Basic Statistics for Data Science

Arif Istiake Sunny

Senior AI Engineer

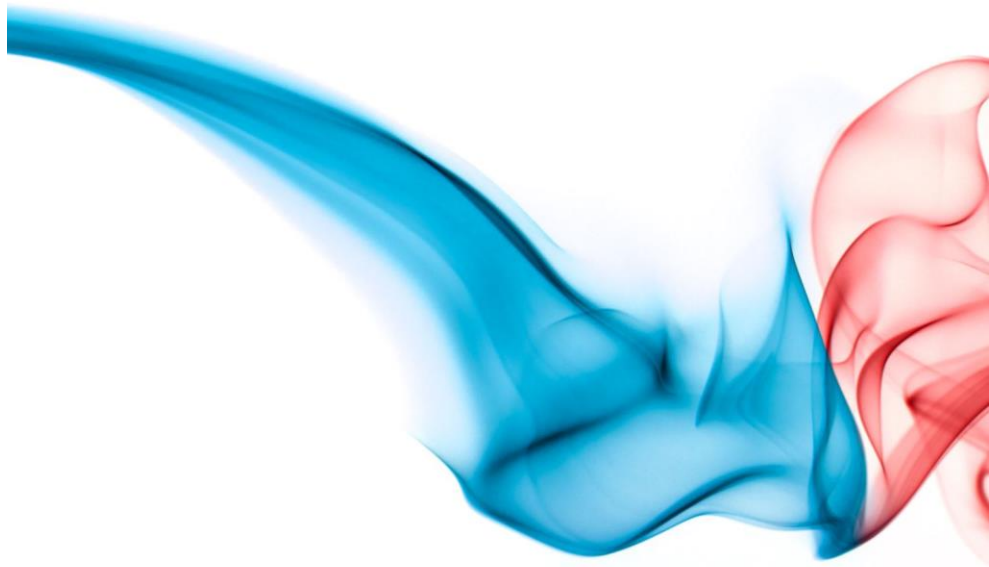
Brain Station 23 PLC

Email: sunny1509006@gmail.com

Mobile: 01732009493

Agenda

- What Is Statistics?
- What Is Data Science?
- Qualitative Data
- Quantitative Data
- Frequency Distribution
- Data Presentation
- Bar Diagram
- Histogram
- Pie Chart
- Central Tendency
- Mean
- Median
- Mode
- Measures of Dispersion
- Variance



Why is basic statistical knowledge required in Data Science?

What Is Statistics?

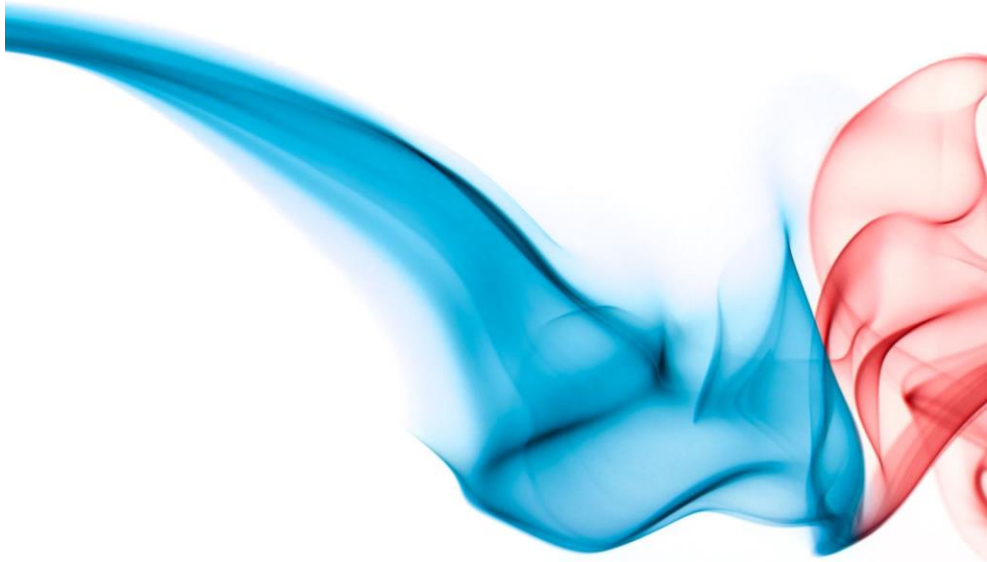
Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.

