $d(i, j)$ represents **dissimilarity** between objects *i* and *j*

Formulas:

- **Dissimilarity:** $d(i, j)$
- **Similarity:**

$$sim(i, j) = 1 - dissimilarity(i, j)$$

or

$$sim(i, j) = 1 - d(i, j)$$

Categorical (Nominal) Dissimilarity

Example:

{Red, Green, Blue}

Classes:

class1, class2, class3

Formula:

$$d(i, j) = \frac{p - m}{p}$$

Where:

- **p** = total number of attributes or features
- **m** = total number of matches

Example Table:

ID	Type-1	Type-2
1	10	A
2	30	B
3	10	A
4	20	C

Example Calculations:

1. $d(2,1) = \frac{2-0}{2} = 1$ → **Completely dissimilar**
2. $d(3,1) = \frac{2-2}{2} = 0$ → **Completely similar**

Interpretation:

If all attribute values match between two objects, **d = 0** (completely similar).

If none match, **d = 1** (completely dissimilar).

Binary Attributes

Definition:

Binary attributes have only two possible states — e.g.

- 1 = True / Male
- 0 = False / Female

Symmetric Binary Attribute
Both 0 and 1 are equally important (e.g., gender, yes/no).

Asymmetric Binary Attribute
1 indicates presence, and 0 indicates absence (e.g., disease symptoms, purchased item).

Example Table

Name	T1	T2	T3	T4	T5	T6
X	1	0	1	0	0	0
Y	1	1	0	0	0	0
Z	1	1	0	1	0	0

Dissimilarity (Asymmetric Case)

Let:

- M_{11} = both 1
- M_{10} = object i = 1, object j = 0
- M_{01} = object i = 0, object j = 1
- M_{00} = both 0

Then:

$$d(i,j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$$

Example:
$d(i,j) = \frac{0+1}{2+0+1} = \frac{1}{3}$

Similarity

$$sim(i,j) = 1 - d(i,j)$$

or equivalently,

$$sim(i,j) = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}$$

Symmetric Binary Similarity

If both 1s and 0s matter, use:

$$sim(i,j) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

This is called the **Simple Matching Coefficient (SMC)**.

Numerical Data

Proximity for Numerical Attributes

Object	A1	A2
P1	0	2
P2	2	0
P3	3	1
P4	5	1

Distance between two objects:

1. **Euclidean Distance:**

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

2. **Manhattan Distance (L1 norm):**

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

3. **Supremum / Chebyshev Distance (L∞ norm):**

$$d(x, y) = \max(|x_1 - y_1|, |x_2 - y_2|)$$

Ordinal Data

Attributes that have a meaningful **order**, but not a fixed numeric scale.

Example: {High, Medium, Low}

Object	T1
1	High
2	Low
3	Medium
4	High

Assign numerical ranks:

High = 1, Medium = 2, Low = 3

Normalize:

$$z = \frac{x - 1}{N - 1}$$

where N = number of ranks.

Example normalization:

- High = (1–1)/(3–1)=0
- Medium = (2–1)/(3–1)=0.5
- Low = (3–1)/(3–1)=1

Supermarket Example (Binary Attributes)

Products = {Sugar, Coffee, Tea, Rice, Egg, Bread, Biscuit}

- **C1 = {Sugar, Coffee, Tea, Rice, Egg}**
- **C2 = {Sugar, Coffee, Bread, Biscuit}**

Product	Bread	Biscuit	Sugar	Coffee	Tea	Rice	Egg
C1	0	0	1	1	1	1	1
C2	1	1	1	1	0	0	0

Using asymmetric dissimilarity formula:

$$dis(C1, C2) = \frac{M_{01} + M_{10}}{M_{11} + M_{01} + M_{10}} = \frac{5}{7}$$

If extended to a large dataset with 1000 attributes:

$$dis(C1, C2) = \frac{5}{1000}$$

Key Notes

- Use **asymmetric binary measures** for features where “1” is more meaningful than “0.”
- Use **ordinal proximity** when ranking matters but distance scale is not fixed.
- **Numerical proximity** relies on true measurable distances.

