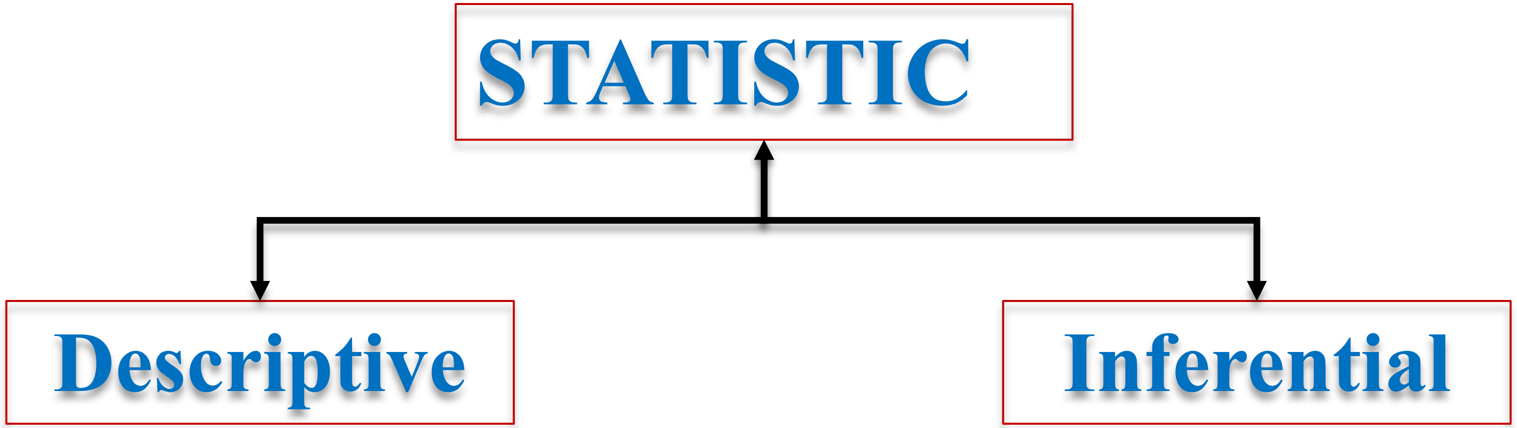
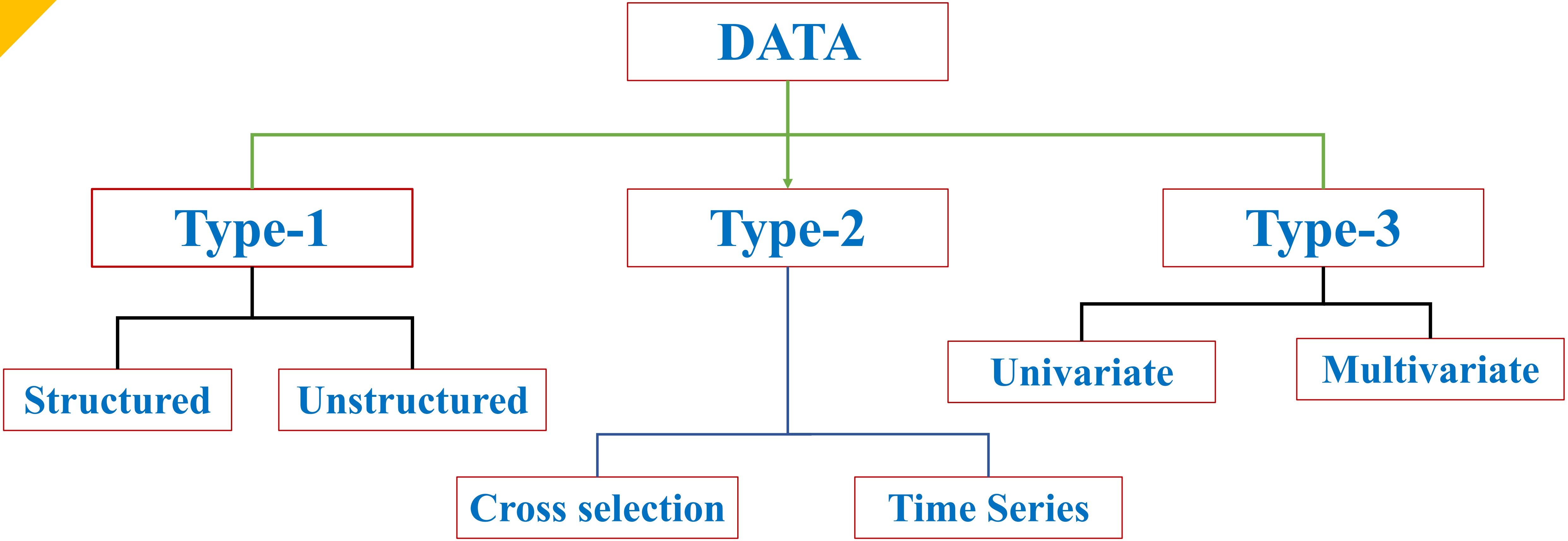
# DESCRIPTIVE STATISTIC