What Is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Examples

- Weather forecasting 🌦️
- Cricket batting average 🏏
- Poll results before elections 🗳️



Types of Statistics

✅ **Descriptive Statistics**
Describes and summarizes data.
➤ Mean, Median, Mode, Standard Deviation, etc.

✅ **Inferential Statistics**
Uses data from a sample to make inferences about a population.
➤ Hypothesis testing, confidence intervals, regression

Types of Data

Type	Description	Examples
Qualitative	Categorical, non-numeric	Gender, Color, Religion
Quantitative	Numeric	Height, Age, Salary

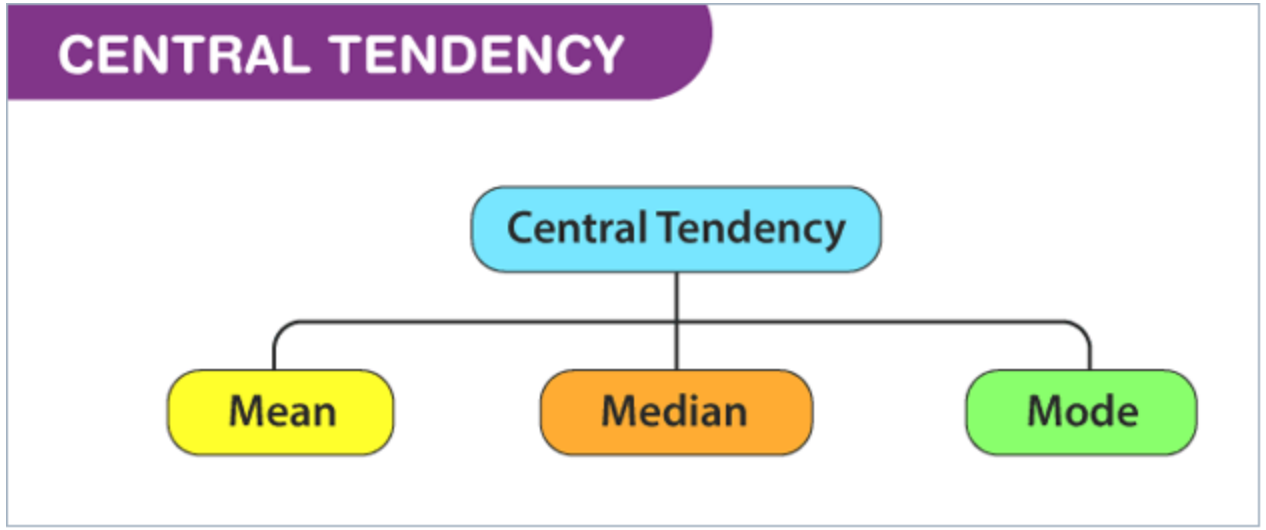
Quantitative is further divided into:

- **Discrete:** Countable (e.g., No. of pets 🐶)
- **Continuous:** Measurable (e.g., Height 📏)

Measures of Central Tendency

Central tendency is defined as “the statistical measure that identifies a single value as representative of an entire distribution or sample.

- ✅ **Mean** – Average
- ✅ **Median** – Middle value
- ✅ **Mode** – Most frequent value

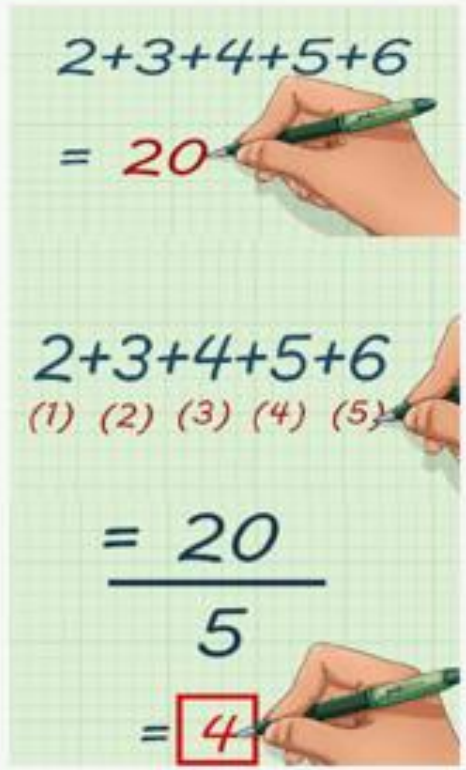


Example:
Test scores: [70, 85, 85, 90, 100]

- Mean = 86
- Median = 85
- Mode = 85

Mean (গড়)

The mean is the mathematical average of a set of two or more numbers.



