

Basic Statistics

WHAT IS STATISTICS?



- **Descriptive Statistics –**

this offers methods to summarised data by transforming raw observations into meaningful information that is easy to interpret and share.

- **Inferential Statistics –**

this offers methods to study experiments done on small samples of chalk out the inferences to the entire population (entire domain)



STATISTICS

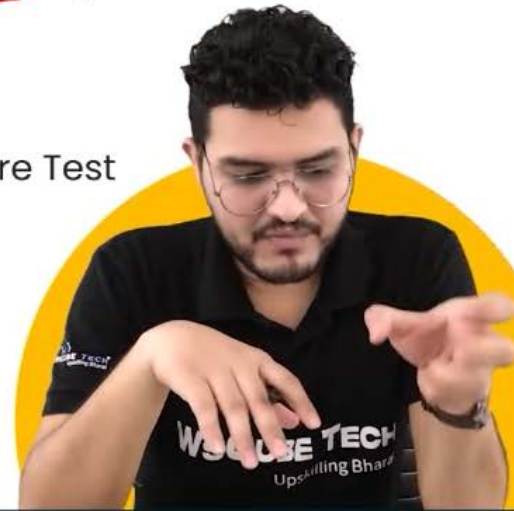


✓ Descriptive

- ▣ Measures of Central Tendency → Mean, Median, Mode
- Measures of Variability → Range, MAD, Var, Std
- Measures of Shape → Skewness
- Percentiles → Boxplot
- Frequency Distribution → plot
- Covariance and Correlation →

Inferential

- Central Limit Theorem *
- Hypothesis Testing
 - Z - Test
 - T - Test
 - Chi Square Test

