
UNIT 1

DESCRIPTIVE DATA ANALYSIS

Structure

- 1.1 Introduction
 - 1.2 Expected Learning Outcomes
 - 1.3 Descriptive analysis
 - 1.3.1 Data Collection
 - 1.3.2 Types of Data
 - 1.3.3 Frequency Distribution of a Variable
 - 1.3.4 Graphical Representation of Frequency Distribution
 - 1.3.5 Measures of Central Tendency
 - 1.3.6 Measures of Dispersion
 - 1.3.7 Summarization of Data
 - 1.4 Summary
 - 1.5 Solutions/Answers
 - 1.6 Further Readings
-

1.1 INTRODUCTION

Descriptive analysis is a foundational aspect of data analysis that involves summarizing and organizing data to uncover meaningful patterns, trends, and insights. Before conducting any advanced statistical modelling or inferential analysis, it is essential to understand the basic characteristics of the data at hand. This unit introduces the key concepts and techniques used to perform descriptive analysis, providing a solid groundwork for deeper statistical exploration.

The unit begins by explaining the process of collecting data, which is the first critical step in any data analysis task. It discusses various kinds of data—including qualitative and quantitative types—and how these classifications influence the choice of analytical methods.

Once data has been collected, organizing it into a frequency distribution helps to identify how values are distributed across different categories or intervals. To enhance interpretability, the unit explores graphical representations of frequency distributions, such as histograms, bar charts, and pie charts, which visually depict the spread and concentration of data values.

Following this, the unit covers techniques for summarization of data, highlighting how large and complex datasets can be distilled into more manageable forms. This leads to a discussion on measures of central tendency, including the mean, median, and mode, which describe the central or typical value in a dataset. Complementing these are the measures of dispersion or variability, such as range, variance, and standard deviation, which provide insights into the spread or consistency of the data.

Through these topics, the unit equips readers with essential tools for understanding and describing data in a clear, concise, and informative manner. These skills are not only crucial

for statistical analysis but also for effective communication of findings in academic, professional, and real-world contexts.

1.2 EXPECTED LEARNING OUTCOMES

After completing this unit, you should be able to:

- Understand the process of collecting data and identify appropriate sources and methods for different types of analysis.
- Distinguish between different kinds of data such as qualitative and quantitative, and understand their relevance in statistical analysis.
- Organize raw data into a frequency distribution to facilitate easier interpretation and analysis.
- Create and interpret various graphical representations of frequency distributions, including bar charts, histograms, and pie charts.
- Summarize datasets effectively using basic descriptive statistics.
- Calculate and interpret measures of central tendency such as mean, median, and mode to describe the central value of a dataset.
- Calculate and interpret measures of dispersion or variability, such as range, variance, and standard deviation, to understand the spread of data.
- Apply descriptive analysis techniques to extract meaningful insights and support data-driven decision-making.

1.3 DESCRIPTIVE DATA ANALYSIS

Descriptive Data Analysis is the process of examining, summarizing, and visualizing data to understand its main characteristics before applying any formal statistical modeling or hypothesis testing. It helps in making sense of raw data by organizing it into meaningful patterns and trends. This method is widely used in research, business, healthcare, education, and many other fields for preliminary data interpretation. It forms the backbone of any data-driven inquiry. By systematically collecting, organizing, visualizing, and summarizing data, it allows analysts to extract meaningful insights that inform decision-making and further statistical procedures. Whether dealing with student scores, sales figures, or survey responses, the principles covered in this unit—ranging from frequency distributions to measures of central tendency and variability—equip learners with essential tools for effective data analysis.

Data Collection is the very first step in the process of analysis, thus before any analysis can begin, data must be collected systematically. Data collection involves gathering raw information from various sources, which can include surveys, experiments, observations,

administrative records, and digital sources (e.g., web data, transaction logs). Example: A teacher wants to analyse students' performance in a mathematics test. They collect marks out of 100 from each student in a class of 40.

But for collecting the data systematically one should have the knowledge of the various Types of Data, because understanding the type of data is crucial for selecting appropriate analysis methods. Data can be broadly classified as Qualitative or Categorical data type and Quantitative or Numerical data type.

The Qualitative (Categorical) Data is further classified as follows:

- Nominal: Categories with no inherent order (e.g., gender, religion)
- Ordinal: Categories with a logical order (e.g., education level: high school, undergraduate, postgraduate)

The Quantitative (Numerical) Data is further classified as follows:

- Discrete: Countable values (e.g., number of children)
- Continuous: Any value within a range (e.g., height, weight)

Now, after collection of suitable data, it is generally required to analyse the data where in the frequency distribution is the point to begin with. A frequency distribution shows how often each different value in a set of data occurs. It helps in identifying patterns such as concentration of data, outliers, and data spread.

Further, the Graphical Representation of Frequency Distribution i.e. Visualizing data helps you quickly identify patterns, trends, and outliers. Several types of graphs are used in descriptive analysis, a few are listed below:

- *Bar charts* are used for categorical data.
- *Histograms* show the distribution of continuous numerical data.
- *Pie charts* display proportions of categories.
- *Line graphs* show trends over time.

These graphs are quite useful in summarizing the data, which involves condensing the data into key statistics or visual formats that convey the main features. Summarizing the data helps you focus on important features without listing every data point. Summaries can be in the form of tables, graphs, or statistics like the average, range, and median etc., which are in fact the Measures of Central Tendency. These measures tell you what is typical or average in your dataset. The three main measures of central tendency are as follows:

- *Mean* is the arithmetic average. Add all values and divide by the number of values.
- *Median* is the middle value when data is sorted.
- *Mode* is the value that appears most frequently.

Each measure gives different insights into the dataset. The mean tells you the overall average, the median shows the central value, and the mode highlights the most frequent score.

Further, to look into the quality of decision one needs to Measure the Dispersion or Variability. These measures of dispersion or variability, describe the spread or variability in a dataset. The measures of dispersion are classified as follows:

- *Range*: Difference between the maximum and minimum values
- *Variance*: Average of the squared differences from the Mean
- *Standard Deviation (SD)*: Square root of the variance; shows how much data deviates from the mean

A higher standard deviation indicates more variability in the data, while a lower one suggests the data points are closer to the mean.

These concepts are discussed in more detail in the subsequent sections, and as you practice these concepts with real-world data, you'll develop a strong foundation in data analysis and become better equipped to interpret and communicate your findings effectively.

Check Your Progress 1

Q1. Explain the concept of Descriptive Data Analysis. Why is it important in data analysis?

.....

.....

Q2. Classify the types of data with suitable examples.

.....

.....

Q3. What are Measures of Central Tendency and Dispersion? Explain briefly.

.....

.....

1.3.1 DATA COLLECTION

Data collection is the first and one of the most important steps in any data analysis. It means carefully gathering the information you need to study a topic or answer a question. Whether you're doing research, running a project, or making business choices, the data you collect is the foundation for your work. If the data is not collected properly, your results might be wrong and lead to bad decisions. But if the data is collected correctly, it helps you understand the situation better and make smarter, more accurate decisions.

One needs to understand that this step is quite crucial, as it is just the beginning of any study/analysis, if error persists in the beginning itself i.e. at the stage of data collection itself, then it may amplify the errors and inaccuracy of results at later stages.

When working with data for analysis, it's important to know where your data is coming from, is it coming from the primary source or the secondary source? Depending on the source of data collection the data is broadly classified into two categories i.e. Primary data and Secondary Data. Knowing the difference between these two helps you decide how to collect the right kind of data for your project or research.

Before understanding the methods of gathering these types of data i.e. Primary data or Secondary data, let's understand the meaning of Primary data and Secondary data respectively.

Primary data is data that you collect yourself, directly from the source, for a specific purpose. This means it is original and has never been used before. You decide what information to collect, how to collect it, and from whom. Since you are directly involved in gathering the data, it is often **more accurate and relevant** to your needs.

However, collecting primary data can take a lot of time and effort. It might also be more expensive, especially if you're surveying a large number of people or conducting experiments.

Let's look at a few common methods used to collect primary data:

- **Surveys and Questionnaires:** These are sets of questions you ask people to get their responses. For example, a company may ask customers to fill out a feedback form after using a product.
- **Interviews:** In this method, you talk to people directly—either face-to-face or over a call—and ask specific questions. For instance, a researcher might interview farmers to learn about their crop practices.
- **Observations:** Here, you watch how people behave or how things happen and take notes. For example, a teacher may observe students during a group activity to understand how they interact.
- **Experiments:** You create a controlled environment to test how changes in one thing affect another. For example, a scientist might test how a new fertilizer affects plant growth.
- **Focus Groups:** This involves a small group of people discussing a topic while a moderator guides the conversation. For instance, a company may gather a group to talk about a new product idea and get their opinions.

Each of these methods gives you fresh, direct information that is well-suited to your specific needs.

Secondary Data is the information that has already been collected and published by someone else, usually for a different purpose. You use this existing data for your own study.

For example, if you are studying population growth, you might use census data already published by the government. This saves you time and money because the data is already available.

Secondary data can come from many sources, such as:

- **Government Reports and Publications:** These include census data, health statistics, or economic reports from official bodies like the Ministry of Health or the Reserve Bank.
- **Institutional Records:** Schools, colleges, hospitals, and companies often maintain records that can be useful. For example, student performance data from a school can help in analyzing academic trends.
- **Research Articles and Journals:** You can use published research papers to support your work or learn from previous studies.
- **Books, Magazines, and Newspapers:** These can offer useful facts, opinions, or background information.
- **Websites and Online Databases:** Many organizations make data freely available online. For example, economic data can be downloaded from the World Bank or websites like data.gov.in.
- **Company Reports:** Financial statements, annual reports, and sales data can help analyze business performance.

*Based on the above information one may consolidate the **differences between the Primary and Secondary Data**. Broadly the difference between the two are as follows:*

- Primary data is collected directly by you for your current study. It is original, often more accurate, but can take more time and effort to gather.
- Secondary data is collected by someone else, usually for a different purpose. It is easy to access and less costly, but may not fully meet your current needs.