Mean

Use:

- To know the overall result
- To determine the skewness and standard deviation
- To determine T-Score, Z-Score

Merits:

- Easy to determine
- Observation has equal weightage
- Most reliable Central Tendency

Mean

Limitations:

- Value of mean will be changed even of one data is changed
- Mean can not be determined if data is qualitative
- Mean is affected by extreme score (Outlier)
- It cannot be computed accurately if any item is missing
- It can not be calculated graphically

Mean

Mean

$$\text{Mean} = \frac{\text{Sum of Data Points}}{\text{Number of Data Points}}$$

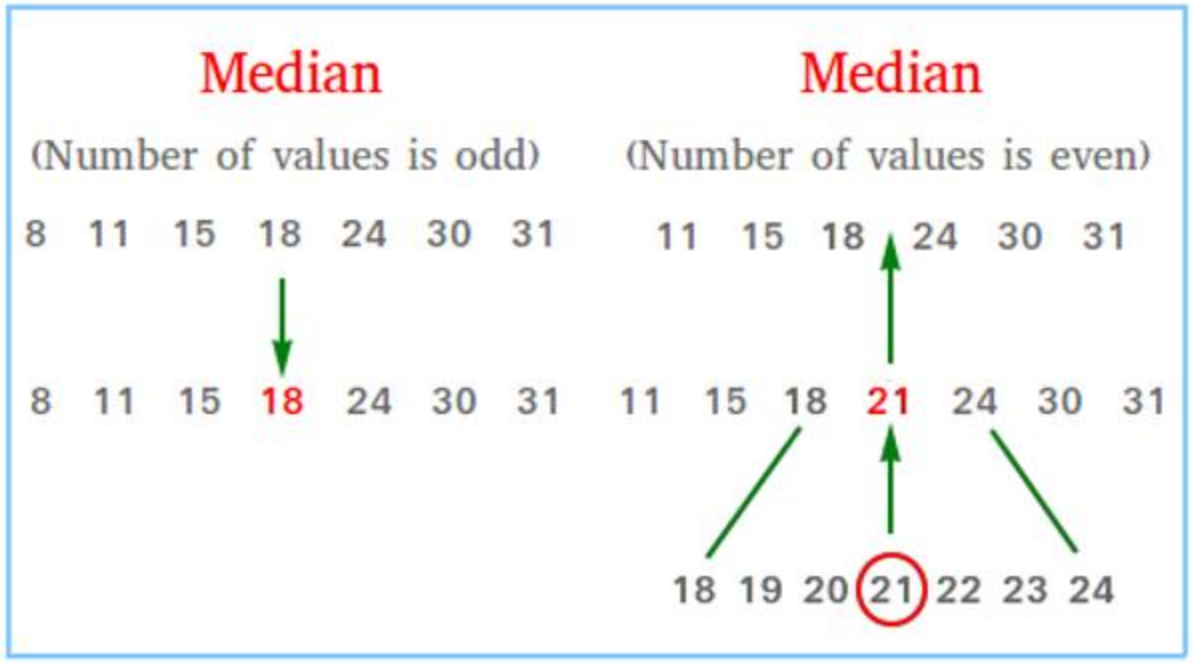
$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Data Set: 6,4,10,3,7

$$\bar{x} = \frac{6 + 4 + 10 + 3 + 7}{5} = \frac{30}{5} = 6$$

Median(মধ্যমা/মধ্যক)

The median is the value that's exactly in the middle of a dataset when it is ordered. It's a measure of central tendency that separates the lowest 50% from the highest 50% of values.



