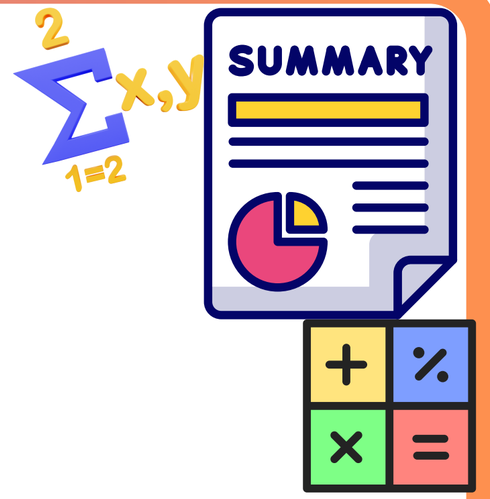



# Day 2
# Preparing Data for Analysis

## Session 2

# Summarizing the data

# Lesson Outcome

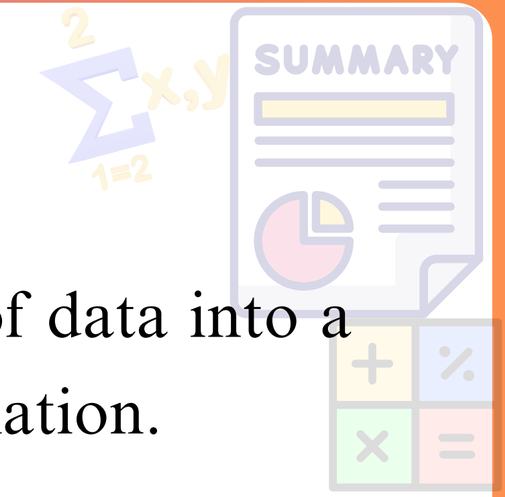
**After completing this session, researchers will be able to**

- Understand the Importance of Data Summarization in Analysis.
- Define and use Tabulation for Organized Data Presentation
- Understand Different Types of Frequency Distributions
- Apply Frequency Distributions in Different Datasets

# Lesson Outline

- Overview of Data Summarization
- Data Summarization Techniques
- Tabulation
- Frequency Distribution
- Cumulative Frequency Distribution

# Data Summarization




## Definition

Data summarization refers to the process of distilling large volumes of data into a more concise and manageable form while retaining the essential information.

## Objectives

- Simplify dataset representation for easier analysis and understanding.
- Offer valuable insights into effectiveness for stakeholders.

# Data Summarization

**Raw Data:**

| Roll | Name | Gender | Day 1 | Day 2 | ... | Day 60 |
|---|---|---|---|---|---|---|
| 1 | Adiyan | M | P | P | ... | A |
| 2 | Mim | F | P | A | ... | P |
| 3 | Sayan | M | A | P | ... | A |
| ... | ... | ... | ... | ... | ... | ... |
| 39 | Ayan | M | P | P | ... | P |
| 40 | Noor | M | P | A | ... | P |

# Data Summarization

**Summarized Data**

| Male | Female |
|---|---|
| 56% | 44% |

| Name | Present (Percentage) | Absent (Percentage) |
|---|---|---|
| Adiyan | 50 (83.3%) | 10 (16.67%) |
| Mim | 45 (75%) | 15 (25%) |
| Sayan | 55 (91.67%) | 5 (8%) |
| .... | .... | .... |
| Noor | 40 (66.67%) | 20 (33.33%) |

# Importance of Data Summarization

**Importance of Data Summarization**

- **Enhances Data Understanding**
- **Facilitates Data Interpretation**
- **Enables Effective Communication**
- **Saves Time and Resources**
- **Supports Decision-Making**

# Importance of Data Summarization

**Enhances Data Understanding:** Summarization distills complex datasets into digestible formats, aiding in grasping the main points and trends efficiently.

**Facilitates Data Interpretation:** By highlighting key insights, summarization enables easier interpretation of data patterns and relationships.

**Enables Effective Communication:** Summarized data provides a clear and concise way to convey findings, making it easier to communicate complex information to diverse audiences.

**Saves Time and Resources:** By condensing large volumes of data, summarization reduces the time and effort required for analysis and reporting.

**Supports Decision-Making:** Summarized data provides actionable insights that support informed decision-making processes across various domains and industries.