STATISTIC

Descriptive

Inferential



DATA

Type-1

Type-2

Type-3

Structured

Unstructured

Cross selection

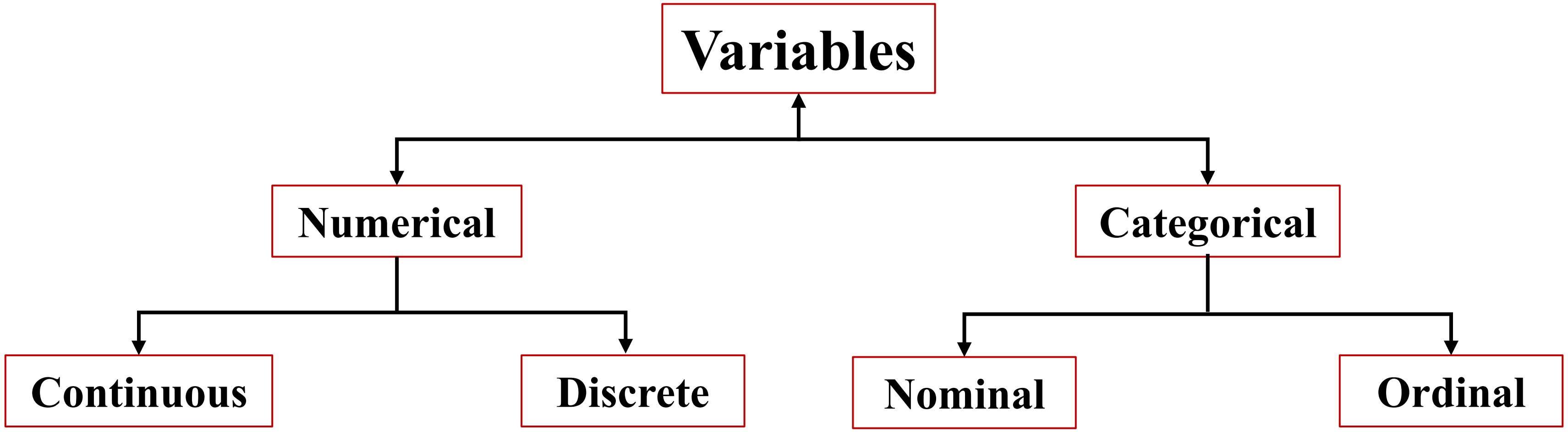
Time Series

Univariate

Multivariate

# Variables :

a **variable** is any characteristic, number, or quantity that can be measured or counted. Variables can vary from one individual or item to another and are fundamental to data collection and analysis.



## Types of Variables in Statistics

### 1.Quantitative Variables (Numerical Variables)

1. **Definition**: Variables that represent measurable quantities and are expressed as numbers. These variables answer questions like "how much?" or "how many?“

2. **Types**:

   1. **Discrete Variables**:

      1. Take on a countable number of distinct values.
      2. Often represent counts of items.
      3. Example: Number of children in a family, number of cars in a parking lot.

   2. **Continuous Variables**:

      1. Can take any value within a given range and are often measured.
      2. Example: Height, weight, temperature, time.

## 1.Qualitative Variables (Categorical Variables)

1. **Definition**: Variables that represent categories or groups and are not inherently numerical. These variables answer questions like "what type?" or "which category?“

2. **Types**:

   1. **Nominal Variables**:

      1. Categories with no natural order or ranking.
      2. Example: Gender (male, female), blood type (A, B, AB, O), colors (red, blue, green).

   2. **Ordinal Variables**:

      1. Categories with a meaningful order or ranking, but the differences between categories are not measurable.
      2. Example: Education level (high school, bachelor’s, master’s, PhD), customer satisfaction (satisfied, neutral, dissatisfied).

# Population & Sample:

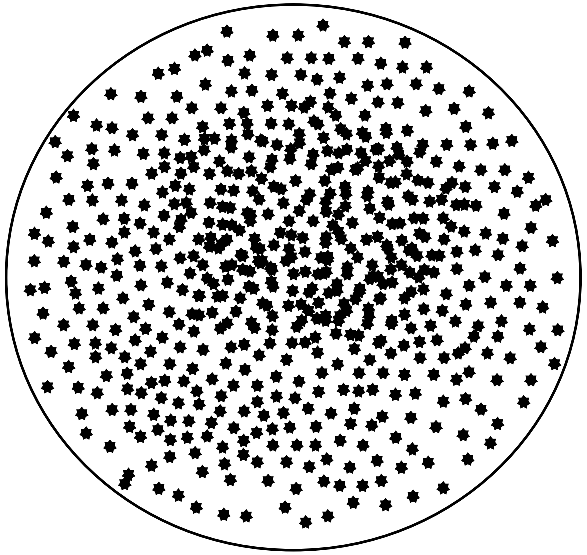
**Population** and **Sample** are fundamental concepts in statistics:

**1.Population**:

1. The population refers to the entire set of individuals, items, or data points that you are interested in studying or drawing conclusions about.

1. It includes every member of the group that you are examining, whether it’s people, objects, events, etc.

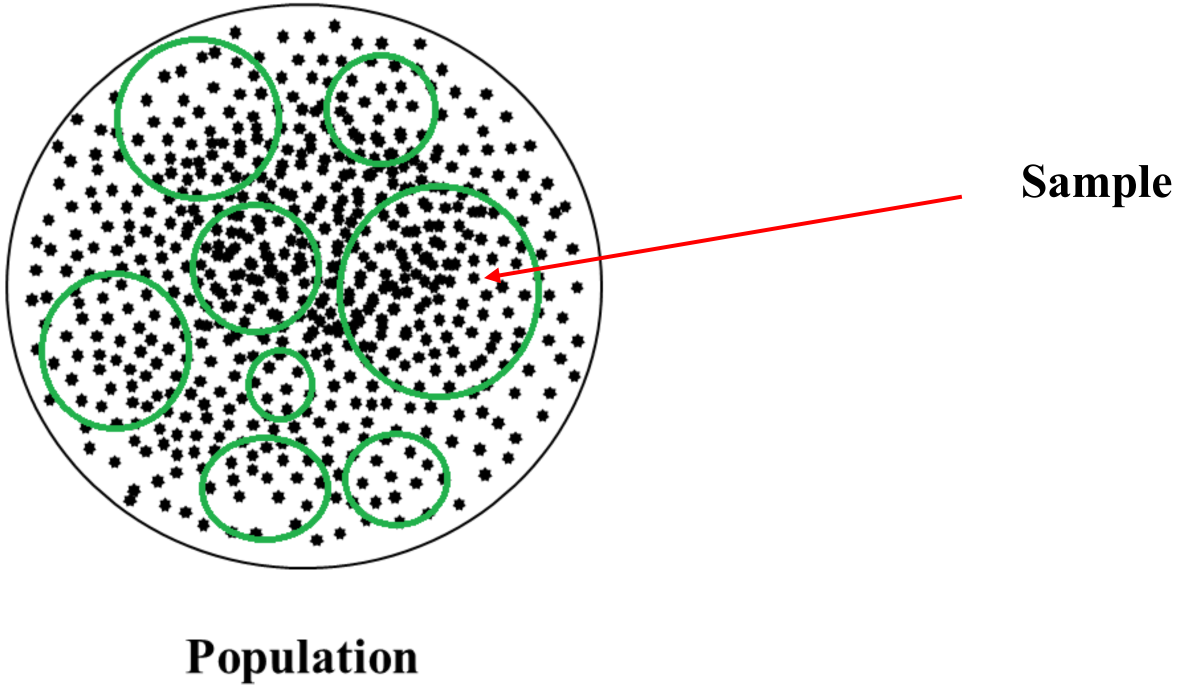
2. For example, if you’re studying the average height of adults in a country, the population would include all adults in that country.