Median

Use

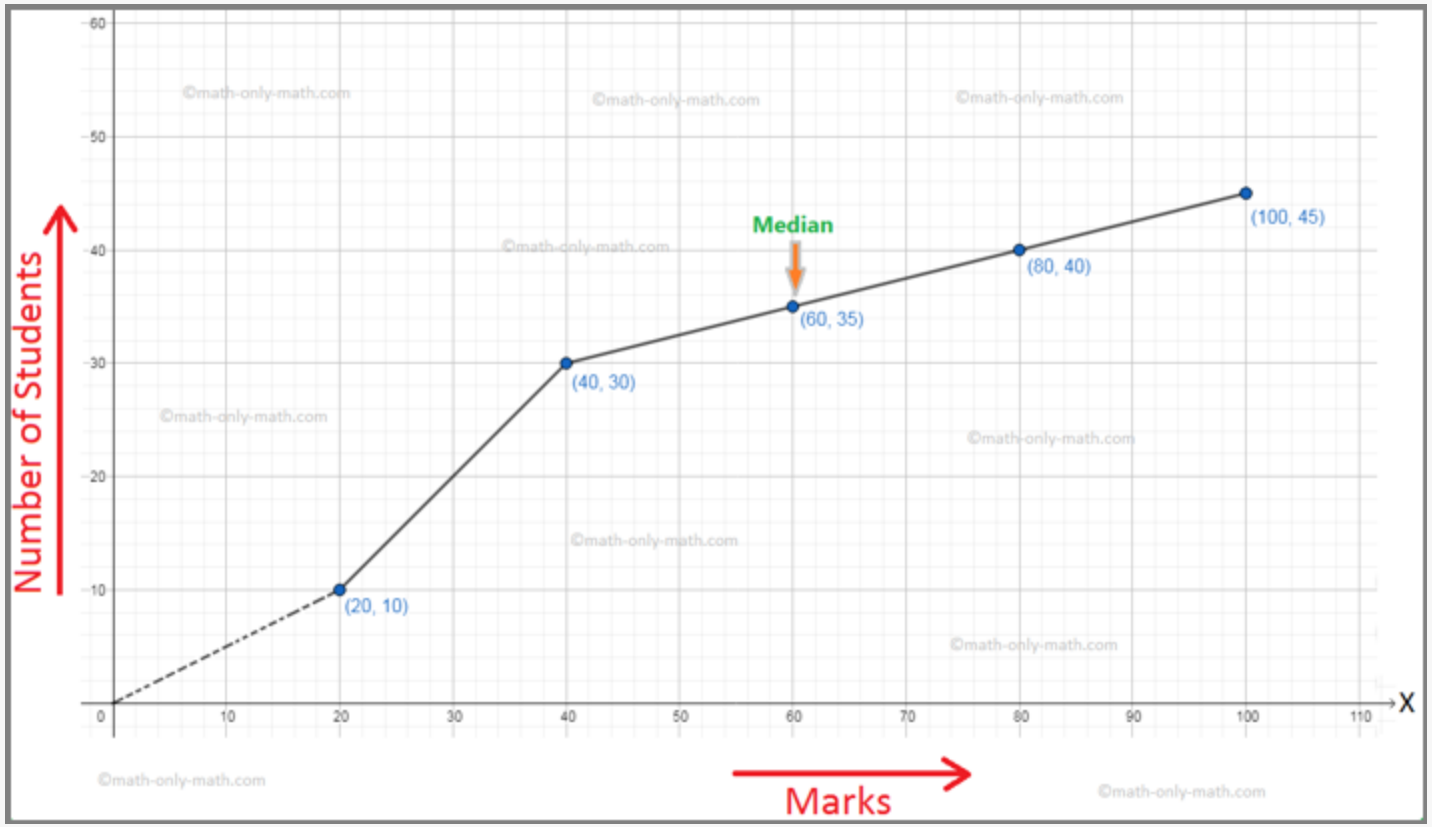
- When we want to know the exact mid point
- When the distribution is skewed.
- To find skewness
- When the distribution contains outliers.
- Also it can be used for the qualitative value

Merits

- It is not at all affected by extreme values (Outliers)
- Median can be determined even any data is missing
- It can be located graphically
- It is easily understood and is easy to calculate. In some cases it can be located merely by inspection.

Median (মধ্যমা/মধ্যক)

Median can be located graphically



Median

Limitations:

- While calculating median, all the data should be arranged in ascending or in descending order. In case of large number of items, it becomes tedious and time consuming.
- Median provides correct result in case of odd observation. When the number is observation is even it fails to obtain accurate result.
- It does not use all the data/score

Median

How to find the Median:

If ‘n’ is odd: $\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$

If ‘n’ is even: $\text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$

How to find the Median:

For Odd Number

Find the median of this set of data values.