Data Preprocessing

Definition:

Data preprocessing is the process of cleaning, transforming, and organizing raw data to make it suitable for analysis or model training.

Key Steps

1. **Handling Missing Values**
    - Techniques:
        - **Imputation:** Replace missing values using mean, median, or mode.
        - **Deletion:** Remove records with missing values (only if few).
        - **Interpolation:** Estimate based on surrounding data.
2. **Noisy Data Handling**
    - Noise = random error or variance in data.
    - Techniques:
        - **Binning (smoothing)**
        - **Clustering**
        - **Regression smoothing**
3. **Dimensionality Reduction**
    - Reduce the number of features while preserving information.
    - Example method: **PCA (Principal Component Analysis)**

Binning (Smoothing Technique)

Purpose: To smooth noisy data by grouping values into bins.

Steps:

1. Sort data.
2. Partition into bins (equal width or equal frequency).
3. Replace values by:
   - **Bin Mean**
   - **Bin Median**
   - **Bin Boundaries**

Example:

Data: 4, 8, 15, 21, 21, 24, 25, 28, 34

Bins (3 bins of 3 elements):

- Bin1: 4, 8, 15
- Bin2: 21, 21, 24
- Bin3: 25, 28, 34

1. Mean Smoothing

- Bin1 → 9, 9, 9
- Bin2 → 22, 22, 22
- Bin3 → 29, 29, 29

2. Boundary Smoothing

- Bin1 → 4, 4, 15
- Bin2 → 21, 21, 24
- Bin3 → 25, 25, 34

Pearson Correlation

Measures linear relationship between two variables.

$$-1 \leq r \leq +1$$

- **r = +1** → perfectly positively correlated
- **r = -1** → perfectly negatively correlated
- **r = 0** → no linear correlation

Histogram vs Bar Chart

Chart Type	Data Type	Description
Histogram	Continuous variables	Represents frequency distribution of numeric data
Bar Chart	Categorical variables	Represents discrete categories

Normalization

Normalization rescales data to a fixed range (commonly [0, 1]).

1. Min–Max Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Scales data to range [0, 1].

Example:

x	1	2	3	4
x'	0	1/3	2/3	1

2. Z-Score Normalization (Standardization)

$$z = \frac{x - \mu}{\sigma}$$

where

μ = mean,

σ = standard deviation.

Transforms data to have:

- Mean = 0
- Standard deviation = 1

Example:

Data = 1, 2, 3, 4

Mean = 2.5, Std Dev = 1.12

x	z-score
1	-1.34
2	-0.45
3	0.45
4	1.34

Outlier Detection (from Z-score)

If

$$|z| > 3$$

→ The point is considered an **outlier**.

Other method: **IQR (Interquartile Range)**

Discretization

Definition: Converting **continuous attributes** into **discrete** or categorical intervals.

Used for:

- Simplifying models
- Enabling algorithms that require categorical input

Example:
Ages (continuous) → Bins like:

- 0–18: “Child”
- 19–35: “Adult”
- 36–60: “Middle-aged”
- 60+: “Senior”

Summary Table

Technique	Purpose	Example
Imputation	Handle missing values	Replace nulls with mean/median
Binning	Smooth noisy data	Mean/median/boundary
Normalization	Scale data	Min–max or Z-score
Dimensionality Reduction	Reduce features	PCA
Discretization	Continuous → categorical	Age ranges

Assign Data to Bins

Bin	Range	Values
1	18–31	18, 21, 28, 30
2	31–44	35, 40, 42
3	44–57	50
4	57–70	60, 70

Summary

- **Equal-width binning** → divides by *range*.
- **Equal-frequency binning** → divides by *number of records per bin*.
- Equal-width is simple but **can be skewed** if data is unevenly distributed.