**Population**

# Sample:

- A sample is a subset of the population that is selected for the actual study.
- The purpose of selecting a sample is to make inferences about the population, as studying the entire population may be impractical or impossible.
- For example, instead of measuring the height of every adult in a country, you might measure the height of a representative group of 1,000 adults. This group would be your sample.

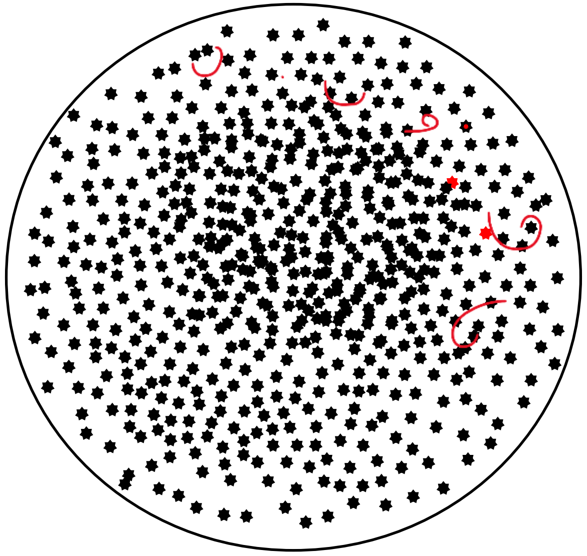


- **Example 1: Studying College Students**
- **Population**: All college students in the United States.
- **Sample**: A group of 500 college students selected from 10 different universities across the United States.
- **Example 2: Testing a New Medication**
- **Population**: All people suffering from a specific medical condition (e.g., diabetes).
- **Sample**: 200 patients with diabetes who are selected to participate in a clinical trial for the new medication.
- **Example 3: Survey on Customer Satisfaction**
- **Population**: All customers who purchased a product from an online store in the last year.
- **Sample**: 1,000 customers randomly selected from the store's customer database to fill out a satisfaction survey.
- **Example 4: Measuring Air Quality**
- **Population**: All air particles in a city over the course of a year.
- **Sample**: Air samples collected from 50 different locations within the city over a period of one week.
- **Example 5: Voting Intentions in an Election**
- **Population**: All eligible voters in a country.
- **Sample**: 2,000 voters randomly selected from different regions to participate in a pre-election poll.

## Simple Random Sampling

- Every individual in the population(N) has an equal chance of being selected.

- Example: Drawing names from a hat where each name has an equal chance of being picked.



Population

## Systematic Sampling

- Selecting every nth individual from a list of the population.
- Example: Selecting every 10th person on an alphabetical list of names.

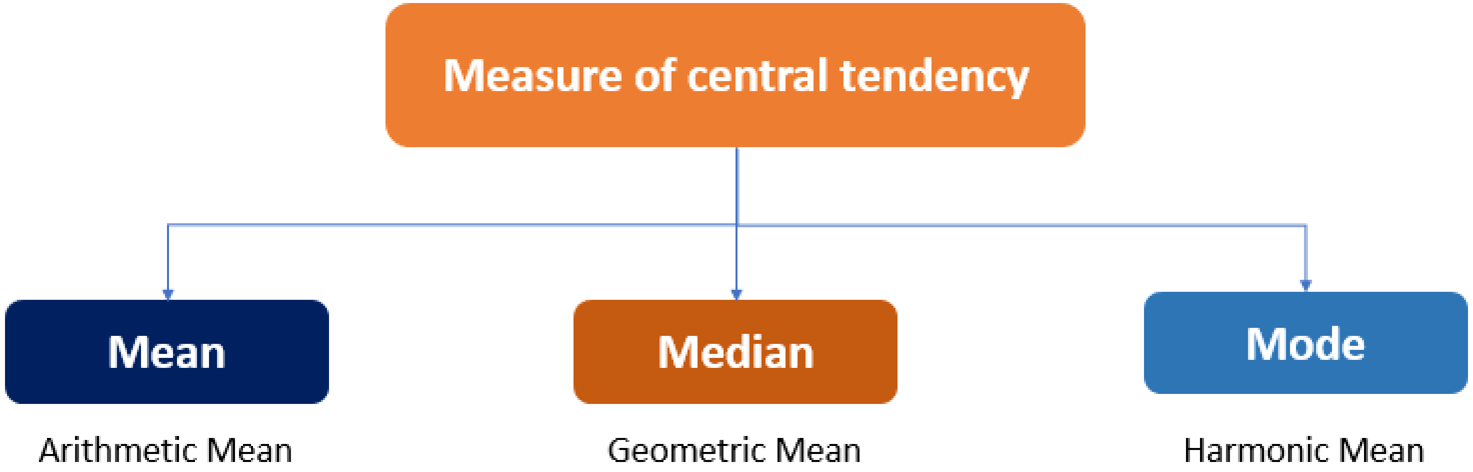
## c. Stratified Sampling

- The population is divided into subgroups (strata) based on specific characteristics (e.g., age, gender), and a random sample is taken from each stratum.
- Example: Dividing a population into males and females, and then randomly selecting individuals from each group.

## Cluster Sampling

- The population is divided into clusters (often geographically), and then entire clusters are randomly selected for study.
- Example: Randomly selecting several schools in a city and surveying all students in those schools.

# Measures of central tendency :



**Measures of central tendency** are statistical metrics used to identify the central point or typical value in a dataset. They provide a summary of the data and are essential for understanding the overall distribution of the values.

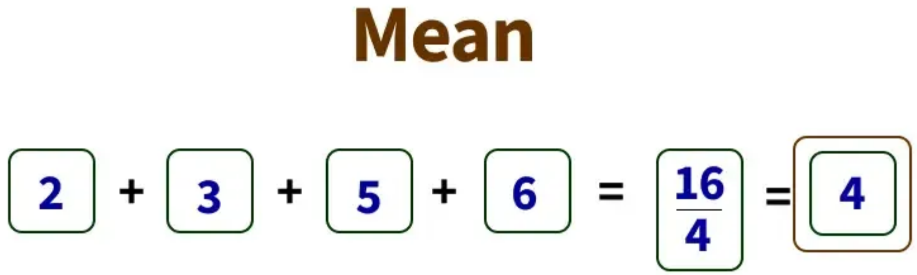
The three most common measures of central tendency are the –

# Mean (Arithmetic Mean)

- **Definition**: The mean is the sum of all values in a dataset divided by the number of values.