47 35 37 32 38 39
36 34 35

Solution:

Lowest value to the highest value:

32 34 35 35 36 37
38 39 47

The number of values, n, in the data set $= 9$

$$\text{Median} = \frac{1}{2}(9+1) \text{ th value}$$

$$= 5\text{th value}$$

$$= 36$$

For Even Number

Find the median of the following data set:

12 18 16 21 10 13 17 19

Solution:

Arrange the data values in order from the lowest value to the highest value:

10 12 13 16 17 18 19 21

The number of values in the data set is 8, which is even.

$$\therefore \text{Median} = \frac{4\text{th data value} + 5\text{th data value}}{2}$$

$$= \frac{16+17}{2}$$

$$= \frac{33}{2}$$

$$= 16.5$$

Mode(প্রচুরক)

The mode is the value that appears most frequently in a data set.

Example:

In {6, 3, 9, 6, 6, 5, 9, 3} the Mode is 6, as it occurs most often.

Mode(প্রচুরক)

Use

- The mode represents the value(s) that occurs most often in a dataset.
- The mode tells us the most common value in categorical data when the mean and median can't be used.

Merits

- It can be determined by observation
- It is not affected by outliers
- Also it can be used for the qualitative value
- It can be presented graphically

Mode(প্রচুরক)

Limitations:

- We cannot find the mode of the equal series
- Mode may not be exist
- Sometime there are more than one mode
- Mode gives us an idea of where the "center" of a dataset is located, but it can be misleading compared to the mean or median.

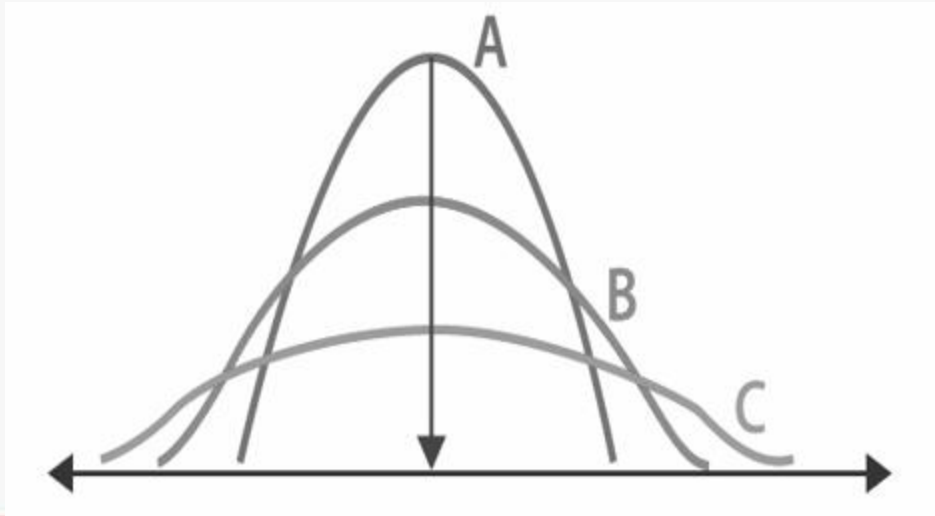
Measures of Dispersion

- **Range** = Max - Min
- **Variance** = Average squared deviation from the mean
- **Standard Deviation (SD)** = Square root of variance