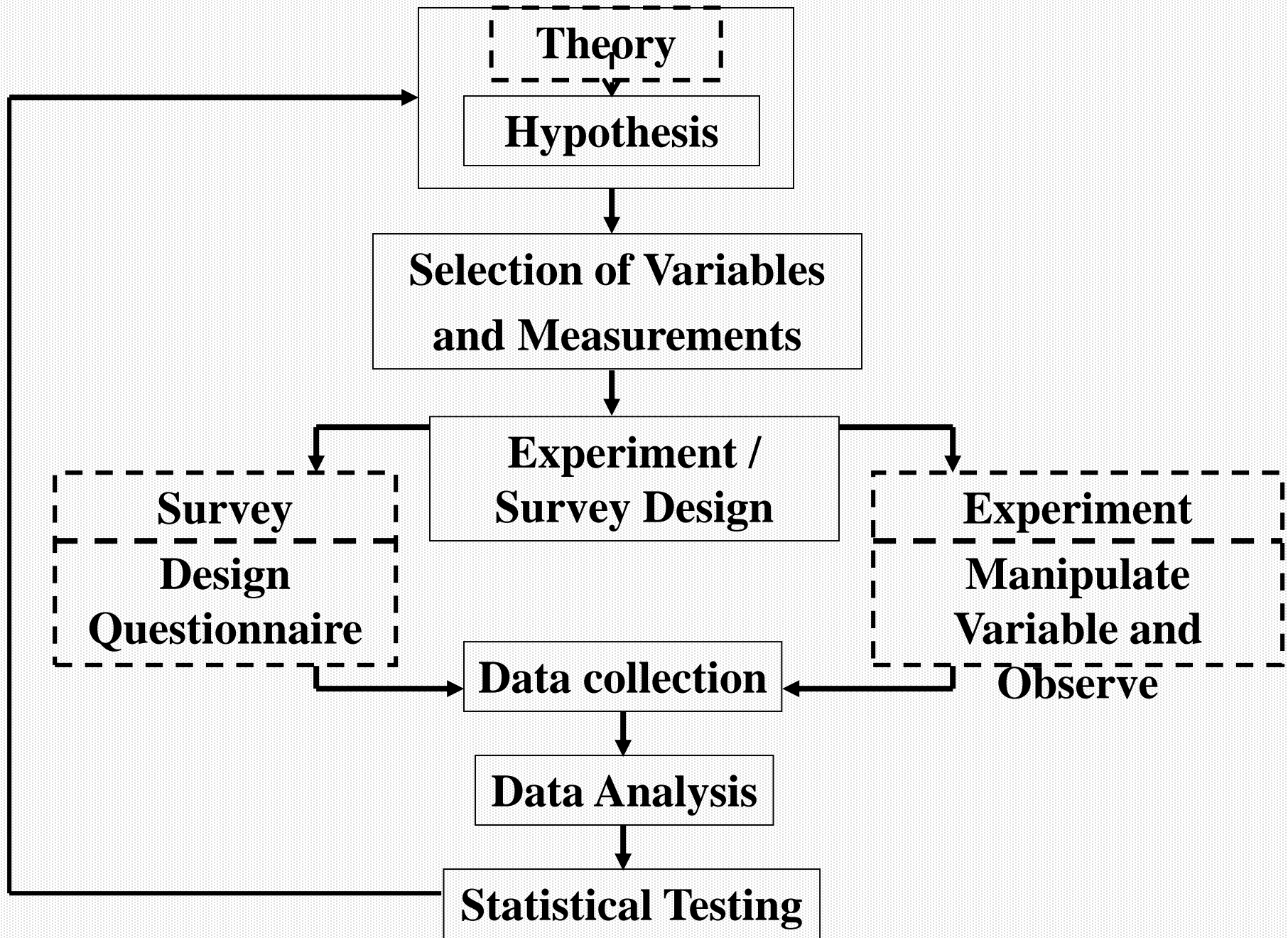


Become a **Data Analyst** in Just 20 Weeks? **Apply Now at** www.wscubetech.com/data-analytics-course

Qualitative and Quantitative Approaches

Qualitative	Quantitative
(Usually) Non-probability based sample	Typically a probability-based sample
Non-generalizable	Generalizable
Answers Why? How?	Answers How many? When? Where?
Researcher is the instrument	Various tools, instruments employed

Quantitative Analysis Procedure



Basics of Statistics

Definition: Science of collection, presentation, analysis, and reasonable interpretation of data. Statistics presents a rigorous scientific method for gaining insight into data.

Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable 'age' can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26

Data Presentation

Two types of statistical presentation of data - **graphical and numerical.**

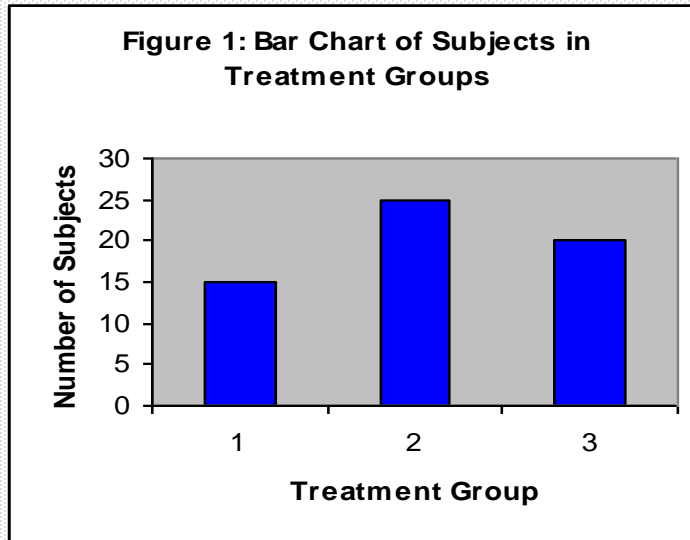
Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Overall pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an *outlier*.

Bar diagram and Pie charts are used for categorical variables.

Histogram, Box-plot are used for numerical variable.

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.