- **Formula**:

$$\text{Mean} = \frac{\sum (\text{All values})}{\text{Number of values}}$$



The Mean Number is 4

- **Example**:

  If the test scores of five students are 80, 85, 90, 75, and 95, the mean score is:

$$\text{Mean} = \frac{80 + 85 + 90 + 75 + 95}{5} = \frac{425}{5} = 85$$

- **Usefulness**: The mean provides a good overall summary when the data is symmetrically distributed without extreme outliers.

# 2. Median

- **Definition**: The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there's an even number of observations, the median is the average of the two middle numbers.

- **Steps to Find**:
  - Arrange the data in order.
  - If the number of observations (n) is odd, the median is the $\left(\frac{n+1}{2}\right)$th value.
  - If the number of observations (n) is even, the median is the average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th values.

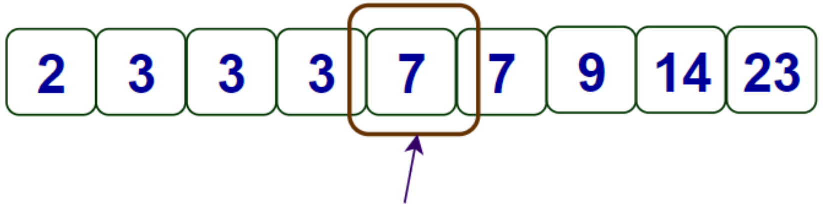
- **Example**:

  For the dataset 75, 80, 85, 90, 95 (odd number of values), the median is 85.

  For the dataset 75, 80, 85, 90 (even number of values), the median is:

$$\text{Median} = \frac{80+85}{2} = 82.5$$

Median

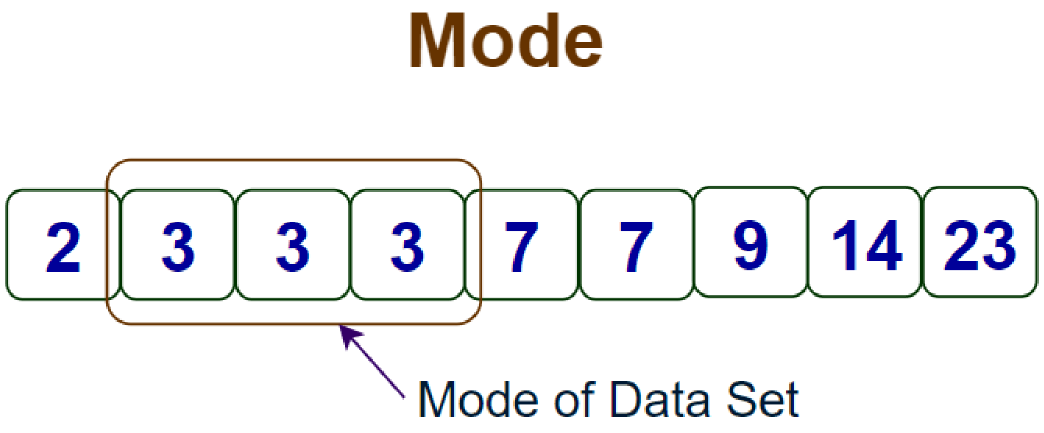


Median of Data Set when N is odd

- **Usefulness**: The median is useful in skewed distributions or when outliers are present, as it is not affected by extreme values.

# Mode

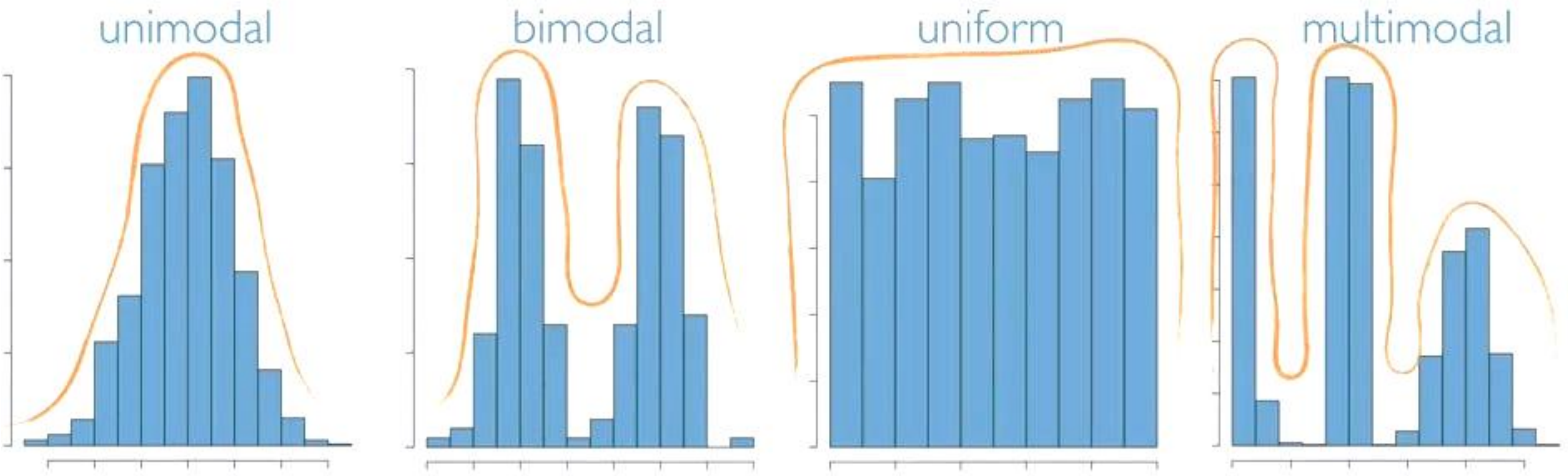
- **Definition**: The mode is the value(s) that appears most frequently in a dataset.
- **Example**: In the dataset 2, 4, 4, 6, 8, 8, 8, 10, the mode is 8 because it appears most often. A dataset can be **unimodal** (one mode), **bimodal** (two modes), or **multimodal** (more than two modes).
- **Usefulness**: The mode is helpful in identifying the most common value, particularly in categorical data or in distributions with multiple peaks.



## Comparison of the Measures

- **Mean**: Best for symmetric distributions without outliers.
- **Median**: Best for skewed distributions or when outliers are present.
- **Mode**: Best for categorical data or when identifying the most common value is important.