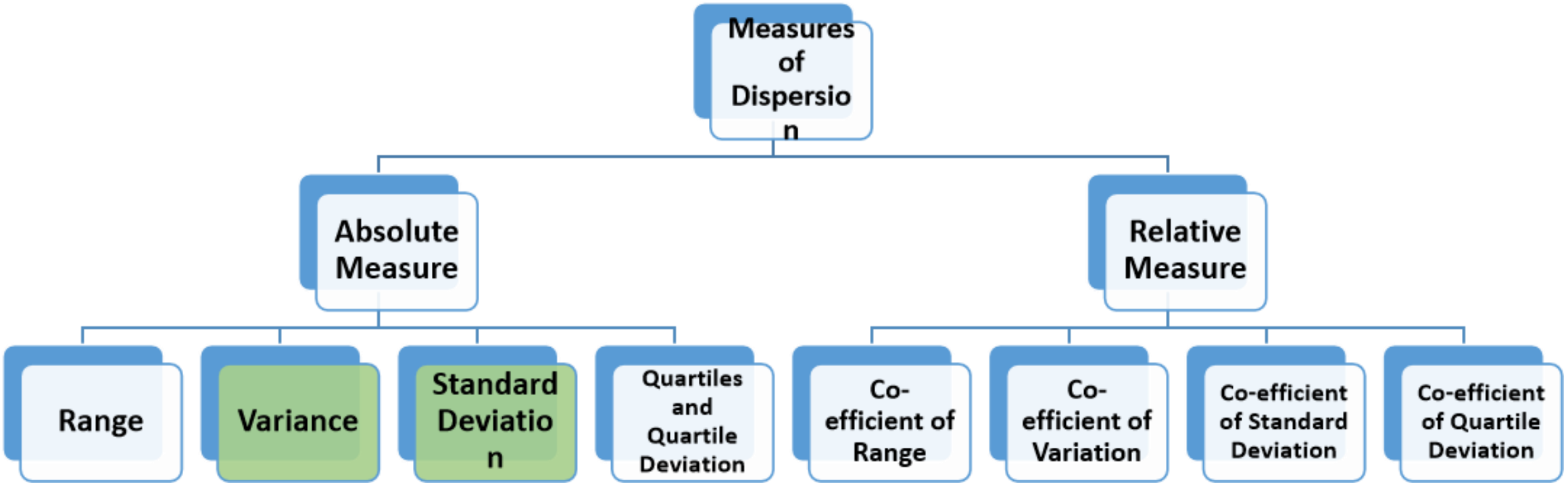
Measures of Dispersion

Literal meaning of dispersion is scatter ness. Dispersion is the degree of the scatter ness or deviation of each value in the data set from a measure of central tendency usually the mean.

- The more similar the scores are to each other, the lower Measures of Dispersion will be
- The less similar the scores are to each other, the higher Measures of Dispersion will be