# The Benefit of Data Summarization

- Simplifies complex data for easier understanding.
- Accelerates decision-making processes by providing concise insights.
- Saves time and resources by condensing large datasets.
- Enables effective communication of key findings to stakeholders.
- Identifies trends, patterns, and outliers for actionable insights.

# The Different Types of Summarization Techniques

## Methods for Summarizing Data

### 1. Tabular Summarization

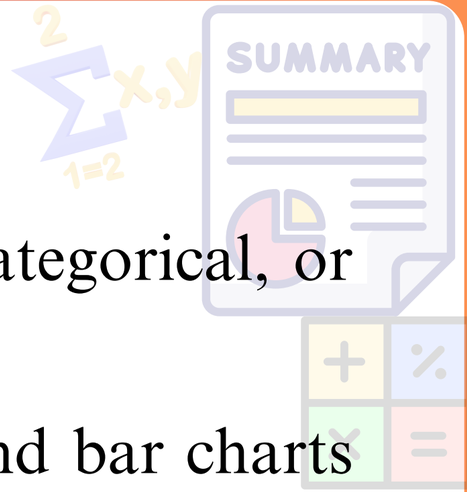
#### Frequency Distribution

- Cumulative Distribution
- Relative Frequency Distribution
- Percentage Distribution

### 2. Visual Summarization

- Bar graph
- Pie chart
- Line chart
- Histogram
- Dot plots
- Box-whisker plots
- Q-Q plots

# Choosing the Right Summarization Techniques




- **Understand Data Types:** Identify whether the data is numerical, categorical, or mixed, as this informs the choice of summarization techniques.
- **Frequency Distributions:** Employ histograms for numerical data and bar charts for categorical data to visualize the distribution and frequency of values.
- **Cross-Tabulation:** For categorical data, create cross-tabulations or contingency tables to summarize relationships between variables.
- **Data Visualization**: Choose appropriate graphs, charts, or maps based on the data's nature and the insights you want to communicate

# Tabular Summarization

Tabular summarization or tabulation is a structured method of organizing raw data into rows and columns, with each row representing a category or variable and each column representing a specific attribute.

**Raw Data:** Data Recorded in the sequence in which they are collected before they are presented worked are called raw data.

| Player | Minutes | Points | Rebounds | Assists |
|---|---|---|---|---|
| A | 39 | 20 | 6 | 5 |
| B | 30 | 29 | 7 | 6 |
| C | 22 | 7 | 7 | 2 |
| D | 26 | three | 3 | 9 |
| E | 20 | 19 | 8 | 2 |
| F | 9 | 6 | 14 | 14 |
| G | 14 | 12 | 8 | 3 |
| I | 22 | 33 | ? | 5 |
| J | 34 | 8 | 1 | 3 |
| K | 1 | | 4 | |

# Tabular Summarization

**Example:** Suppose a primary school wants to summarize the attendance records of its students over a certain period, such as a school term. They collect daily attendance data for each student and want to summarize it in a way that provides a clear overview of attendance patterns.

| Student ID | Name | Total Present | Total Absent | Percentage of Attendance |
|---|---|---|---|---|
| 101 | Akher | 75 | 5 | 94.74% |
| 102 | Ruhul | 70 | 10 | 87.50% |
| 103 | Aziza | 65 | 15 | 81.25% |
| 104 | Afroza | 80 | 0 | 100.00% |
| 105 | Afrin | 72 | 8 | 90.00% |

# Tabular Summarization

## Frequency Distribution

A frequency distribution is the organization of the raw data in table form, using classes and frequencies. **Purpose** of the frequency distribution

1. **To organize the data**
2. **To summarize the data**

**1. The Relative Frequency:** The Relative frequency of categories is obtained by dividing the categories of those categories by sum of the all categories. The formula for the relative frequency of a category is:

$$\text{Relative Frequency} = \frac{\text{Frequency of the Category}}{\text{Total Frequency)}}$$

# Tabular Summarization

**2. Cumulative frequency:** Cumulative frequency is a statistical concept that represents the total frequency of observations less than or equal to a particular value in a dataset. It is used to analyze the distribution of data and understand the proportion of observations that fall below a certain threshold.

**3. Percentage Distribution:** The percentage of a categories is obtained by multiplying the relative frequency of that categories by 100.

$$\text{Percentage} = \frac{\text{Frequency of the Category}}{\text{Total Frequency}} \times 100$$

# Tabular Summarization

Two types of frequency distribution :

1. Frequency distribution for qualitative data
2. Frequency distribution for quantitative data

## Frequency Distribution for Qualitative Data

The frequency distribution of qualitative data is the list of all categories and the number of elements that belong to each of the categories.