# modality

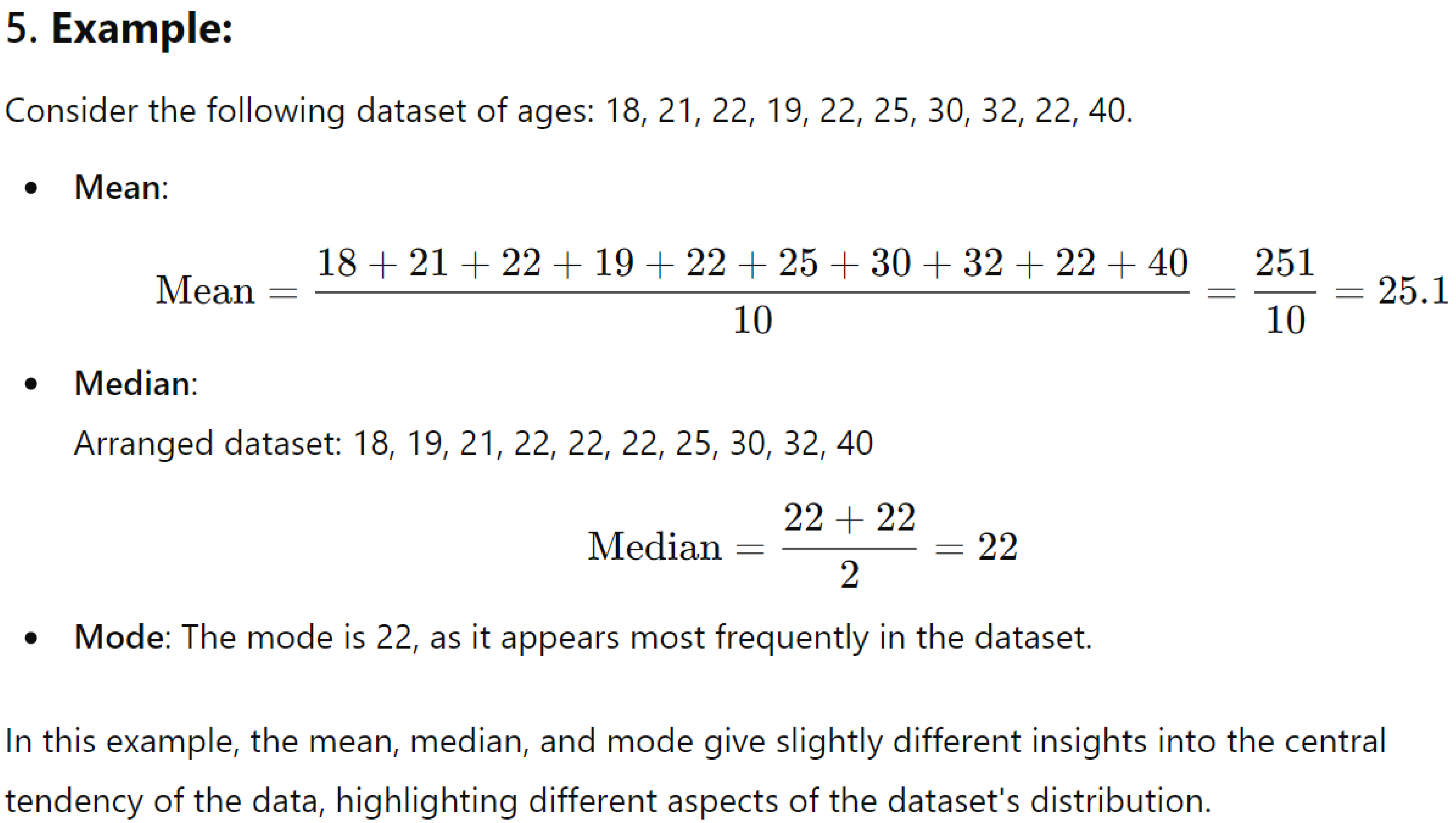


unimodal

bimodal

uniform

multimodal



### 5. Example:

Consider the following dataset of ages: 18, 21, 22, 19, 22, 25, 30, 32, 22, 40.

- **Mean**:

$$\text{Mean} = \frac{18+21+22+19+22+25+30+32+22+40}{10} = \frac{251}{10} = 25.1$$

- **Median**:

  Arranged dataset: 18, 19, 21, 22, 22, 22, 25, 30, 32, 40

$$\text{Median} = \frac{22+22}{2} = 22$$

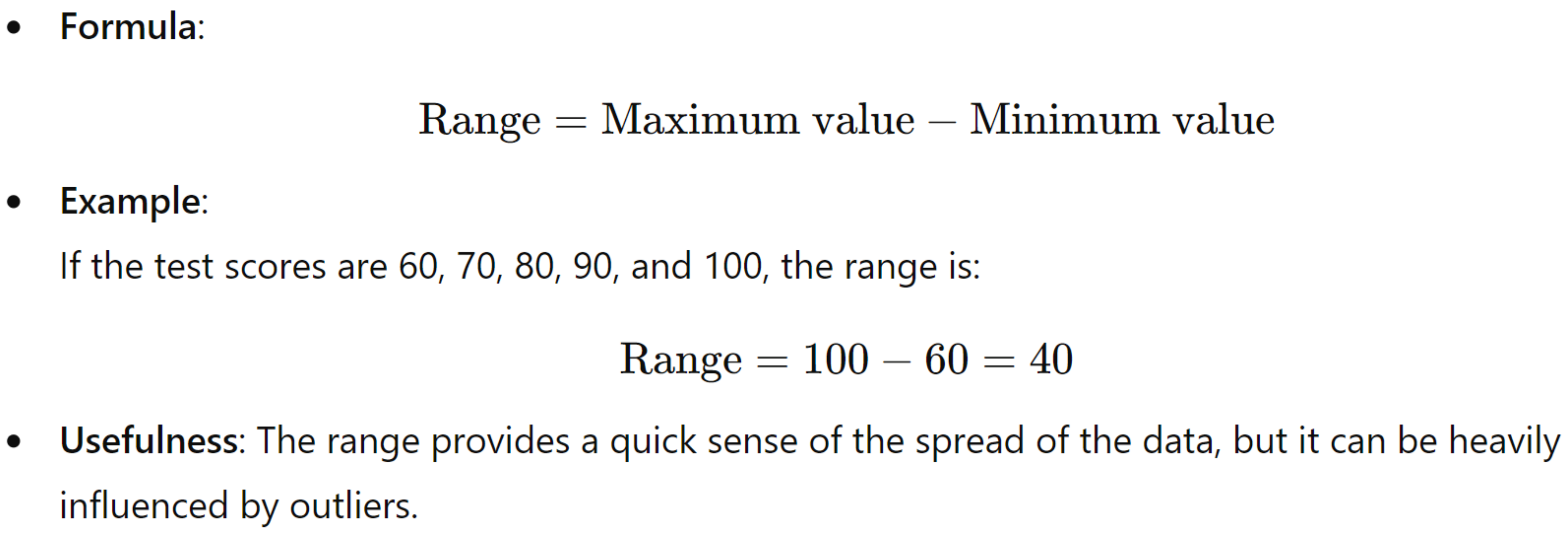
- **Mode**: The mode is 22, as it appears most frequently in the dataset.