In general, there is a question that among the primary data and the secondary data, Which One Should You Use? The answer to this query about the choice among primary and secondary data depends on your goals, time, and budget.

- If you need specific, detailed, and up-to-date information, and you have the resources to collect it yourself, go for primary data.
- If good-quality data already exists and is suitable for your purpose, secondary data can save you time and effort.

In many cases, a good research project uses both types of data—primary data to get fresh insights, and secondary data to support or compare with existing information.

Understanding the difference between primary and secondary data helps you plan your data collection more effectively. Primary data gives you control and relevance, while secondary data gives you speed and convenience. By choosing the right type of data—or combining both—you can ensure your analysis is strong, reliable, and meaningful.

☛ Check Your Progress 2

Q1. What is Data Collection? Why is it considered a crucial step in data analysis?

.....

.....

Q2. Differentiate between Primary Data and Secondary Data.

.....

.....

Q3. Explain any three methods of collecting Primary Data with examples.

.....

.....

1.3.2 TYPES OF DATA

Before you begin any kind of data analysis, it's important to understand the type of data you're working with. This is because different types of data require different methods of analysis, visual representation, and interpretation. We learned from the above section 1.4 that depending on the source of data collection the data is broadly classified in to two categories i.e. Primary data and Secondary Data, and we learned about them. Now we are in position to take a deeper dive and understand the various types of the data which contributes towards the data analysis. Although a brief introduction for the types of data was made in the introduction section 1.1, now we will elaborate and discuss these data types here.

Broadly, data can be classified into two main categories: qualitative (or categorical) data and quantitative (or numerical) data. Each of these has subtypes, and some data types have more structure than others.

Qualitative Data (Categorical Data): It describes qualities or characteristics. This type of data is non-numeric and is used to classify data into categories or groups. You cannot perform mathematical operations like addition or averaging on this data, but you can count how often each category occurs.

There are two main types of qualitative data:

a) Nominal Data: Nominal data refers to data that labels categories without any order. These categories are simply names or labels used to identify different groups. There is no ranking or natural order among the values. *Examples: Gender: Male, Female, Other; Blood Group: A, B, AB, O; Types of Fruits: Apple, Banana, Mango.* Here, one category isn't greater or smaller than the other—just different.

b) Ordinal Data: Ordinal data represents categories that do have a logical order, but the differences between the categories aren't measurable. You can rank them, but you can't tell exactly how much more or less one is compared to another. *Examples:*

Education Level: *High School, Undergraduate, Postgraduate;* **Customer Satisfaction:** *Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied;* **Socioeconomic Status:** *Low, Middle, High.*

You can say that someone with a postgraduate degree has a higher qualification than someone with just high school, but you can't measure exactly how much more educated they are based on these labels.

Quantitative Data (Numerical Data): It consists of numbers and represents values that can be measured or counted. You can perform mathematical operations on this data, and it provides more depth for analysis.

Quantitative data is further divided into:

a) Discrete Data: Discrete data includes **countable values**. These are often whole numbers and usually answer questions like "how many?" *Examples:* Number of students in a class: 25 ; Number of books in a library: 5,000 ; Number of children in a family: 2, 3, 4

Discrete values are separate and distinct. You can't have 2.5 students or 3.8 cars.

b) Continuous Data: Continuous data can take **any value within a range**, and it's often used to measure things. It includes decimals and fractions. *Examples:* Height: 172.5 cm ; Weight: 68.2 kg ; Temperature: 36.6°C.

Since continuous data can take infinite values within a range, it provides a lot of flexibility for analysis.

Like *Nominal and Ordinal data* under the category of Qualitative data, we are also having the *Interval and Ratio data* under the category of Quantitative data.

The **Interval data** is numerical data where the difference between values is meaningful, and intervals are equal. However, it lacks a true zero point. This means that while you can add and subtract interval data, you cannot multiply or divide it meaningfully in terms of ratios. *Examples:* Temperature in Celsius or Fahrenheit ; IQ scores ; Dates on a calendar

Let's take temperature as an example. The difference between 10°C and 20°C is the same as between 20°C and 30°C—that is, 10°C. But you can't say 20°C is "twice as hot" as 10°C, because 0°C doesn't represent the absence of temperature—it's just another point on the scale.

The **Ratio data** is also numerical and has all the properties of interval data, but with a meaningful or absolute zero point. Because of this, all mathematical operations, including multiplication and division, are valid. *Examples:* Height (e.g., 180 cm) ; Weight (e.g., 70 kg) ; Age (e.g., 25 years) ; Income (e.g., ₹50,000)

With ratio data, you can say that a person who weighs 60 kg is twice as heavy as someone who weighs 30 kg. This is because zero weight means no weight, making the scale absolute.

Now let's look at a more precise classification based on the level of measurement, which includes Nominal, Ordinal, Interval, and Ratio data types. This classification helps determine what types of statistical techniques you can use.

Levels of Measurement: The Four Data Types

1. Nominal – Categories without order (qualitative)
2. Ordinal – Categories with order but no fixed interval (qualitative)
3. Interval – Numeric data with equal spacing, but no true zero (quantitative)
4. Ratio – Numeric data with equal spacing and a true zero (quantitative)

Important:

- It is to be noted that the interval data is almost always continuous. But the Ratio data can be either discrete or continuous, depending on whether you're counting or measuring something.
- Interval and Ratio data are two subtypes of quantitative (numerical) data. Both deal with numbers and allow for mathematical operations, but they differ in one key way: ratio data has a true zero point, while interval data does not.
- Both ratio and interval data can be either continuous or discrete, but they are most commonly associated with continuous data.

In this section we learned that understanding data types—nominal, ordinal, interval, and ratio—is a fundamental skill in data analysis. Qualitative data helps classify and group information, while quantitative data allows for precise measurement and mathematical analysis. The level of measurement determines what kind of analysis you can perform. By correctly identifying your data type, you ensure that your analysis is appropriate, accurate, and meaningful.

Whether you're just starting out or working on a complex project, always begin by asking: What type of data am I working with? The answer will guide the rest of your analysis.

☛ Check Your Progress 3

Q1. Explain the four levels of measurement in data analysis i.e. Nominal, Ordinal, Ratio & interval data. Why are they important?

.....
.....

1.3.3 FREQUENCY DISTRIBUTION OF A VARIABLE

A frequency distribution is a powerful tool that transforms raw data into an organized format, making it easier to analyse. Whether you use a simple table for small datasets or a grouped one for larger values, frequency distributions help identify patterns, trends, and key summary statistics like the most frequent (modal) class, range, and central values.

In short, frequency distribution is your first step toward understanding and visualizing data effectively.

When you collect a set of raw data, it can be messy and hard to understand at first glance—especially if the number of observations is large. To make sense of such data, one of the first steps in descriptive data analysis is to organize it into a frequency distribution.

A frequency distribution is a way to arrange data values to show how often each value or range of values (called a class) occurs. This helps reveal patterns, trends, and how the values are spread across the dataset.

Imagine you conducted a survey among 30 students to find out how many hours they studied in a week. The responses might look like this:

[4, 7, 6, 5, 4, 9, 10, 7, 6, 4, 8, 7, 6, 5, 5, 4, 8, 9, 6, 5, 6, 4, 7, 6, 5, 10, 8, 7, 6, 9]

Looking at this list directly can be overwhelming. It's hard to draw conclusions from such raw data.

Let's now understand how to create a Discrete *Frequency Distribution* from the data above.

Step 1: Identify the data range:

Minimum value = 4
Maximum value = 10

So, we'll group the data based on the hours studied, from 4 to 10.

Step 2: Count how many times each value occurs

Study Hours (x)	Frequency (f)
4	5
5	5
6	7
7	5
8	3
9	3
10	2

This table shows how many students studied for each number of hours.

From this table, you can quickly observe:

- Most students (7) studied for 6 hours.
- Fewest students (2) studied for 10 hours.
- The data is spread around 4–10 hours.

This kind of representation makes it easy to analyse the data and identify central tendencies (like mean or median), spot patterns, or decide the next steps in your analysis (e.g., plotting a graph or calculating measures of spread).

In the above example the dataset is quite limited so the creation of frequency distribution is quite simple and is achieved in just 2 steps, as discussed above.

But, If the data range is wide or there are too many individual values, we can group them into class intervals i.e. in case of larger datasets, we need to opt for *Grouped Frequency Distribution*, discussed below

Let's say you surveyed 50 people about their monthly spending on groceries in ₹ (rupees), and the values range from ₹500 to ₹2000. Instead of listing each value, you can group them like this:

Monthly Spending (₹)	Frequency
500–799	6
800–1099	10
1100–1399	15
1400–1699	12
1700–1999	7

This is a *grouped frequency distribution*. It's useful when individual values are not as important as the **range** they fall into.

To understand the Grouped Frequency Distribution in a better way, let's understand how to create it by an example in a stepwise manner.

Let's go through the steps one by one using an example:

Example Scenario: Suppose you conducted a test and collected the **marks obtained by 40 students** out of 100. Here's the (unordered) data:

72, 55, 48, 89, 90, 68, 77, 52, 69, 73, 64, 83, 95, 70, 66, 71, 60, 58, 85, 80, 74, 56, 63, 61, 86, 67, 84, 76, 54, 59, 62, 65, 79, 78, 50, 81, 75, 88, 53, 57

Step 1: Identify the Range

- Find the **lowest** and **highest** values.
 - Lowest = 48
 - Highest = 95
- **Range = 95 – 48 = 47**

Step 2: Decide the Number of Classes

A good rule of thumb is to use between **5 and 10 classes**, depending on the data size. For 40 values, let's go with **8 classes**.

Step 3: Calculate Class Width ; = Range/Number of Classes = $47/8 \approx 5.875$

We round this up to the next whole number: **6**

So each class will cover an interval of **6 marks**.