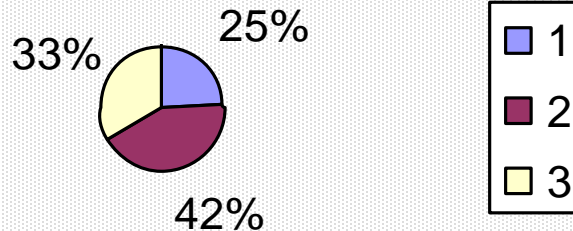


Treatment Group	Frequency
red	15
black	25
blue	20
Total	60

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.

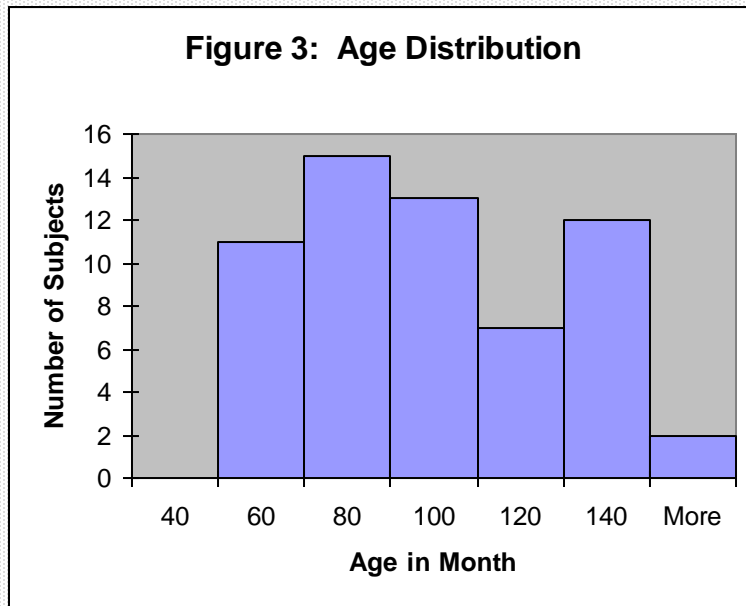
Figure 2: Pie Chart of Subjects in Treatment Groups



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

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**.



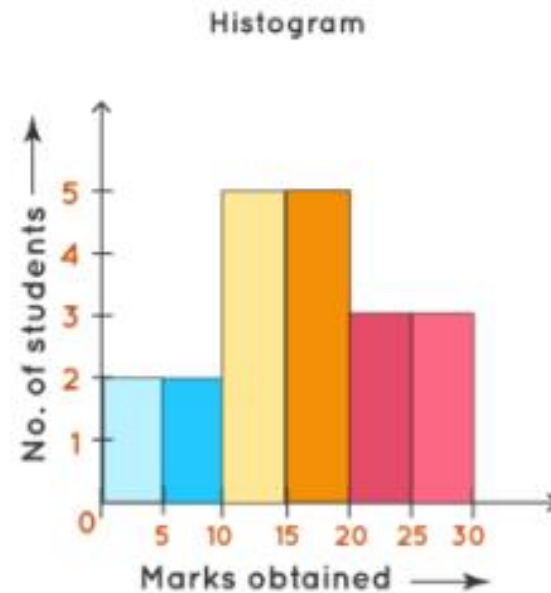
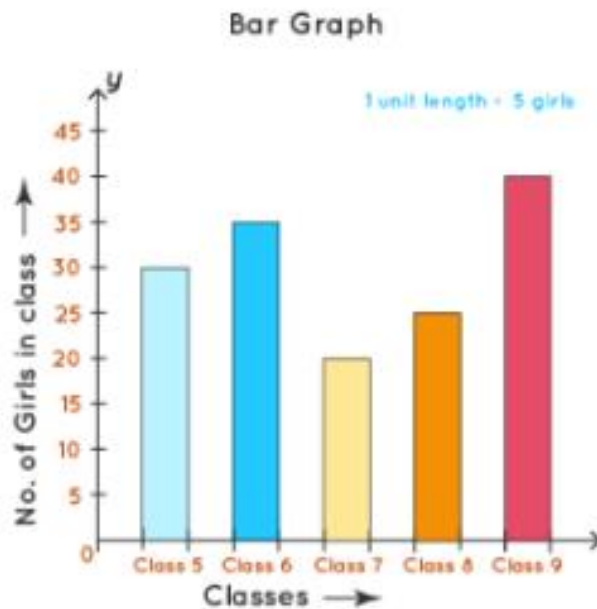
Age	Frequency
40-60	11
60-80	15
80-100	13
100-120	7
120-140	12
More	2