In this example, the mean, median, and mode give slightly different insights into the central tendency of the data, highlighting different aspects of the dataset's distribution.

# Measures of dispersion

**Measures of dispersion** describe the spread or variability within a dataset. They indicate how much the data values differ from the central tendency (mean, median, mode) and from each other. These measures are crucial for understanding the distribution and reliability of the data. Here are the main measures of dispersion:

1. **Range:**



- **Formula**:

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

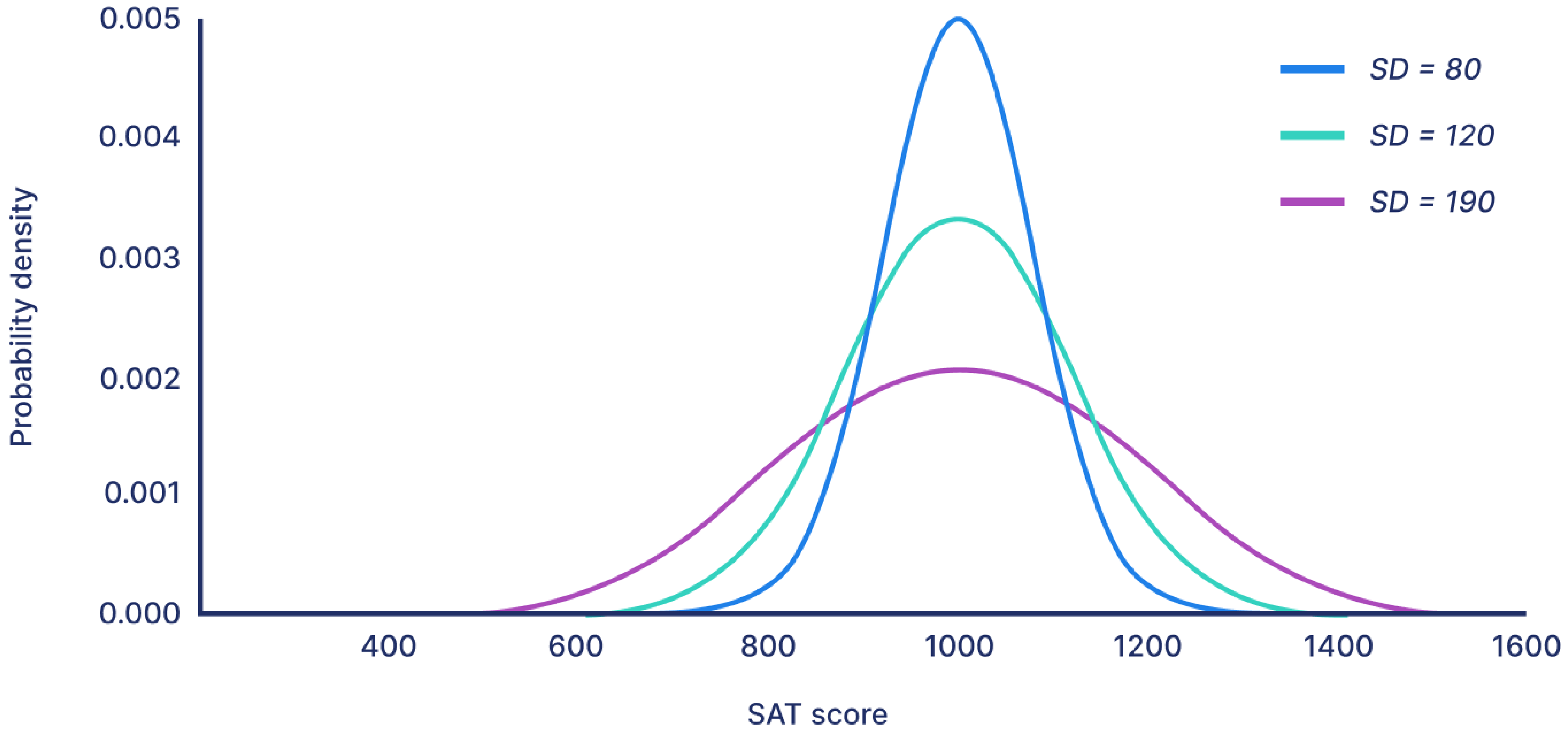
- **Example**:

  If the test scores are 60, 70, 80, 90, and 100, the range is:

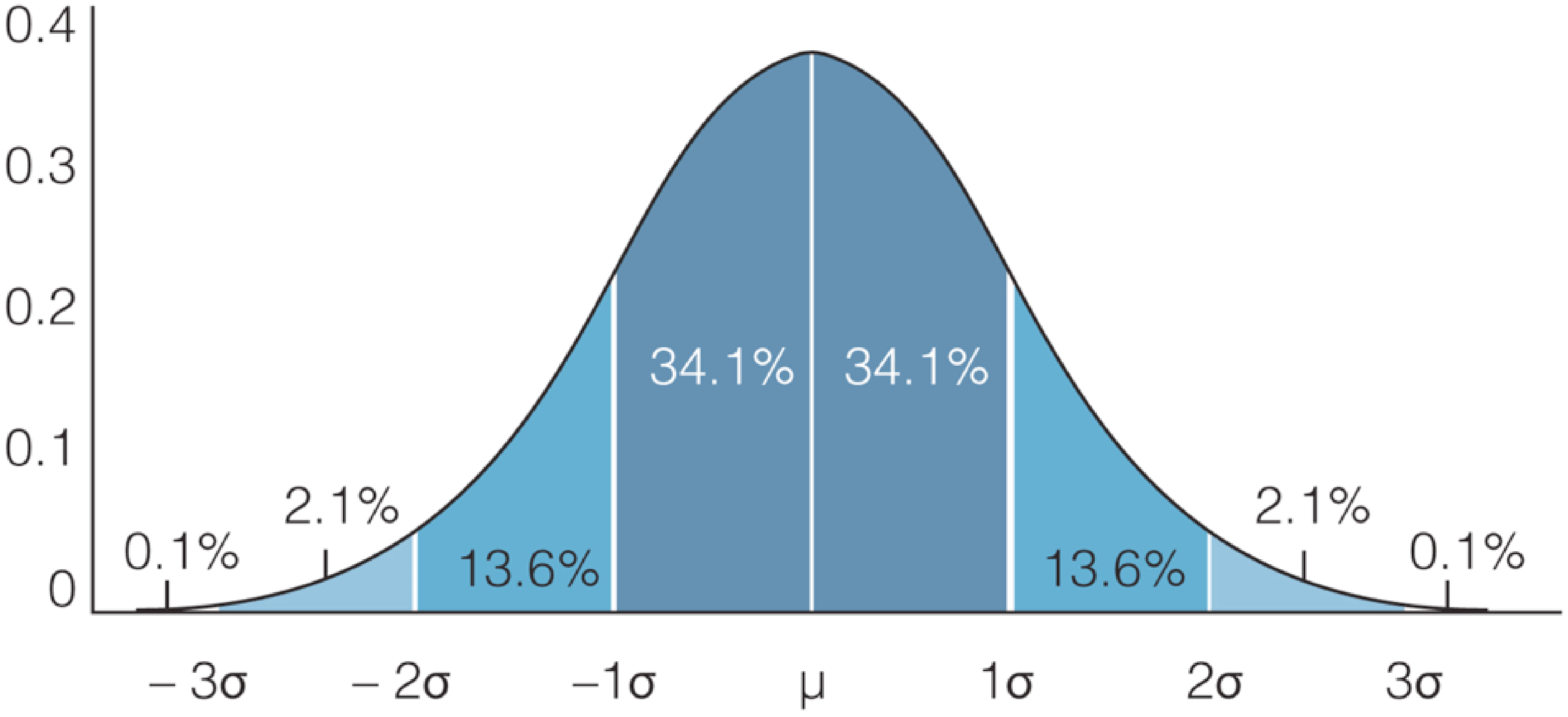
$$\text{Range} = 100 - 60 = 40$$

- **Usefulness**: The range provides a quick sense of the spread of the data, but it can be heavily influenced by outliers.

## Normal distributions with different standard deviations



# Distribution of Variance



μ = Expected Value
-1σ to 1σ = 1 Standard Deviation (ie: ~2/3 of the time, your results/variance will fall within this range)
-2σ to 2σ = 2 Standard Deviations (ie: 95% of the time, your results/variance will fall within this range)
-3σ to 3σ = 3 Standard Deviations (ie: 99.7 of the time, your results/variance will fall within this range)

## Variance

- **Definition**: Variance measures the average squared deviation of each data point from the mean. It gives an idea of how much the values in the dataset deviate from the mean.

- **Formula**:

$$\text{Variance}(\sigma^2) = \frac{\sum(x_i - \mu)^2}{n}$$

Where:

- $x_i$ = each individual value
- $\mu$ = mean of the dataset

**Example**:

For the dataset 60, 70, 80, 90, 100:

Bessel Correlation
Degree of freedom

- Mean = 80
- Variance = $\frac{(60-80)^2+(70-80)^2+(80-80)^2+(90-80)^2+(100-80)^2}{5} = \frac{400+100+0+100+400}{5} = 200$

**Usefulness**: Variance is useful for understanding the degree of spread in the data, but because it's in squared units, it's less interpretable directly compared to the standard deviation.

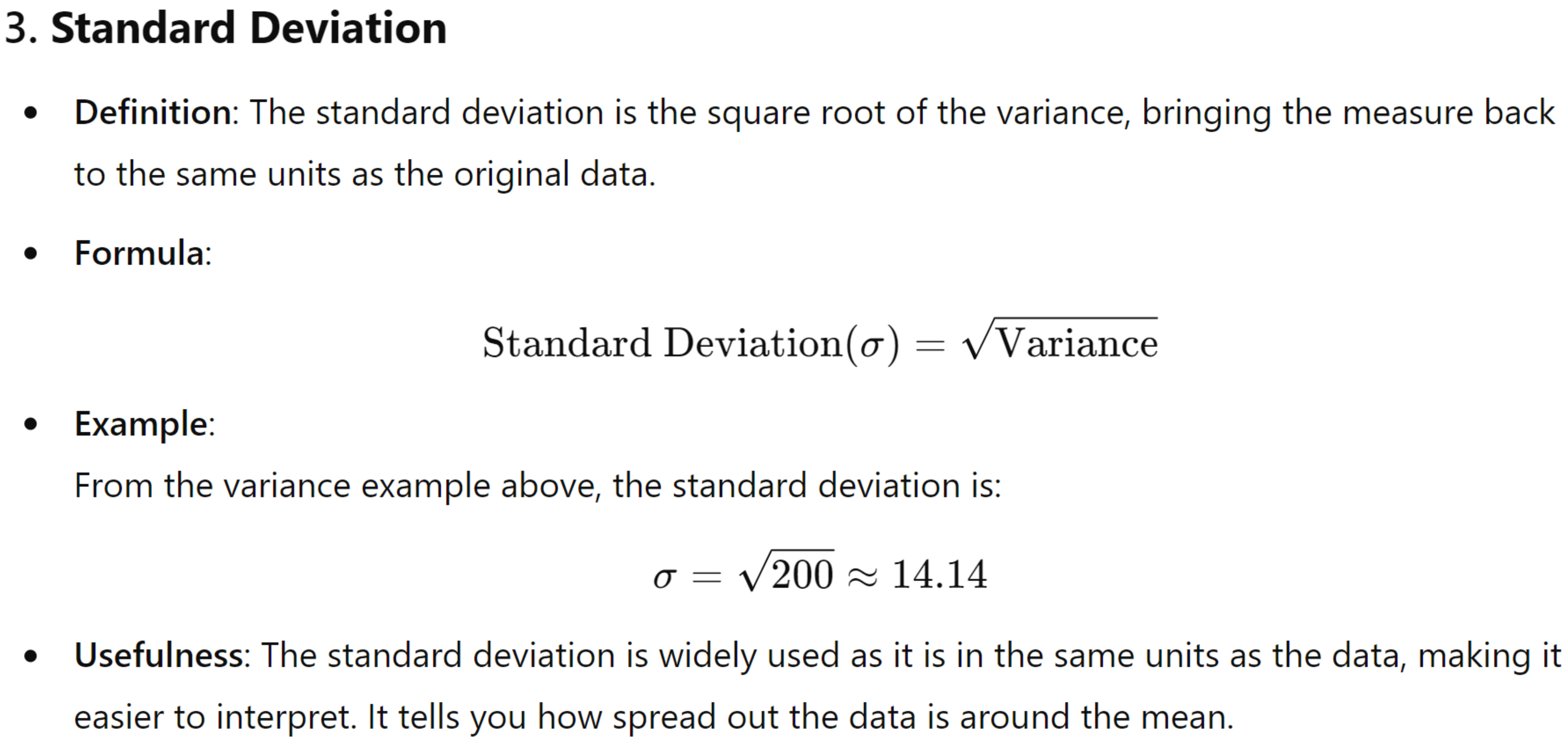
| X | X | x - X | (x – X)2 |
|---|---|---|---|
| 1 | 3 | -2 | 4 |
| 2 | 3 | -1 | 1 |
| 2 | 3 | -1 | 1 |
| 3 | 3 | 0 | 0 |
| 3 | 3 | 0 | 0 |
| 4 | 3 | 1 | 1 |
| 4 | 3 | 1 | 1 |
| 5 | 3 | 2 | 4 |
| Total= 8 | | | 12 |