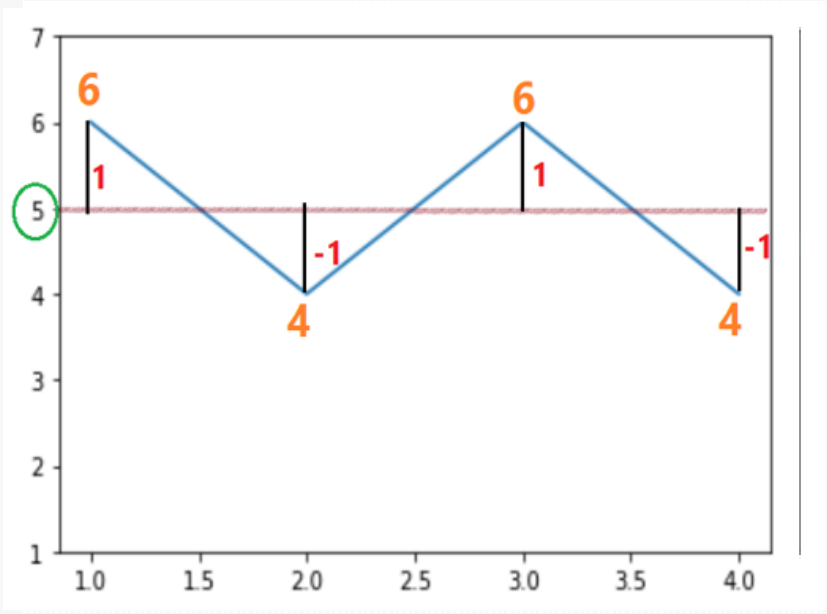
Measures of Dispersion



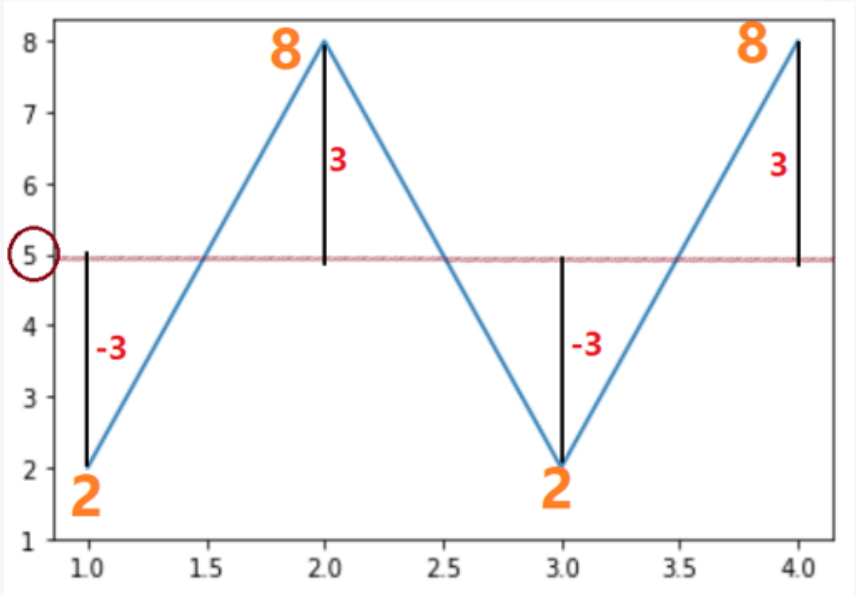
Variance

The variance is a measures that indicates how much data scatter around the mean.

6, 4, 6, 4 **Mean = 5**



2, 8, 2, 8 **Mean = 5**



Variance

6, 4, 6, 4 Mean = 5

Total Distance = $(6-5) + (4-5) + (6-5) + (4-5) = 0$ X

Total Distance = $(6-5)^2 + (4-5)^2 + (6-5)^2 + (4-5)^2 = 4$

$$\text{Variance} = \frac{(6-5)^2 + (4-5)^2 + (6-5)^2 + (4-5)^2}{4} = 1$$

Variance

Population Data **x1, x2, ...** x_n **Mean =** μ

Variance $\sigma^2 = \frac{(x1 - \mu)^2 + (x2 - \mu)^2 \ldots\ldots (x_n - \mu)^2}{N}$

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

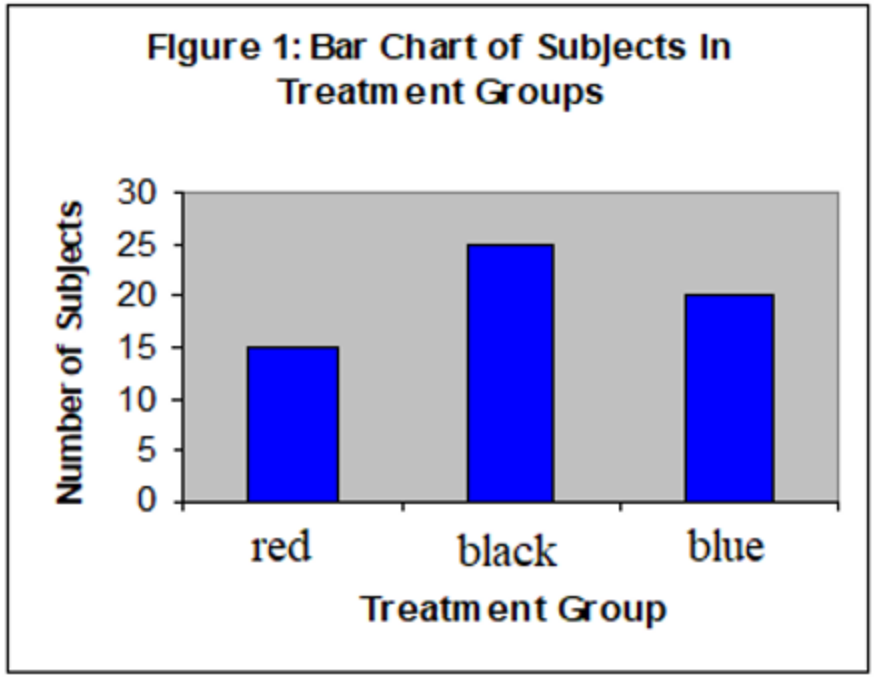
Sample Data $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Data Representation

- **Tabular Form** (Frequency tables)
- **Graphical Form**
  - Bar chart 📊
  - Histogram 📉
  - Pie chart 🥧
  - Box plot 📦
  - Line graph 📈