Step 4: Create Class Intervals

Start from the lowest value (rounded down, if needed). Let's start at **48**. The intervals will be:

- 48–53
- 54–59
- 60–65
- 66–71
- 72–77
- 78–83
- 84–89
- 90–95

Note: The intervals are **continuous**, meaning there is no gap between them.

Step 5: Frequency Distribution Table (Completed)

Now, go through the dataset and count how many values fall into each class.

Class Interval	Frequency
48–53	4
54–59	6
60–65	6
66–71	6
72–77	6
78–83	5
84–89	5
90–95	2
Total	40

Each frequency value represents how many student marks fall within that class interval. The total (sum of frequencies) equals 40, matching the total number of students in the dataset.

Steps of Creating a continuous frequency distribution involves:

1. Finding the range of the data.
2. Deciding on the number of class intervals.
3. Calculating the class width.
4. Creating non-overlapping, continuous intervals.
5. Tallying the frequency of data in each class.

When studying frequency distribution, you may come across a few related terms like Relative and cumulative frequencies. Let's further extend our discussion, and understand the concept of Relative Frequency and Cumulative Frequency.

- **Relative Frequency:** The proportion or percentage of the total that each class represents.
 - Formula: $\text{Relative Frequency} = (\text{Class Frequency} / \text{Total Frequency})$

- **Cumulative Frequency:** The running total of frequencies up to a certain class.
 - Useful for percentile and median calculations.

Enhanced Frequency Distribution Table

Given:

Total number of students = 40

We'll compute:

- **Relative Frequency** = Frequency $\div$ Total
- **Cumulative Frequency** = Running total of frequencies

Class Interval	Frequency (f)	Relative Frequency (f/40)	Cumulative Frequency
48–53	4	$4 / 40 = 0.10$	4
54–59	6	$6 / 40 = 0.15$	10
60–65	6	$6 / 40 = 0.15$	16
66–71	6	$6 / 40 = 0.15$	22
72–77	6	$6 / 40 = 0.15$	28
78–83	5	$5 / 40 = 0.125$	33
84–89	5	$5 / 40 = 0.125$	38
90–95	2	$2 / 40 = 0.05$	40
Total	40	1.00	—

Here, in the above table the Relative Frequency shows the proportion or percentage of students in each class. Whereas the Cumulative Frequency shows how many students scored up to and including a given class range.

1.3.4 GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTION

Graphical representation of frequency distribution is a method used to visually summarize and interpret data. Instead of examining rows and columns of numbers in a table, graphs provide a clear picture of how data is spread across different values or intervals. This makes it easier to identify trends, patterns, outliers, and the overall shape of the distribution. Several types of graphs are commonly used depending on the nature of the data and the purpose of the analysis.

One of the most frequently used graphs for continuous numerical data is the **histogram**. A histogram consists of adjacent (touching) bars where the horizontal axis (X-axis) represents the class intervals, and the vertical axis (Y-axis) shows the frequencies—the number of data values

that fall within each interval. Each bar's height corresponds to the frequency for that class. For example, if students' exam marks are grouped into intervals like 40–49, 50–59, 60–69, and so on, a histogram can quickly show how many students fall into each score range. Since the data is continuous, there are no gaps between bars.

Steps to Create a Histogram for Grouped (Continuous) Data

Creating a histogram from grouped data involves a systematic process that helps visualize the distribution of a dataset. Let's walk through the steps using the example of marks obtained by 40 students, distributed across class intervals.

Step 1: Organize the Data

The first step is to prepare a frequency distribution table. This table should include clearly defined, non-overlapping class intervals and their corresponding frequencies. For example:

Class Interval	Frequency
48–53	4
54–59	6
60–65	6
66–71	6
72–77	6
78–83	5
84–89	5
90–95	2

Make sure that the class intervals are **continuous**, meaning there are no gaps between them, and that the frequencies are accurately counted from the raw data.

Step 2: Draw the Axes

Next, draw two perpendicular axes to create your graph. The **horizontal axis (X-axis)** represents the class intervals, such as 48–53, 54–59, and so on. These intervals should be equally spaced and in ascending order. The **vertical axis (Y-axis)** represents the frequency or the number of observations in each class. Choose a scale for the Y-axis that comfortably accommodates the highest frequency in your table; in this case, the highest frequency is 6.

Step 3: Draw Bars

Now, for each class interval, draw a bar such that:

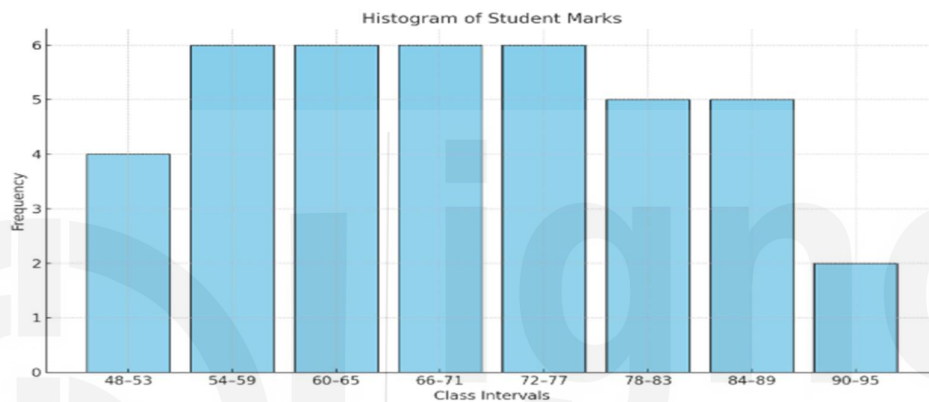
- The **width** of each bar is equal (because our class intervals are of equal width — 6 marks in this example).

- The **height** of each bar corresponds to the **frequency** for that interval.

For example, the bar for the 48–53 interval should have a height of 4, while the bar for 54–59 should rise to 6, and so on. Since the data is continuous, ensure that the bars are **adjacent to each other without any gaps**.

Step 4: Label the Histogram

Once the bars are drawn, label your chart clearly. Give the histogram a descriptive **title**, such as “Histogram of Student Marks.” The **X-axis** should be labeled as “Class Intervals” or “Marks,” and the **Y-axis** should be labeled as “Frequency” or “Number of Students.” This labeling helps in interpreting the graph easily.



Here's the histogram representing the distribution of student marks based on the continuous frequency distribution table:

Step 5: Analyze the Histogram

After constructing the histogram, analyze its shape and pattern. Look for where most of the values are concentrated, which indicates the **central tendency** of the data. You can also assess whether the distribution is **symmetrical**, **skewed** to the left or right, or has multiple peaks (**bimodal**). Additionally, observe any **outliers** or gaps in the data. This visual analysis can provide quick insights into the nature of the dataset.

Another useful graphical tool is the **frequency polygon**, which is similar in purpose to a histogram but uses a line graph instead of bars. In a frequency polygon, midpoints of each class interval are plotted on the X-axis, and frequencies are plotted on the Y-axis. These points are then connected with straight lines to form a polygon. Often, additional points with zero frequency are added at the beginning and end to close the polygon. This graph is especially helpful when comparing multiple distributions on the same plot, making it easier to identify differences in shape and spread.

Steps to Create a Frequency Polygon

Step 1: Prepare the Frequency Distribution Table

Create a table with class intervals and their corresponding frequencies. Also, compute the **midpoints** of each class interval.

Class Interval	Frequency (f)	Midpoint (x)
48–53	4	50.5
54–59	6	56.5
60–65	6	62.5
66–71	6	68.5
72–77	6	74.5
78–83	5	80.5
84–89	5	86.5
90–95	2	92.5

Midpoint is calculated using the formula = (Lower Limit + Upper Limit)/2

Step 2: Add Extra Midpoints at Both Ends (Optional but Recommended)

To close the polygon at the baseline, add one class interval before the first and one after the last with **zero frequency**.

- Before 48–53: midpoint = 44.5, frequency = 0
- After 90–95: midpoint = 98.5, frequency = 0

Now, the extended list of midpoints and frequencies becomes:

Midpoint (x)	Frequency (f)
44.5	0
50.5	4
56.5	6
62.5	6
68.5	6
74.5	6
80.5	5
86.5	5
92.5	2
98.5	0

Step 3: Plot the Points

On graph paper or using software (e.g., Excel, Python, or any graphing tool):

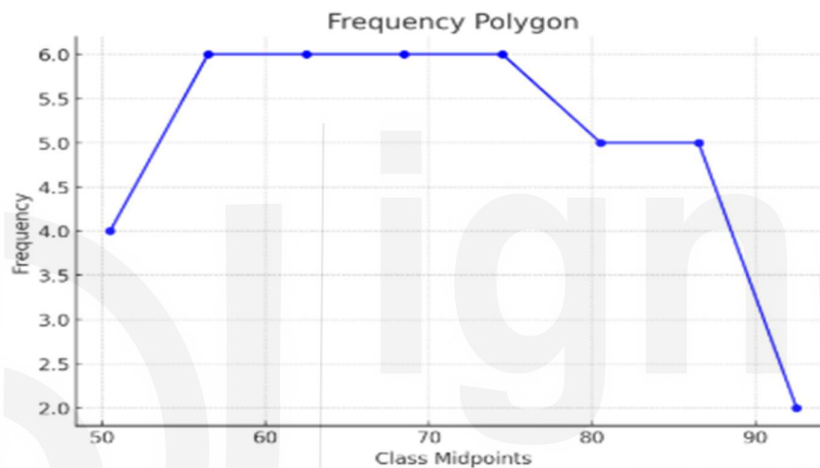
- Plot each point with a midpoint **on the X-axis** and **frequency on the Y-axis**.
- For example, plot the points (50.5, 4), (56.5, 6), and so on.

Step 4: Connect the Points with Line Segments

- Draw straight lines to connect the plotted points in order from left to right.
- Make sure to include the starting and ending points with zero frequency to anchor the polygon to the baseline.