Bar Graph vs Histogram

The main differences between a bar chart and a histogram are as follows:

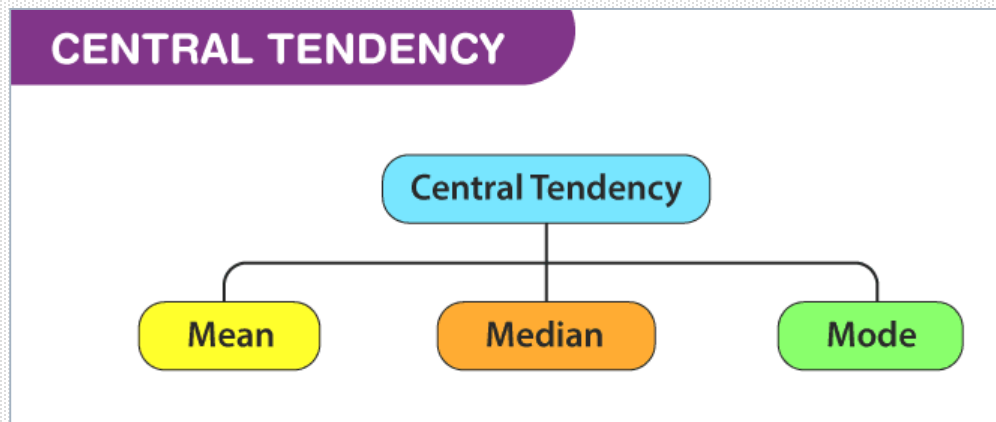
Bar Graph	Histogram
Equal space between every two consecutive bars.	No space between two consecutive bars. They should be attached to each other.
The x-axis can represent anything.	The x-axis should represent only continuous data that is in terms of numbers.

Bar Graph vs Histogram



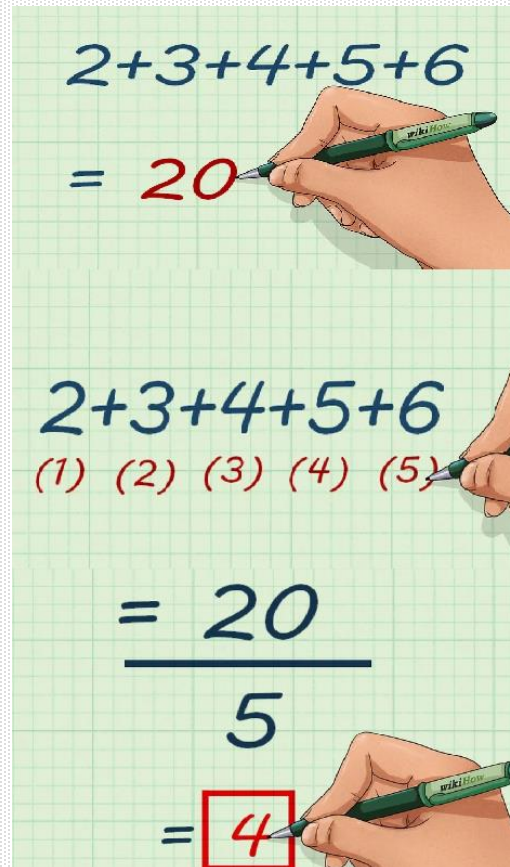
Central Tendency

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



Mean (গড়)

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



The illustration shows the calculation of the mean for the numbers 2, 3, 4, 5, and 6. It is divided into three parts:

- Top part:** The numbers are summed: $2+3+4+5+6 = 20$. The number 20 is written in red.
- Middle part:** The numbers are listed with indices below them: $2+3+4+5+6$ with (1), (2), (3), (4), and (5) respectively. The number 5 is written in red.
- Bottom part:** The sum is divided by the count: $\frac{20}{5} = 4$. The number 4 is written in red and enclosed in a red box.