Example of qualitative data, Gender of 20 students.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| F | M | M | M | F | M | M | F | M | M |
| M | M | F | M | M | F | F | M | F | M |

# Frequency Distribution for Qualitative Data

The frequency distribution of Gender of 20 students is bellowed:

| Gender | Tally marks | Frequency | Cumulative Frequency | Relative Frequency | Percentage |
|---|---|---|---|---|---|
| Male | 𝍸 𝍸 \| | 13 | 13 | $\frac{13}{20} = 0.65$ | $\frac{13}{20} \times 100 = 65\%$ |
| Female | 𝍸 \|\| | 7 | 20 | $\frac{7}{20} = 0.35$ | $\frac{7}{20} \times 100 = 35\%$ |
| Total | 𝍸 𝍸 𝍸 𝍸 | 20 | 20 | 1.00 | 100 |

# Frequency Distribution for Quantitative Data

Frequency distribution of quantitative data is the list of all values/Classes and the number of values that belong to each of the values/class.

Types of frequency distribution of quantitative data:

1. **Ungrouped Frequency distribution**
2. **Grouped Frequency distribution**

## 1. Ungrouped Frequency Distribution

An ungrouped frequency distribution for quantitative data is the list of all values and the number of values that belong to each of the values.

It is constructed when the list of the value of the variable is short.

**Example**: The following data represents the number of family members of 50 families (The list is always short).

# Ungrouped frequency distribution

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 3 | 5 | 6 | 2 | 4 | 5 | 6 | 3 | 4 |
| 2 | 5 | 4 | 5 | 5 | 6 | 2 | 3 | 4 | 5 |
| 5 | 6 | 3 | 5 | 7 | 5 | 8 | 6 | 5 | 7 |
| 4 | 8 | 5 | 2 | 6 | 3 | 4 | 5 | 3 | 8 |
| 7 | 5 | 6 | 5 | 6 | 5 | 4 | 7 | 8 | 5 |

Constructing a frequency distribution for this set of ungrouped data.

# Ungrouped frequency distribution

| Value | Tally | Frequency | Cumulative Frequency | Relative Frequency | Percentage |
|---|---|---|---|---|---|
| 2 | IIII | 04 | 04 | 0.08 | 8.0% |
| 3 | 𝍸 I | 06 | 10 | 0.12 | 12% |
| 4 | 𝍸 | 05 | 15 | 0.10 | 10% |
| 5 | 𝍸 𝍸 IIII | 14 | 29 | 0.28 | 28% |
| 6 | 𝍸 III | 08 | 37 | 0.16 | 16% |
| 7 | IIII | 04 | 41 | 0.08 | 8.0% |
| 8 | III | 03 | 44 | 0.06 | 6.0% |

## 2. Grouped frequency distribution

A grouped frequency distribution for quantitative data consists of all classes or groups and their respective values, used when the variable list is long. However, these types of distributions are less easily identifiable for numerical data.

Determining the intervals of a frequency distribution for numerical data requires answers to certain questions.

i. How many intervals should be use?

ii. How wide should each interval begin?

iii. Where does the first interval begin?

iv. Where does the last interval begin?

# Grouped frequency distribution

The general rules for preparing frequency distributions that make it easier for us to answer these types of questions to summarize data and to communicate results.

**Step 1: Determine the number of classes.**

Determine the value of k, the number of classes, k may be determined by the Sturge's rules

$$k = 1 + 3.322 * log_2 (n)$$

Where ***n*** is the number of observations.

A useful way to determine the number of classes (k) is the "2 to the k rule". This guide suggests we select the smallest number (k) that is greater than the number of observations (n) that is $2^k \geq n.$

**Step 2: Determine the class interval**

Generally, the class interval is the same for all classes. The classes all taken together must cover at least the distance from the minimum value in the data up to the maximum value. Expressing these words in a formula:

$$w = \frac{Maximum\ value - Minimum\ value}{k}$$

Where, w is rounded up to the closest continent number.

**Step 3: Set the individual class limits:** State clear class limits so you can put each observation into only one category. This means you must avoid overlapping or unclear class limits.

- **Class Limits can be two Types:**

1. **Discrete Class Limits**

# Grouped frequency distribution

Discrete class limits are distinct, non-overlapping values that define the boundaries of each class or interval in a data set, typically countable and limiting the number of values.