Step 5: Label and Title the Graph

- Add labels for axes: **X-axis:** Class Midpoints **Y-axis:** Frequency
- Title the graph: *Frequency Polygon for Student Marks*



What Does a Frequency Polygon Show? : The Frequency Polygon allows you to see the **shape** of the distribution more clearly than a histogram, especially when comparing multiple datasets. Peaks and symmetry are easier to detect.

Now, For analyzing cumulative trends, the **ogive or cumulative frequency curve** is an effective choice. There are two types of ogives: the "less than" ogive, which plots cumulative frequencies against the upper boundaries of the class intervals, and the "more than" ogive, which uses the lower boundaries. For example, if you want to find how many students scored less than 70 marks, the "less than" ogive can give you a quick visual answer. Ogives are particularly useful for locating the median, quartiles, and percentiles in a data set.

Steps to Create an Ogive (Cumulative Frequency Curve)

Step 1: Prepare the Frequency Distribution Table

Use the same frequency table used for the histogram and compute the **cumulative frequencies**.

Class Interval Upper Boundary Frequency (f) Cumulative Frequency (Less Than Type)

48–53	53	4	4
54–59	59	6	10
60–65	65	6	16
66–71	71	6	22
72–77	77	6	28
78–83	83	5	33
84–89	89	5	38
90–95	95	2	40

Cumulative Frequency is the running total of frequencies up to each class.

For example: $4 + 6 = 10$, $10 + 6 = 16$, and so on.

Step 2: Plot the Points

Each point is plotted with:

- **Upper Class Boundary on the X-axis**
- **Cumulative Frequency on the Y-axis**

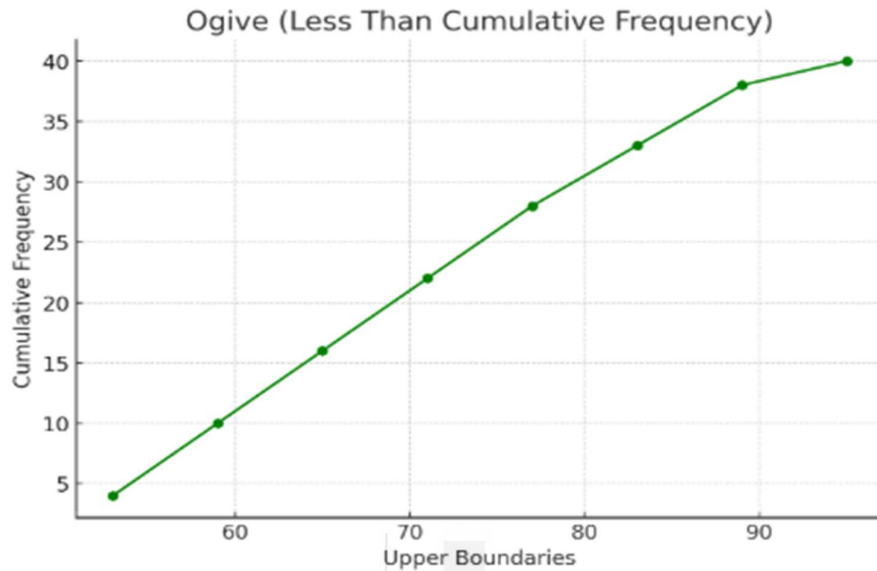
So, plot the points:

(53, 4), (59, 10), (65, 16), (71, 22), (77, 28), (83, 33), (89, 38), (95, 40)

You may optionally begin from a lower value such as (47, 0) to anchor the curve to the X-axis.

Step 3: Connect the Points Smoothly

- Join the plotted points using a smooth curve or straight lines.
- Ensure the curve starts from zero and rises as you move right, because cumulative frequency always increases or remains constant.



Step 4: Label and Title the Graph

- **X-axis:** Upper Boundaries of Class Intervals **Y-axis:** Cumulative Frequency
- **Title:** *Ogive (Less Than Cumulative Frequency Curve) for Student Marks*

What Does Ogive (Cumulative Frequency Curve) Show? : The Ogive helps you to Understand how data accumulates across intervals. Also it estimates the median, quartiles, and percentiles. And helps to quickly answer questions like “How many students scored less than 70?”

To further analyze distribution and detect spread or outliers, one can use a box plot, also known as a **box-and-whisker plot**. This graphical representation displays the five-number summary of a dataset: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box represents the interquartile range (middle 50% of the data), and the lines extending from the box (whiskers) indicate variability outside the upper and lower quartiles. Any values lying outside these whiskers may be considered outliers. For instance, a box plot of students' marks can show not only the median score but also whether any students scored exceptionally low or high.

Steps to Create a Box-and-Whisker Plot (Box Plot)

A Box-and-Whisker Plot visually displays the **distribution**, **spread**, and **skewness** of a dataset through five summary statistics:

- **Minimum**
- **First Quartile (Q1)**
- **Median (Q2)**
- **Third Quartile (Q3)**
- **Maximum**

Step 1: Reconstruct the Raw Data from Grouped Data

Since a Box Plot needs raw data or an approximation of it, reconstruct a **pseudo dataset** using the **midpoints** of each class interval and the corresponding **frequency**. Here's how:

Class Interval	Frequency	Midpoint
48–53	4	50.5
54–59	6	56.5
60–65	6	62.5
66–71	6	68.5
72–77	6	74.5
78–83	5	80.5
84–89	5	86.5
90–95	2	92.5

Total observations (N) = 40

Construct Pseudo Raw Data: i.e. expand the data based on the frequency:

$50.5 \times 4 \rightarrow$ positions 1–4
 $56.5 \times 6 \rightarrow$ positions 5–10
 $62.5 \times 6 \rightarrow$ positions 11–16
 $68.5 \times 6 \rightarrow$ positions 17–22
 $74.5 \times 6 \rightarrow$ positions 23–28
 $80.5 \times 5 \rightarrow$ positions 29–33
 $86.5 \times 5 \rightarrow$ positions 34–38
 $92.5 \times 2 \rightarrow$ positions 39–40

Step 2: Arrange the Data in Ascending Order

Arrange the values from smallest to largest. Since the data is already binned using midpoints, and each midpoint is repeated according to its frequency, it's already structured to represent an ordered dataset.

Step 3: Find the Five-Number Summary i.e. Min., Max, Q1, Q2, and Q3

Using the pseudo raw data given above:

- **Minimum:** Smallest value = 50.5
- **Maximum:** Largest value = 92.5
- **Median (Q2):** Middle value (20th and 21st observation average) ≈ 68.5
- **First Quartile (Q1):** 25th percentile (10th observation) ≈ 58.0
- **Third Quartile (Q3):** 75th percentile (30th observation) ≈ 80.5

So, the five-number summary is:

- **Min** = 50.5
- **Q1** = 58.0
- **Median (Q2)** = 68.5
- **Q3** = 80.5
- **Max** = 92.5

Let's understand how to determine the Quartiles Q1, Q2 and Q3 ; especially in the context of a dataset like the one we've been working with (40 student marks). You can use the same method for any dataset.

So, What Are Quartiles? we have 3 quartiles in general Q1, Q2, and Q3; their respective meanings are as follows:

- **Q1 (First Quartile):** The value that separates the **lowest 25%** of the data.
- **Q2 (Median):** The middle value; separates the **lower 50%** from the **upper 50%**.
- **Q3 (Third Quartile):** The value that separates the **lowest 75%** from the top 25%.

Now, we understand How to Determine Q1, Q2 and Q3 (Manually)

Assume you have $N = 40$ data points, arranged in **ascending order**.

Step A: Find the Positions

- $Q1 \text{ position} = (N+1)/4 = (41)/4 = 10.25^{\text{th}}$ item
- $Q2 \text{ position} = (N+1)/2 = (41)/2 = 20.5^{\text{th}}$ item
- $Q3 \text{ position} = 3(N+1)/4 = 3(41)/4 = 30.75^{\text{th}}$ item

These are **positions**, not values yet.

Step B: Estimate the Quartile Values

To get the values at 10.25th and 30.75th positions, you interpolate between the actual values at the 10th, 11th (for Q1), 20th, 21st (for Q2), and 30th, 31st (for Q3).

Suppose the ordered pseudo data looks like:

(Refer to Pseudo Raw Data given above)

$50.5 \times 4 \rightarrow$ positions 1–4

$56.5 \times 6 \rightarrow$ positions 5–10

$62.5 \times 6 \rightarrow$ positions 11–16

$68.5 \times 6 \rightarrow$ positions 17–22

$74.5 \times 6 \rightarrow$ positions 23–28

$80.5 \times 5 \rightarrow$ positions 29–33

Finding Quartile-1(Q1), from the pseudo raw data given we can see that there are

4 values of 50.5

6 values of 56.5 → positions 5 to 10
6 values of 62.5 → positions 11 to 16

So the **10th value = 56.5**, and **11th value = 62.5**

Thus, $Q1 = 56.5 + 0.25 \times (62.5 - 56.5) = 56.5 + 1.5 = 58.0$

(Q2) Median position = $(N+1)/2 = 41/2 = 20.5$ th value

We want the **value at the 20.5th position** in the ordered pseudo data.

Now from the pseudo raw data given we can see that both
20th value = 68.5 and 21st value = 68.5

Thus, $Q2 = (68.5 + 68.5)/2 = 68.5$

Since the 20.5th value lies between 68.5 and 68.5, the median is **68.5**

Similarly, **30th value = 80.5**, **31st value = 86.5**

$Q3 = 80.5 + 0.75 \times (86.5 - 80.5) = 80.5 + 4.5 = 85.0$

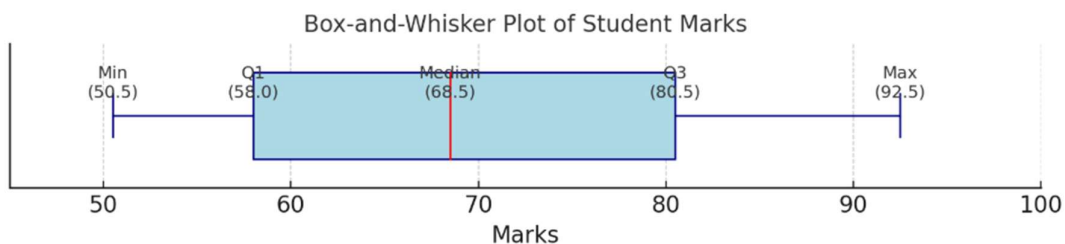
So, $Q1 \approx 58.0$, $Q2 \approx 68.5$ and $Q3 \approx 85.0$ in this case.