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 if one data point is changed
- Mean cannot be determined if data is qualitative
- Mean is affected by extreme score (Outlier)
- It cannot be computed accurately if any item is missing
- It cannot be calculated graphically

Mean

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

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

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

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

Mean

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

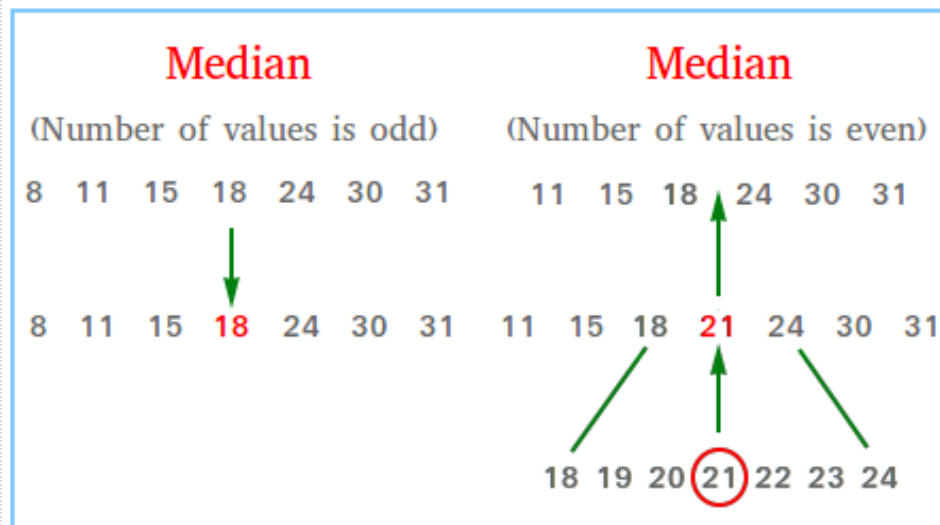
$\sum x_i \longrightarrow x_1 + x_2 + x_3 + \dots + x_n$

$n \longrightarrow$ Total number of elements in a group

$\mu \longrightarrow$ Mean

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

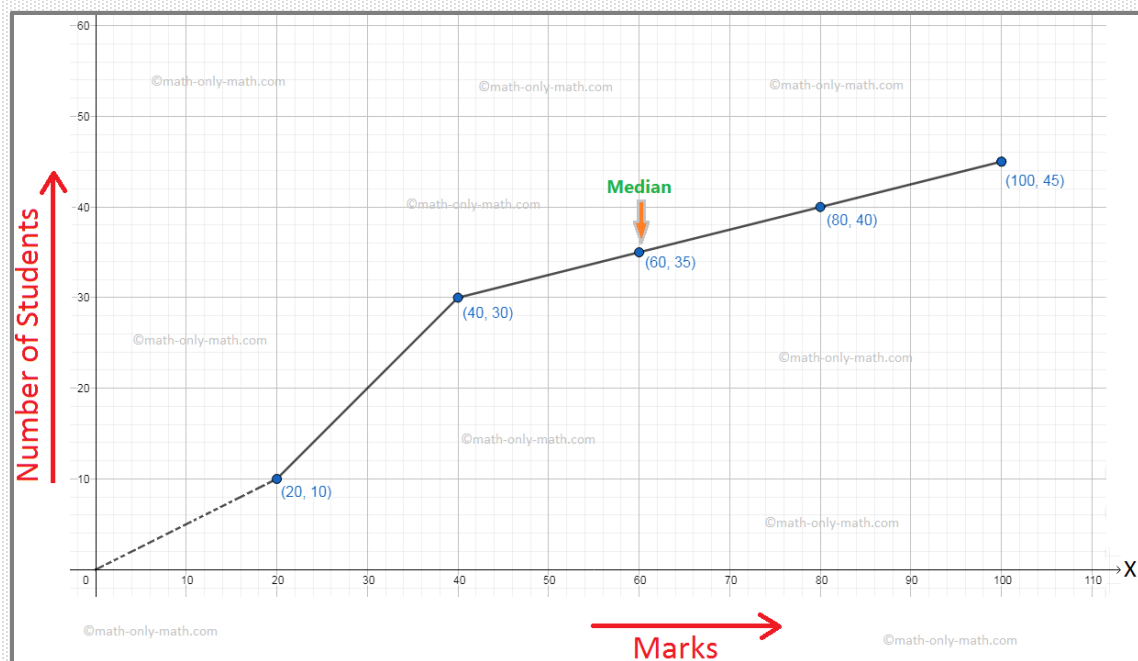
- When we want to know the exact mid point
- When the distribution is skewed.
- To find skewness
- When the distribution contains outliers.
- Also it can be used for the qualitative value