$$\text{Standard Deviation}(\sigma) = \sqrt{\text{Variance}}$$

**= 1.5**

**= 1.224**

$$\text{Variance}(\sigma^2) = \frac{\sum(x_i - \mu)^2}{n}$$ **= 12/8** **= 1.5**



### 3. **Standard Deviation**

- **Definition**: The standard deviation is the square root of the variance, bringing the measure back to the same units as the original data.
- **Formula**:

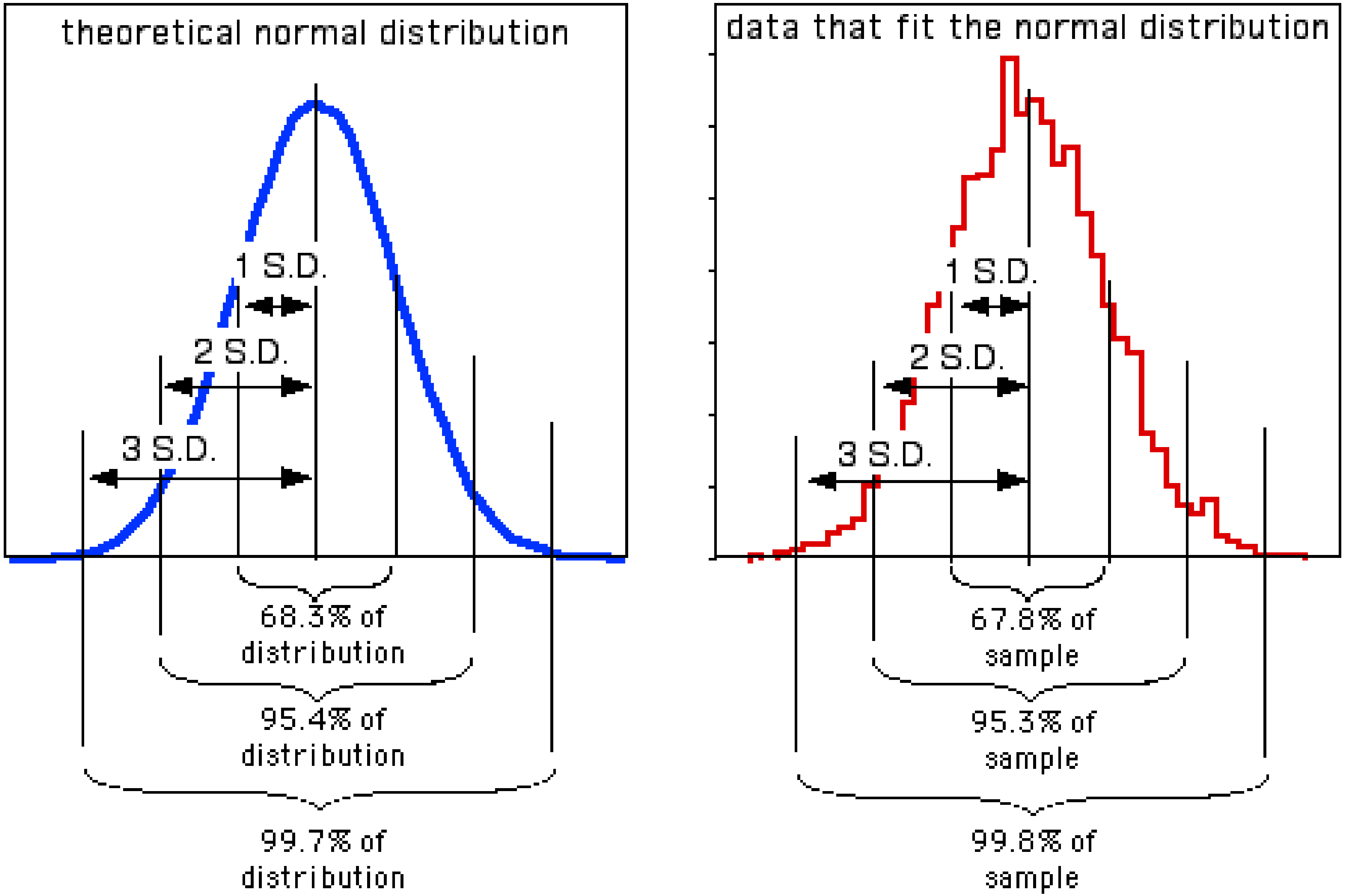
$$\text{Standard Deviation}(\sigma) = \sqrt{\text{Variance}}$$

- **Example**:

  From the variance example above, the standard deviation is:

$$\sigma = \sqrt{200} \approx 14.14$$

- **Usefulness**: The standard deviation is widely used as it is in the same units as the data, making it easier to interpret. It tells you how spread out the data is around the mean.



theoretical normal distribution
1 S.D.
2 S.D.
3 S.D.
68.3% of distribution
95.4% of distribution
99.7% of distribution
data that fit the normal distribution
1 S.D.
2 S.D.
3 S.D.
67.8% of sample
95.3% of sample
99.8% of sample