Step 4: Draw the Box-and-Whisker Plot

- Draw a horizontal number line that covers the range of your data (e.g., from 45 to 100).
- Mark the five-number summary values on the line.
- Draw a rectangular box from Q1 to Q3 .
- Draw a vertical line inside the box at the median .
- Draw whiskers (horizontal lines) from:
 - Min to Q1
 - Q3 to Max

Step 5: Label the Plot

- Label the five-number summary on the axis.
- Title your plot appropriately: “*Box-and-Whisker Plot of Student Marks*”



Here is the **Box-and-Whisker Plot** based on the student marks data. It visually displays the five-number summary:

Minimum: 50.5 ; **Q1 (First Quartile):** 58.0 ; **Median (Q2):** 68.5 ;
Q3 (Third Quartile): 80.5 ; **Maximum:** 92.5

This plot helps you understand the **spread**, **center**, and **symmetry** of the data, as well as spot potential outliers (though none appear here).

In short, The box plot helps to:

- Visualize the **spread and central tendency**.
- Detects **symmetry** or **skewness**.
- Identify **potential outliers** (values far from the whiskers).
- Quickly compare multiple datasets if needed.

Each of these graphical methods serves a specific purpose and is chosen based on the data type and the kind of insight the analyst wants to gain. Histograms, frequency polygons, and ogives are best suited for continuous numerical data. Bar charts and pie charts are ideal for categorical data. Box plots provide a deeper look into data distribution and help identify variability and outliers.

When dealing with categorical or discrete data, bar charts are commonly used. In a **bar chart**, each category is represented by a separate bar, and unlike a histogram, the bars do not touch. The X-axis lists the categories (such as blood groups: A, B, AB, and O), while the Y-axis represents the frequency of each category. For example, a bar chart can display how many people in a survey belong to each blood group. Thus, Bar charts are useful for showing comparisons among **discrete categories** using rectangular bars. The height (or length) of each bar represents the value of the corresponding category.

Example: Create a bar chart for the data collected under a survey which shows the number of students preferring different programming languages:

Language	Number of Students
Python	30
Java	20
C++	10
JavaScript	15
R	5

Stepwise – Solution to create Bar chart:

Step 1: Choose Axes : Decide which variable will go on each axis:

- **X-axis:** Categories (e.g., programming languages)
- **Y-axis:** Frequencies or counts (e.g., number of students)

Step 2: Draw the Axes and Label Them : Draw two perpendicular lines:

- Label the X-axis with the category names (Python, Java, etc.).
- Label the Y-axis with a numerical scale (0 to 35, for example, with intervals of 5).

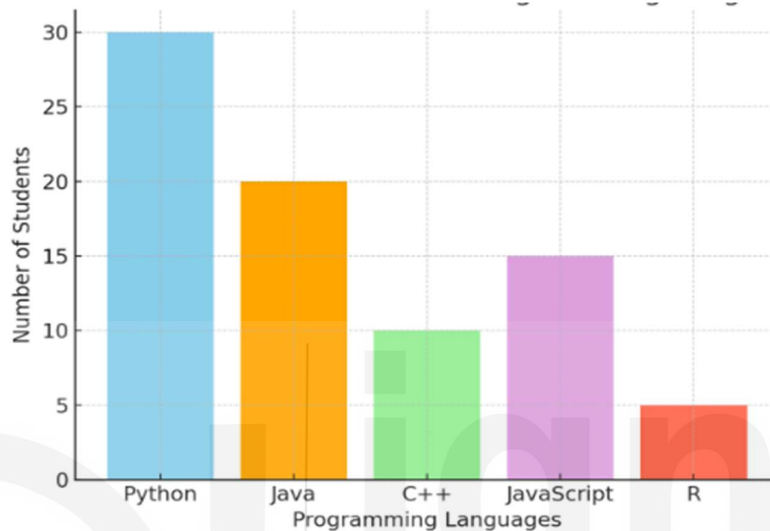
Step 3: Draw Bars : Draw a **separate vertical bar for each category**, ensuring that:

- Bars are of **equal width**
- There is **equal spacing** between bars
- Heights correspond exactly to the values

For instance:

- Python bar will be tallest (30 units high)
- R bar will be shortest (5 units high)

Step 4: Add Title and Color (Optional) :Title the chart: "*Students' Preferred Programming Languages*". Optionally, use different colors for bars to enhance readability.



Another widely used representation for categorical data is the **pie chart**, which shows data as slices of a circular pie. Each slice represents a category, and the size of the slice corresponds to the proportion of that category in the whole dataset. The angle of each slice is calculated by the formula: $(\text{Category Frequency} / \text{Total Frequency}) \times 360^\circ$.

So, a pie chart is a circular graph divided into slices to illustrate numerical proportions. It is best used when you want to show **percentage contributions** of each category to a whole.

Let's understand how to generate a pie chart, the stepwise solution to the generation of pie-chart for the above example is as follows

Stepwise – Solution to create Pie chart:

Step 1: Collect Data: Use the same example as above:

Language	Number of Students
Python	30
Java	20
C++	10
JavaScript	15
R	5

Step 2: Calculate Total: Add up all the values to get the **total number of students**.

$$\text{Total} = 30 + 20 + 10 + 15 + 5 = 80$$

Step 3: Calculate Percentage or Angle for Each Category :

To calculate the angle for each slice of the pie:

$$\text{Angle} = (\text{Value} / \text{Total}) \times 360^\circ$$

For example:

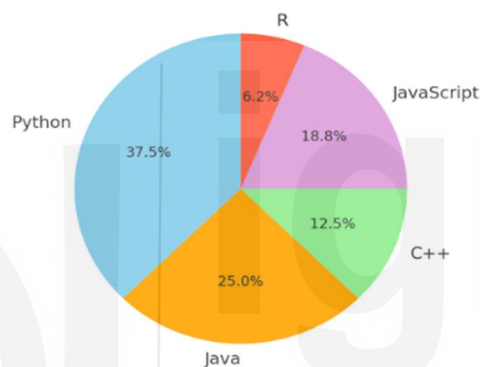
- Python: $(30/80) \times 360 = 135^\circ$
- Java: $(20/80) \times 360 = 90^\circ$
- C++: $(10/80) \times 360 = 45^\circ$
- JavaScript: $(15/80) \times 360 = 67.5^\circ$
- R: $(5/80) \times 360 = 22.5^\circ$

Step 4: Draw the Circle and Plot Slices

- Draw a circle using a compass or software.
- Use a protractor to mark each angle from the center.
- Label each slice with its category name and percentage.

Step 5: Add Title and Color (Optional)

Title it something like *"Language Preferences Among Students"*



Both Bar Chart and Pie-Chart are powerful tools for visualizing categorical data

- Use **bar charts** to compare discrete categories clearly and directly.
- Use **pie charts** to show the **proportion** each category contributes to the total.

Understanding these graphical tools is essential for effective data analysis, as they transform raw numbers into meaningful visual stories that are much easier to interpret and communicate.

1.3.5 MEASUREMENT OF CENTRAL TENDENCY

Measures of central tendency help us summarize a large amount of data using a single representative value. This makes it easier to understand and analyse the overall trend in the data. For instance, it's not practical to remember the individual incomes of millions of people in India. However, knowing the **average income** gives us a general idea of the income level of the population as a whole.

These measures also make it easier to **compare different data sets**. For example, the **average sales** in April can be compared with those of previous months to understand whether performance has improved or declined.

A **good measure of central tendency** is one that effectively represents a dataset with a single, meaningful value. To be truly useful, such a measure should have certain key characteristics. First, it should be **easy to understand**, so that people without technical backgrounds can interpret it. It should

also be **simple to calculate**, using basic mathematical operations. A reliable measure must be **based on all the observations** in the dataset, not just a few selected ones, to ensure it truly reflects the data. Additionally, the measure should be **uniquely defined**, meaning that it always gives the same result for a given dataset. It should also allow for **further algebraic treatment**, so it can be used in more complex analyses and calculations. Finally, it should be **resistant to extreme values (outliers)**; that is, it should not be heavily influenced by unusually large or small numbers that might distort the overall picture. In practice, several types of measures of central tendency are commonly used in fields like **business, economics, and industry**. These include:

- **Arithmetic Mean** – the most common average.
- **Weighted Arithmetic Mean** – used when different values have different levels of importance.
- **Median** – the middle value when the data is arranged in order.
- **Quartiles** – values that divide the data into equal-sized intervals, such as quartiles or percentiles.
- **Mode** – the value that appears most frequently.
- **Geometric Mean** – useful in cases involving rates of growth or ratios.
- **Harmonic Mean** – often applied in averaging ratios or rates, like speed.

Understanding these measures and their characteristics helps in selecting the right one depending on the nature and context of the data.

Let's Begin with **Calculation of the Central Tendency for the Ungrouped Data:**

1. Arithmetic Mean: The arithmetic mean, often referred to as the mean or average, is the most widely used and easily understood measure of central tendency. In statistical terms, the word "average" generally refers to any value that represents the center of a data set. Specifically, the arithmetic mean is calculated by adding all the values in the dataset and then dividing the total by the number of observations. It can be expressed symbolically as follows:

$$\bar{X} = \frac{\sum x}{N}$$

Here, $\sum x$ represents the total of all observed values, while N stands for the number of observations in the dataset.