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 of observations 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: $\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ term}$

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

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

$$\begin{aligned}\text{Median} &= \frac{1}{2}(9+1) \text{ th value} \\ &= 5\text{th value} \\ &= 36\end{aligned}$$

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.

$$\begin{aligned}\therefore \text{Median} &= \frac{4\text{th data value} + 5\text{th data value}}{2} \\ &= \frac{16+17}{2} \\ &= \frac{33}{2} \\ &= 16.5\end{aligned}$$

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)

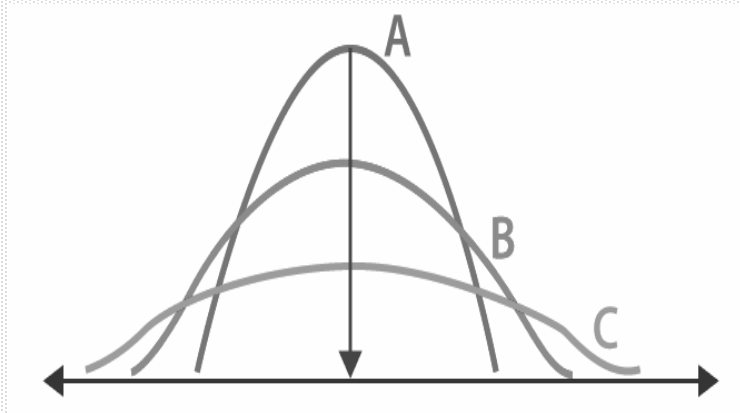
Limitations:

- 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.