Data Presentation –Categorical Variable

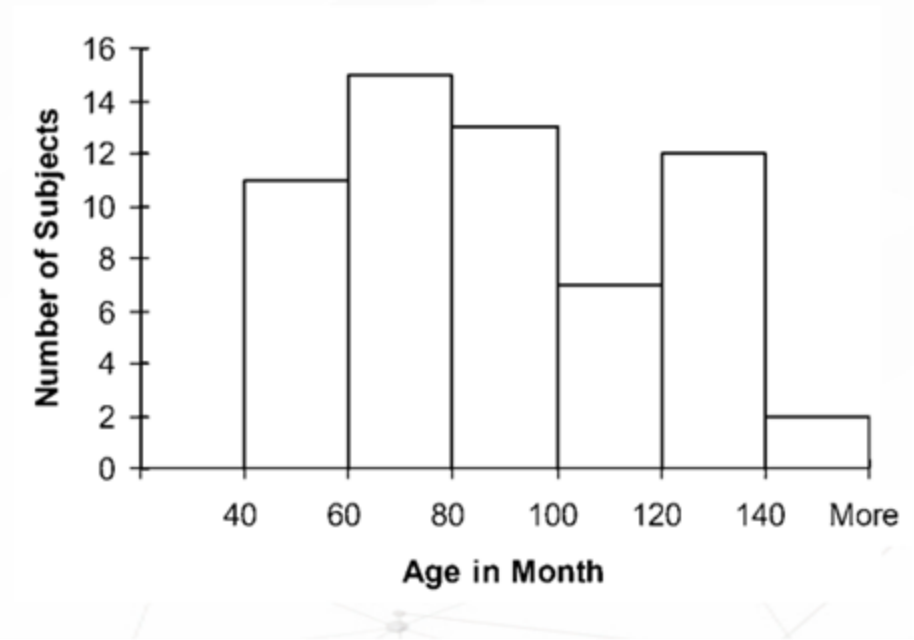
Bar Diagram: A bar diagram (or a bar graph) is a rectangular bar shaped statistical graphic which is divided into several bar to illustrate numerical proportion.



Treatment Group	Frequency
red	15
black	25
blue	20
Total	60

Graphical Presentation –Numerical Variable

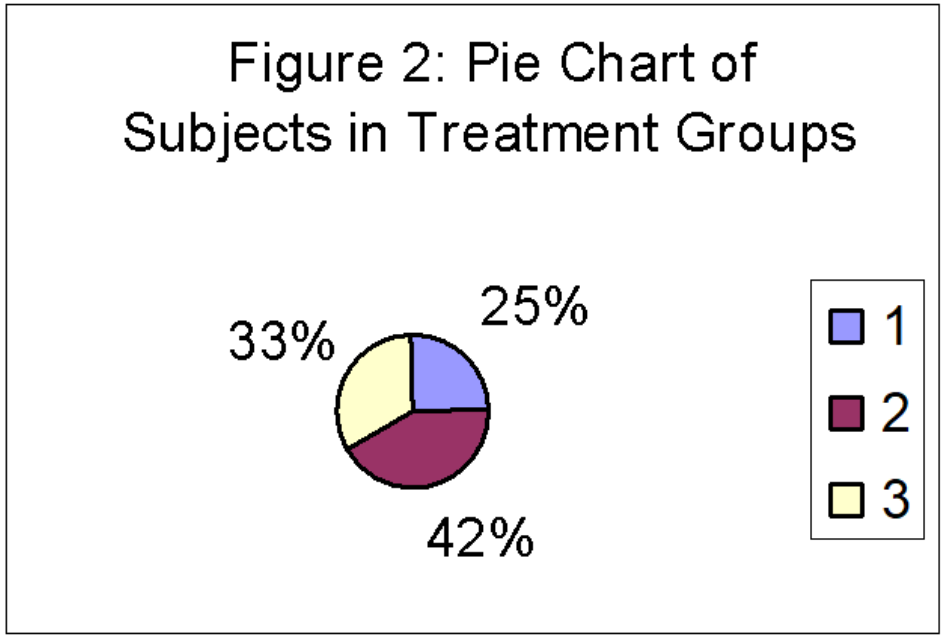
Histogram is a graphical representation of the distribution of numerical data. Overall pattern can be described by its **shape, center, and spread**. The following age distribution is **right skewed**. The center lies between **80 to 100. No outliers**.



Age	Frequency
40-60	11
60-80	15
80-100	13
100-120	7

Data Presentation – Categorical Variable

Pie Chart: A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion.



Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.417	41.7
3	20	(20/60)=0.333	33.3
Total	60	1.00	100