Let's take an example to understand this better. Suppose The marks obtained by a student in five subjects are: 60, 70, 80, 90, 100

To find the **arithmetic mean**, we add all the marks and divide by the number of students

Thus, the arithmetic mean is $\bar{X} = (60 + 70 + 80 + 90 + 100) / 5 = 400 / 5 = 80$ i.e. the average marks are 80

2. Weighted Arithmetic Mean: Used when some values carry more importance or weight than others.

Example:

Grades and credit weights: Math (90, weight 4), Science (80, weight 3), English (70, weight 2)

Weighted Mean = $(90 \times 4 + 80 \times 3 + 70 \times 2) / (4 + 3 + 2) = 740 / 9 \approx 82.22$

3. **Median:** The median is the middle value in an ordered dataset.

Example 1: Data: 12, 15, 20, 25, 30 → Median = 20

Example 2: Data: 10, 20, 30, 40 → Median = $(20 + 30)/2 = 25$

4. **Quartiles:** Quartiles divide ordered data into four equal parts.

Example: Data: 10, 20, 30, 40, 50, 60, 70, 80

Q1 = 25, Q2 (Median) = 45, Q3 = 65

5. **Mode:** The mode is the value that appears most frequently.

Example: Data: 3, 4, 4, 5, 6, 6, 6, 7 → Mode = 6

6. **Geometric Mean:** Useful for multiplicative data like growth rates.

Example: Growth rates: 10%, 20%, 30% → Convert to factors: 1.10, 1.20, 1.30

Geometric Mean = $\sqrt[3]{1.10 \times 1.20 \times 1.30} \approx \sqrt[3]{1.716} \approx 1.19$ or 19% growth per year

7. **Harmonic Mean:** Best used when averaging rates like speed.

Example: Speed to and from a place: 60 km/h and 40 km/h

Harmonic Mean = $2 \times 60 \times 40 / (60 + 40) = 4800 / 100 = 48$ km/h

So far, we have learned how to calculate the arithmetic mean for **ungrouped data**. Now, let's understand how this changes when dealing with **grouped data**. In a frequency distribution, the data values are grouped into class intervals. Since we don't know the exact values within each class, we use the **midpoint** of each class interval as a representative value for that class. Based on this approach, the **arithmetic mean for grouped data** is calculated using these midpoints. It is defined as follows:

$$\bar{X} = \frac{\sum f \cdot x}{N}$$

Where X is midpoint of various classes, f is the frequency for corresponding class and N is the total frequency, i.e. $N = \sum f$.

1.3.6 MEASUREMENT OF DISPERSION

In above section 1.3.5 we learned about the measure of central tendency of data, now we extend our discussion towards the understanding of the dispersion of data and how it is measured. It is to be noted that the degree to which numerical data tend to spread about an average value is called the *variation or dispersion of data*.

Actually, there are two basic kinds of a measure of dispersion

- (i) Absolute measures and
- (ii) Relative measures.

The *absolute measures of dispersion* are used to measure the variability of a given data expressed in the same unit, while the *relative measures* are used to compare the variability of two or more sets of observations.

Following are the different measures of dispersion:

- 1. Range
- 2. Quartile Deviation

3. Mean Deviation

4. Standard Deviation and Variance

*Before extending our discussion for the different measures of dispersion, firstly we need to understand the **significance of Measures of dispersion**, actually they are needed for the following purposes:*

- They show how much the data values vary or spread around the average.
- They help determine how reliable or representative an average value is.
- Less variation means the average represents the data well; more variation means it may be unreliable.
- They help identify the causes of variation so that it can be controlled (e.g., quality control in production).
- They allow comparison of two or more data sets in terms of consistency or uniformity.
- Smaller dispersion indicates greater uniformity and consistency in data.
- They form the basis for many statistical methods such as correlation, hypothesis testing, analysis of variance, and quality control techniques.

Further The characteristics of a good measure of dispersion are similar to those of a good measure of central tendency. Thus, it is reiterated that an ideal measure of dispersion should have the following qualities:

- It should be easy to understand.
- It should be simple to calculate.
- It should be clearly and precisely defined.
- It should consider all observations in the dataset.
- It should be suitable for further algebraic or mathematical analysis.
- It should remain stable across different samples.
- It should not be excessively influenced by extreme values.

Let's begin our discussion on the different measures of dispersion i.e. Range, Quartile Deviation, Mean Deviation and Standard Deviation and Variance; range is the first one to begin with

1. Range: Range is the simplest measure of dispersion. It represents the difference between the maximum and minimum values in a dataset.

$$\text{Range} = \text{Maximum value } (X_{\max}) - \text{Minimum value } (X_{\min})$$

Example: for the Data: 5, 10, 15, 20, 25; the maximum value is 25 and the minimum value is 5 thus the Range = $25 - 5 = 20$

The Merits and Demerits of Range are as follows :

Merits of Range

1. It is the simplest to understand;

2. It can be visually obtained since one can detect the largest and the smallest observations easily and can take the difference without involving much calculations; and
3. Though it is crude, it has useful applications in areas like order statistics and statistical quality control.

Demerits of Range

1. It utilizes only the maximum and the minimum values of variable in the series and gives no importance to other observations;
 2. It is affected by fluctuations of sampling;
 3. It is not very suitable for algebraic treatment;
 4. If a single value lower than the minimum or higher than the maximum is added or if the maximum or minimum value is deleted range is seriously affected; and
 5. Range is the measure having units of the variable and is not a pure number.
- That's why sometimes coefficient of range is calculated by

$$\text{Coefficient of Range: } \frac{(X_{\max} - X_{\min})}{(X_{\max} + X_{\min})}$$

2. Quartile Deviation: Quartile Deviation measures dispersion using the middle 50% of the data. It is also called the Semi-Interquartile Range. $(Q_3 - Q_1)$ gives the inter quartile range. The semi inter quartile range which is also known as Quartile Deviation (QD) is given by

$$\text{Quartile Deviation (QD)} = (Q_3 - Q_1) / 2$$

Relative measure of Q.D. known as Coefficient of Q.D. and is defined as

$$\text{Coefficient of QD} = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$$

Example-1: For the Data: 2, 4, 6, 8, 10, 12, 14; the first Quartile $Q_1 = 4$, and the third Quartile are $Q_3 = 12$. Therefore, $QD = (12 - 4) / 2 = 4$

Example-2: Find the Quartile Deviation for the following data.

Class Interval	Frequency
0–10	3
10–20	5
20–30	7
30–40	9
40–50	4

Solution:

<u>Class Interval</u>	<u>Frequency (f)</u>	<u>Cumulative Frequency (cf)</u>
0 – 10	3	3
10 – 20	5	8
20 – 30	7	15
30 – 40	9	24
40 – 50	4	28

Step 1: Calculate Total Frequency (N)= 3+5+7+9+4 = 28

Step 2: Determine Quartile Positions for $Q_1 = N/4 = 7$ and for $Q_3 = 3N/4 = 21$

Step 3: Determine Quartile Classes

Q_1 class → Cumulative frequency just greater than 7 → 10–20

Q_3 class → Cumulative frequency just greater than 21 → 30–40

i.e. Q_1 class = 10–20 and Q_3 class = 30–40

Step 4: We know the Formula for Quartile (grouped data) is $Q = L + ((kN/4 - cf) / f) \times h$

Where:

L = lower limit of quartile class

cf = cumulative frequency before quartile class

f = frequency of quartile class

h = class width

Step 5: Now Calculate Q_1 and Q_3

Calculate Q_1 : For class 10–20 we have $L = 10, cf = 3, f = 5, h = 10$

Therefore $Q_1 = 10 + ((7 - 3) / 5) \times 10 = 18$

Calculate Q_3 : For class 30–40 we have $L = 30, cf = 15, f = 9, h = 10$

Therefore $Q_3 = 30 + ((21 - 15) / 9) \times 10 = 36.67$

Step 6: Finally calculate Quartile Deviation (QD) = $(Q_3 - Q_1) / 2$

Quartile Deviation (QD) = $(36.67 - 18) / 2 = 9.34$

3. Mean Deviation:

Mean deviation is defined as the average of the sum of the absolute values of deviations from any arbitrary value such as mean, median, mode, etc. It is often suggested to calculate mean deviation from the median because it gives the least value when measured from the median.

The deviation of an observation X_i from the assumed mean A is defined as: $(x_i - A)$

Therefore, mean deviation can be defined as: $MD = (1/n) \sum |x_i - A|$

Note : It is to be noted that the quantity $\sum |x_i - A|$ is minimum when A is the median.

Mean deviation from mean is defined as: $MD = (1/n) \sum |x_i - \bar{x}|$

Example: Given Data: 2, 4, 6, 8 has Mean = 5,

therefore, Mean Deviation = $(3 + 1 + 1 + 3) / 4 = 2$

Mean deviation from median is defined as: $MD = (1/n) \sum |x_i - \text{Median}|$

For a frequency distribution, the formulae are:

Mean Deviation from Mean: $MD = \sum f_i |x_i - \bar{x}| / \sum f_i$

Mean Deviation from Median: $MD = \sum f_i |x_i - \text{Median}| / \sum f_i$

where all symbols have their usual meanings.