**Example**: Consider a data set representing the number of students who scored certain marks in an exam:

| Score Range | Number of Students |
|---|---|
| $0 - 10$ | 3 |
| $11 - 20$ | 5 |
| $21 - 30$ | 7 |

Here, the discrete class limits are 0, 10, 11, 20, 21, 30.

**2. Continuous Class Limits**

Continuous class limits define interval boundaries in continuous data sets, allowing for potential overlap and encompassing values like height, weight, temperature, and time.

# Grouped frequency distribution

**Example**: Consider a data set representing the heights of students in a classroom:

| Height Range (in cm) | Number of Students |
|---|---|
| $150 - 160$ | 5 |
| $161 - 170$ | 7 |
| $171 - 180$ | 3 |

Here, the continuous class limits are 150 (lower limit of the first class), 160 (upper limit of the first class), 161 (lower limit of the second class), 180 (upper limit of the last class).

**Step 4:** Tally the value into the classes and determine the number of observations in each class.

# Grouped frequency distribution

**Example:** Given the height(cm) data of 20 grade one and two students, we aim to create various concise frequency distributions for data summarization.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **65** | **98** | **55** | **62** | **79** | **59** | **51** | **90** | **72** | **56** |
| **70** | 62 | 66 | 80 | 94 | 79 | 63 | 73 | 71 | 85 |

**Solution:**

**Step 1: Determining the number of classes.**

Using Sturges' rule:

$$k = 1 + 3.322 * log_2(20) = 5.32 \approx 6$$

**Step 2: Determining the class interval**

Range = Maximum value – Minimum value

Range = 98 – 51 = 47

# Grouped frequency distribution

$$\text{Width of each interval} = \frac{Maximum - minimum}{k} = \frac{98 - 51}{6} = 7.83$$

**Creating Intervals:** We start from the minimum value and add the width successively to create.

| Class | Tally | Frequency | Cumulative Frequency | Relative Frequency | Percentage |
|---|---|---|---|---|---|
| 51 - 59 | \|\|\| | 3 | 3 | 0.15 | 15.00% |
| 59 - 67 | 𝍸 | 5 | 8 | 0.25 | 25.00% |
| 67 - 75 | \|\|\| | 3 | 11 | 0.15 | 15.00% |
| 75 - 83 | \|\| | 2 | 13 | 0.10 | 10.00% |
| 83 - 91 | \|\|\| | 3 | 16 | 0.15 | 15.00% |
| 91 - 99 | \|\|\|\| | 4 | 20 | 0.20 | 20.00% |

# Grouped frequency distribution

**Interpretation:**

**Which class has the maximum number of students?**

Class 59 to 67 has the maximum number of students.

**Which class has the minimum number of students?**

Class 75 to 83 has the minimum number of students.

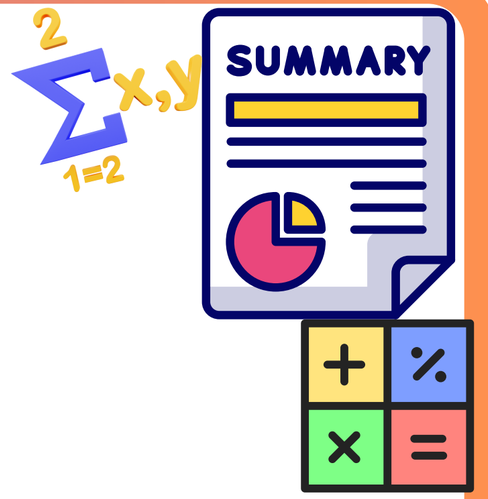
**How many students are below are equal to the height of 75 cm?**

A total of 11 students have a height below 75 cm.

**Which percentage of the students have a height between 83 to 91 cm?**

15% of the total students fall in the height class 83 to 91 cm.

| Class | Frequency | Cumulative Frequency | Relative Frequency | Percentage |
|---|---|---|---|---|
| 51 - 59 | 3 | 3 | 0.15 | 15.00% |
| 59 - 67 | 5 | 8 | 0.25 | 25.00% |
| 67 - 75 | 3 | 11 | 0.15 | 15.00% |
| 75 - 83 | 2 | 13 | 0.10 | 10.00% |
| 83 - 91 | 3 | 16 | 0.15 | 15.00% |
| 91 - 99 | 4 | 20 | 0.20 | 20.00% |




# Thank You