Example: For the data given below,

x:	1	2	3	4	5	6	7
f:	3	5	8	12	10	7	5

Determine

(i) Mean deviation from mean and

(ii) Mean deviation from Median

Solution :

(i) Mean Deviation from Mean

Step 1: Calculate Mean

$$\sum f = 50$$

$$\sum fx = (1 \times 3) + (2 \times 5) + (3 \times 8) + (4 \times 12) + (5 \times 10) + (6 \times 7) + (7 \times 5) = 3 + 10 + 24 + 48 + 50 + 42 + 35 = 212$$

$$\text{Therefore, Mean } (\bar{x}) = \sum fx / \sum f = 212 / 50 = 4.24$$

Step 2: Calculate $|x - \bar{x}|$ and $f|x - \bar{x}|$

x	f	$ x - 4.24 $	$f x - 4.24 $
1	3	3.24	9.72
2	5	2.24	11.20
3	8	1.24	9.92
4	12	0.24	2.88
5	10	0.76	7.60
6	7	1.76	12.32
7	5	2.76	13.80

$$\text{Find } \sum f |x - \bar{x}| = 67.44$$

$$\text{Therefore, Mean Deviation from Mean } MD(\text{mean}) = \sum f |x - \bar{x}| / \sum f = 67.44 / 50 = 1.348$$

(ii) Mean Deviation from Median

Step 1: Find Median

Compute Cumulative Frequency:

x:	1	2	3	4	5	6	7
f:	3	5	8	12	10	7	5
cf:	3	8	16	28	38	45	50

$$\text{As } N = 50 \rightarrow N/2 = 25$$

Thus, the 25th observation lies at $x = 4$

Hence, Median = 4

Step 2: Calculate $|x - \text{Median}|$ and $f |x - \text{Median}|$

x	f	$ x-4 $	$f x-4 $
1	3	3	9
2	5	2	10
3	8	1	8
4	12	0	0
5	10	1	10
6	7	2	14
7	5	3	15

As $\Sigma f |x - \text{Median}| = 66$

Mean Deviation from Median

$$\begin{aligned}\text{MD}(\text{median}) &= \Sigma f |x - \text{Median}| / \Sigma f \\ &= 66 / 50 \\ &= 1.32\end{aligned}$$

Thus, we got, Mean Deviation from Mean = 1.348 and Mean Deviation from Median = 1.32

The mean deviation also possesses some merits and demerits; they are as follows:

Merits of Mean Deviation:

1. It utilizes all the observations.
2. It is easy to understand and calculate.
3. It is not much affected by extreme values.

Demerits of Mean Deviation:

1. Negative deviations are converted into positive values.
2. It is not suitable for algebraic treatment.
3. It cannot be calculated for open-end classes.

4. Standard Deviation and Variance

Variance and Standard Deviation are important measures of dispersion used in statistics and data science. They describe how much the data values deviate from the mean.

formally, Variance is the average of the squares of deviations of observations from the arithmetic mean, and Standard deviation is the positive square root of variance.

The formula for variance and standard deviation are as follows :

$$\text{Formula (Ungrouped Data): Variance } (\sigma^2) = \Sigma (x - \bar{x})^2 / n$$

$$\text{Formula: (Ungrouped Data): Standard deviation } (\sigma) = \sqrt{[\Sigma (x - \bar{x})^2 / n]}$$

$$\text{Formula (Frequency Distribution): Variance } (\sigma^2) = \Sigma f (x - \bar{x})^2 / \Sigma f$$

Formula (Frequency Distribution): Standard deviation (σ) = $\sqrt{[\Sigma f(x - \bar{x})^2 / \Sigma f]}$

Example (Ungrouped Data) : Determine the variance in the data: 2, 4, 6, 8

Solution:

Step 1: Calculate Mean ($\bar{x}$) = (2 + 4 + 6 + 8) / 4 = 5

Step 2: Find Deviations

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-3	9
4	-1	1
6	1	1
8	3	9

Calculate $\Sigma(x - \bar{x})^2 = 20$

Step 3: Find Variance (σ^2) = $\Sigma(x - \bar{x})^2 / n$

$$\sigma^2 = 20 / 4 = 5$$

If Variance = 5 then Standard Deviation (σ) = $\sqrt{5} \approx 2.24$

INTERPRETATION: Standard deviation shows that the data values deviate from the mean by about 2.24 units.

Above we have discussed how to calculate the variance for a single variable. If there are two or more populations and the information about the means and variances of those populations are available then we can obtain the combined variance of several populations.

If $n_1, n_2, \dots, n_k$ are the sizes $x_1, x_2, \dots, x_k$ are the means and $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$ are the variances of k populations, then the combined variance is given by the Formula for Variance of Combined Series:

$$\sigma^2 = \frac{[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + \dots + n_k(\sigma_k^2 + d_k^2)]}{(n_1 + n_2 + \dots + n_k)}$$

Where:

$$d_i = \bar{x}_i - \bar{x}$$

$$\text{Mean of Combined Series } (\bar{x}) = (\Sigma n_i \bar{x}_i) / (\Sigma n_i)$$

Here:

σ^2 = variance of the combined series

σ_i^2 = variance of the i -th series

n_i = number of observations in the i^{th} series

$\bar{x}_i$ = mean of the i^{th} series

$\bar{x}$ = mean of the combined series

d_i = deviation of individual mean from combined mean

This formula is used when the means and variances of two or more series are known and the combined variance is required.

Example: Suppose a series of 100 observations has mean 50 and variance 20. Another series of 200 observations has mean 80 and variance 40. Find the combined variance of the given series.

Solution:

Series	(n)	Mean	Variance
1	100	50	20
2	200	80	40

Given:

First Series: $n_1 = 100$; $\bar{x}_1 = 50$; $\sigma_1^2 = 20$

Second Series: $n_2 = 200$; $\bar{x}_2 = 80$; $\sigma_2^2 = 40$

$$\begin{aligned}
 \text{Step 1: Calculate Combined Mean } \bar{x} &= (n_1\bar{x}_1 + n_2\bar{x}_2) / (n_1 + n_2) \\
 &= (100 \times 50 + 200 \times 80) / 300 \\
 &= (5000 + 16000) / 300 \\
 &= 21000 / 300 = 70
 \end{aligned}$$

Step 2: Calculate Deviations from Combined Mean

$$d_1 = \bar{x}_1 - \bar{x} = 50 - 70 = -20$$

$$d_1^2 = 400$$

$$d_2 = \bar{x}_2 - \bar{x} = 80 - 70 = 10$$

$$d_2^2 = 100$$

Step 3: Apply Combined Variance Formula

$$\sigma^2 = [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)] / (n_1 + n_2)$$

$$\sigma^2 = [100(20 + 400) + 200(40 + 100)] / 300$$

$$\sigma^2 = (42000 + 28000) / 300$$

$$\sigma^2 = 70000 / 300$$

$$\sigma^2 = 233.33$$

Thus, Combined Variance = 233.33

The Variance and Standard deviation also possess some merits and demerits; they are as follows:

Variance:

- **Merits:** Variance is a precisely defined measure of dispersion that considers every observation in the dataset, ensuring comprehensive coverage of the data. Its mathematical formulation makes it suitable for algebraic manipulation and further statistical analysis. By squaring deviations, it effectively eliminates the issue of negative values and appropriately emphasizes larger deviations. Owing to these properties, variance has become one of the most widely used and accepted measures of dispersion in statistical studies.

- **Demerits:** Variance may not be appropriate in situations where the mean is not a suitable measure of central tendency, such as in distributions with open-ended classes; in such cases, quartile deviation is often preferred. It is not a unit-free measure, and its computation can be tedious for large datasets, often requiring the use of calculators or computers. Additionally, since variance is expressed in squared units of the original variable, interpreting the magnitude of dispersion becomes less intuitive compared to standard deviation.

Standard Deviation:

- **Merits:** Standard deviation is a well-defined and reliable measure of dispersion that takes into account all observations in a dataset. It is mathematically tractable, allowing for extensive algebraic and statistical applications. The squaring of deviations helps remove negative signs while preserving the influence of extreme values. Because it is expressed in the same unit as the original data, standard deviation is widely regarded as the most practical and popular measure of dispersion.
- **Demerits:** Standard deviation may be unsuitable when the mean does not adequately represent the data, particularly in the presence of open-ended classes, where quartile deviation can be more appropriate. It is not a unit-free measure, which limits comparisons across datasets with different units. Although conceptually straightforward, its calculation for large or complex datasets can be time-consuming and typically requires computational tools.

Note: Standard deviation is more suitable for algebraic and mathematical operations than mean deviation. Mean deviation directly ignores the signs of deviations, which results in the loss of information carried by positive and negative differences. In contrast, standard deviation retains the effect of signs initially, though they are removed automatically through squaring. However, mean deviation is still preferable in situations where the median is a better measure of central tendency than the mean.

There are two more aspects of dispersion one is Root Mean Square Deviation and Other is Coefficient of Variation, both are discussed below:

Root Mean Square Deviation (RMSD): As discussed earlier, standard deviation is the positive square root of the average of the squares of deviations taken from the mean. But, if the deviations are taken from an assumed mean instead of the actual mean, the resulting measure is called *Root Mean Square Deviation*.

Formula (Ungrouped Data): $\text{RMSD} = \sqrt{[\sum (x_i - A)^2 / n]}$

where: A = assumed mean and n = number of observations

For a frequency distribution, the formula is: $\text{RMSD} = \sqrt{[\sum f_i (x_i - A)^2 / \sum f_i]}$

Note: When the assumed mean is equal to the actual mean ($A = \bar{x}$), the Root Mean Square

Deviation becomes equal to the Standard Deviation.

The Coefficient of Variation (CV) is defined as a relative measure of dispersion used to compare the variability of two or more data series. The data series having a smaller coefficient of variation is considered more consistent.

Formula: $CV = (\sigma / \bar{x}) \times 100$

Note: One should be careful while interpreting the coefficient of variation.

For example, the series 10, 10, 10 has zero standard deviation and hence CV is also zero. Similarly, the series 50, 50, 50 also has zero standard deviation and CV equal to zero. Although both series have the same CV, the second series has a higher mean and is therefore considered more consistent.

Example: Suppose batsman A has mean 50 with Standard deviation (SD) 10. Batsman B has mean 30 with Standard deviation (SD) 3. What do you infer about their performance?

Solution: To compare the performance and consistency of the two batsmen, we use the Coefficient of Variation (CV). We Know, $CV = \frac{\text{Standard Deviation}}{\text{Mean}}$

Calculations for Coefficient of Variation for Batsman A & B:

- Coefficient of Variation for Batsman A (CV_A). $= \frac{10}{50} = 0.20$
- Coefficient of Variation for Batsman B (CV_B). $= \frac{3}{30} = 0.10$

Interpretation:

- Batsman A has a higher mean score (50), so he is a better run scorer.
- Batsman B has a lower coefficient of variation (0.10) compared to A (0.20), so he is more consistent.

Final Inference:

- Batsman A is the better performer in terms of average runs.
- Batsman B is the more consistent batsman.

1.3.7 SUMMARIZATION OF DATA

Descriptive summarization of data refers to the process of organizing, simplifying, and presenting raw data in a meaningful manner so that its essential characteristics can be clearly understood. Instead of focusing on individual observations, descriptive summarization highlights overall patterns, typical values, variability, and the general nature of the data

distribution. It forms the foundation of statistical analysis and is widely used in data science to gain an initial understanding of datasets.

The main parameters used for descriptive summarization are broadly grouped into the following categories.

1. Measures of Central Tendency
2. Measures of Dispersion (Variability)
3. Measures of Position (Location)
4. Measures of Shape of Distribution
5. Measures of Relative Variability
6. Frequency-Based Parameters
7. Graphical & Visual Summaries

In the above sections we have already discussed in detail the parameters under these categories. Although a brief discussion on the parameters in each category is given below:

1. Measures of Central Tendency

Measures of central tendency describe the central or most representative value of a dataset. The mean provides the arithmetic average of all observations and is useful for numerical data that does not contain extreme values, although it is sensitive to outliers. The median represents the middle value of an ordered dataset and is more reliable when the data is skewed or contains extreme values. The mode indicates the most frequently occurring value and is especially useful for categorical data or for identifying the most common observation. Together, these measures offer a single value that summarizes the centre of the data.

The parameters listed below, Measures of Central Tendency and describe the central or typical value around which the data is clustered.

(a) Mean

- The arithmetic average of all observations.
- Useful for numerical data without extreme values.
- Sensitive to outliers.

(b) Median

- The middle value when data is arranged in ascending or descending order.
- More reliable for skewed distributions or data with extreme values.

(c) Mode

- The most frequently occurring value.
- Useful for categorical data and identifying the most common observation.

The purpose of Measures of central tendency is to provide a single representative value for the dataset.

2. Measures of Dispersion (Variability)

Measures of dispersion explain how far the data values are spread around the central value. The range gives a basic idea of spread by calculating the difference between the maximum and minimum values. Quartile deviation focuses on the dispersion of the middle 50% of the data

and is less affected by extreme values. Mean deviation calculates the average of absolute deviations from a central value and is simple to understand, though less useful for advanced analysis. Variance measures the average of squared deviations from the mean and plays a significant role in theoretical and statistical modeling. Standard deviation, being the square root of variance, is the most widely used measure of dispersion because it is expressed in the same unit as the data. These measures help evaluate consistency, reliability, and variability within the dataset.

The parameters listed below, describe Measures of Dispersion (Variability) i.e. how spread out the data values are around the central value.

(a) Range

- Difference between maximum and minimum values.
- Provides a quick idea of data spread.

(b) Quartile Deviation (Interquartile Range)

- Measures dispersion of the middle 50% of data.
- Less affected by extreme values.

(c) Mean Deviation

- Average of absolute deviations from mean or median.
- Simpler to understand but less useful for advanced analysis.

(d) Variance

- Average of squared deviations from the mean.
- Used heavily in theoretical and statistical modeling.

(e) Standard Deviation

- Square root of variance.
- Most widely used measure of dispersion in data science.

The Purpose of Measures of dispersion is to help assess consistency, reliability, and variability of data.

3. Measures of Position (Location)

Measures of position indicate the relative standing of individual data values within a dataset. Quartiles divide the data into four equal parts, deciles divide it into ten equal parts, and percentiles divide it into one hundred equal parts. These measures help in understanding how a particular observation compares with the rest of the data and are useful in ranking and comparative analysis.

The parameters listed below, describe Measures of Position (Location) and indicate the relative position of a data value within a dataset

(a) Quartiles

- Divide data into four equal parts (Q1, Q2, Q3).

(b) Deciles

- Divide data into ten equal parts.

(c) Percentiles

- Divide data into one hundred equal parts.

The Purpose of Measures of position is to provide help to compare individual observations relative to the dataset.

4. Measures of Shape of Distribution

Measures of shape describe the overall form and pattern of the data distribution. Skewness indicates the degree of asymmetry in the distribution, showing whether the data is skewed to the right or left. Kurtosis measures the peakedness or flatness of the distribution and provides insight into the presence of extreme values. Together, these measures offer a deeper understanding of the nature of the data beyond central values and variability.

The parameters listed below, describe Measures of Shape of Distribution and describe the pattern and form of the data distribution followed by the data in any dataset

(a) Skewness

- Indicates the **symmetry or asymmetry** of the distribution.
- Positive skew: longer right tail
- Negative skew: longer left tail

(b) Kurtosis

- Measures the **peakedness or flatness** of the distribution.
- Helps identify the presence of extreme values.

The Purpose of Measures of shape is to provide insight into the nature of data distribution beyond averages.

5. Measures of Relative Variability

Measures of relative variability allow comparison between datasets that differ in units or scale. The coefficient of variation, which relates the standard deviation to the mean, is a unit-free measure that helps determine which dataset is more consistent or stable. This is particularly useful when comparing variability across different contexts.

The parameters listed below, describe Measures of Relative Variability and allow comparison between datasets with different units or scales.

Coefficient of Variation (CV)

- Ratio of standard deviation to mean.
- Unit-free measure of consistency.

The Purpose of Relative measures is to provide help to identify which dataset is more consistent or stable.

6. Frequency-Based Parameters

Frequency-based parameters summarize how often data values occur. These include frequency distributions, relative frequencies, and cumulative frequencies. Such parameters help reveal patterns of data concentration, distribution trends, and the overall structure of the dataset.

Frequency-Based Parameters summarize how often values occur, and they include following:

- Frequency tables
- Relative frequency
- Cumulative frequency

The Purpose of Frequency-Based Parameters is to help in understanding the distribution patterns and data concentration.

7. Graphical and Visual Summaries

Graphical and visual summaries complement numerical measures in descriptive summarization. Tools such as bar charts, histograms, pie charts, box plots, and line graphs visually represent data patterns, trends, and outliers. These visual techniques make complex datasets easier to interpret and communicate.

Although Graphical and Visual Summaries are numerical in nature, descriptive summarization is often supported by visual tools like:

- Histograms
- Bar charts
- Pie charts
- Box plots
- Line graphs

The Purpose of Graphical and Visual summaries is to make patterns, trends, and outliers easier to interpret.

Thus, we learned that the Descriptive summarization of data relies on a combination of measures of central tendency, dispersion, position, shape, relative variability, and frequency, supported by graphical representations. Together, these parameters provide a comprehensive statistical overview of a dataset, enabling better understanding, comparison, and informed decision-making in statistics and data science.

1.4 SUMMARY

This Unit-1 titled “Descriptive Data Analysis” is the foundational step in understanding any dataset before applying advanced statistical techniques. It begins with data collection, where information is gathered either as primary data (collected directly) or secondary data (already available). Once collected, it is essential to identify the type of data—qualitative (nominal, ordinal) or quantitative (discrete, continuous, interval, ratio)—as this determines the

appropriate methods of analysis. The data is then organized using frequency distributions, which help in identifying patterns, concentrations, and irregularities within the dataset.

To make data more understandable, it is presented through graphical representations such as bar charts, histograms, pie charts, and line graphs, enabling quick visual interpretation of trends and patterns. Further, the dataset is summarized using measures of central tendency like mean, median, and mode, which provide insights into the typical or average values. Alongside this, measures of dispersion such as range, variance, and standard deviation help in understanding the spread or variability of the data, indicating how much the values deviate from the average.

Finally, data summarization combines these techniques to present information in a concise and meaningful way through tables, graphs, and key statistical values. This process allows analysts to draw initial insights, detect patterns, and prepare the data for further analysis. Overall, mastering these concepts equips learners with essential skills to interpret data accurately and make informed decisions in academic, research, and real-world contexts.

1.5 SOLUTIONS/ANSWERS

☛ Check Your Progress 1

Q1. Explain the concept of Descriptive Data Analysis. Why is it important in data analysis?

Answer: Refer to section 1.3

Q2. Classify the types of data with suitable examples.

Answer: Refer to section 1.3

Q3. What are Measures of Central Tendency and Dispersion? Explain briefly.

Answer: Refer to section 1.3

☛ Check Your Progress 2

Q1. What is Data Collection? Why is it considered a crucial step in data analysis?

Answer: Refer to section 1.3.1

Q2. Differentiate between Primary Data and Secondary Data.

Answer: Refer to section 1.3.1

Q3. Explain any three methods of collecting Primary Data with examples.

Answer: Refer to section 1.3.1

☛ Check Your Progress 3

Q1. Explain the four levels of measurement in data analysis i.e. Nominal, Ordinal, Ratio & interval data. Why are they important?

Answer: Refer to section 1.3.2

1.6 FURTHER READINGS

1. *Handbook of Statistical Analysis and Data Mining Applications* by Robert Nisbet, John Elder, and Gary Miner, published by Elsevier (2nd Edition, 2017, ISBN: 978-0124166455).
2. *Data Mining and Analysis: Fundamental Concepts and Algorithms* by Mohammed J. Zaki and Wagner Meira Jr., published by Cambridge University Press (1st Edition, 2014, ISBN: 978-0521766333).
3. *Core Concepts in Data Analysis: Summarization, Correlation and Visualization* by Boris Mirkin, published by Springer (1st Edition, 2011, ISBN: 978-0857292872).
4. *Applied Predictive Modeling*, Max Kuhn and Kjell Johnson, 1st Edition, Springer, 2013.
5. *An Introduction to Statistical Learning with Applications in R*, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 1st Edition, Springer, 2013.
6. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2nd Edition, Springer, 2009.
7. *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, published by Springer (2nd Edition, 2021, ISBN: 978-1071614174).
8. *Time Series Analysis: Forecasting and Control* by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung, published by Wiley (5th Edition, 2015, ISBN: 978-1118675021)