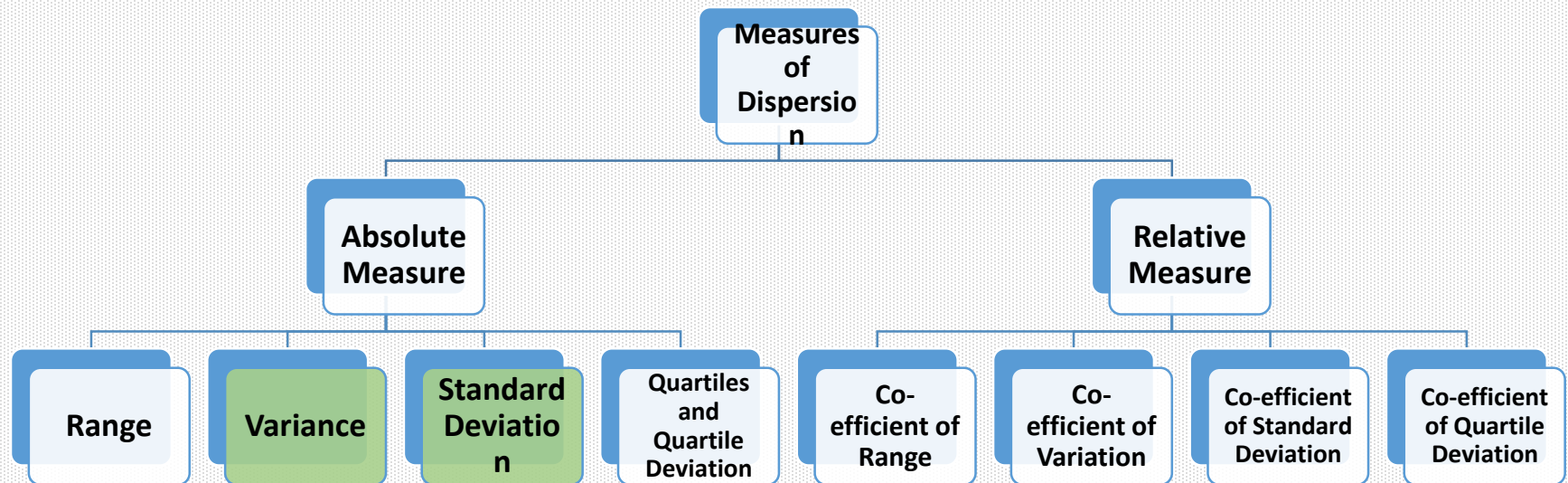
Measures of Dispersion

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



Measures of Variability

- ✓ Range
- ✓ Mean Absolute Deviation
- ✓ Variance
- ✓ Standard Deviation



Activate Windows
Go to Settings to activate Windows.

Become a **Data Analyst** in Just 20 Weeks? **Apply Now at** www.wscubetech.com/data-analytics-course

RANGE



Range is the difference between the maximum and minimum values in the dataset.
It provides a simple measure of the spread of the data, but it can be sensitive to outliers.

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



MEAN ABSOLUTE DEVIATION

The mean absolute deviation of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$



Become a **Data Analyst** in Just 20 Weeks? **Apply Now at** www.wscubetech.com/data-analytics-course

VARIANCE



Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

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



Become a **Data Analyst** in Just 20 Weeks? **Apply Now at** www.wscubetech.com/data-analytics-course

STANDARD DEVIATION

The standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

```
data = [19,10,32,22,43]
```

```
import numpy as np
```

```
np.std(data, ddof=1)
```

```
import statistics as st
```

```
st.stdev(data)
```

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

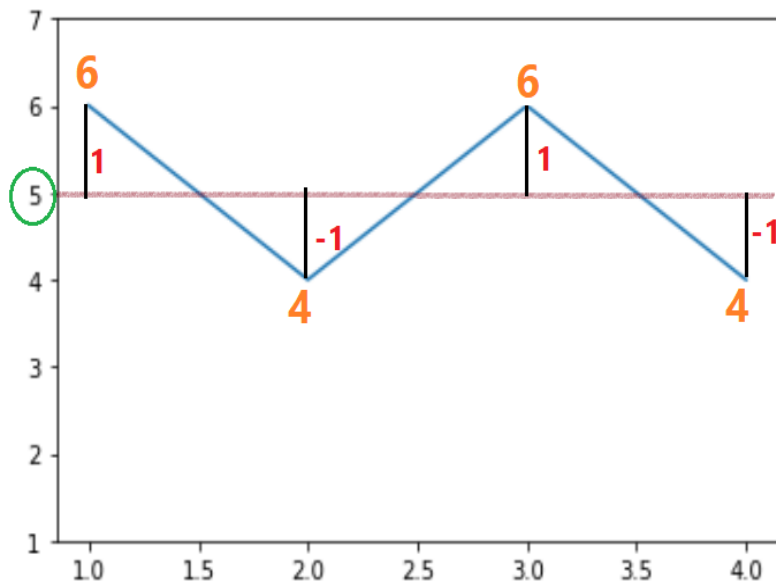
μ = the population mean



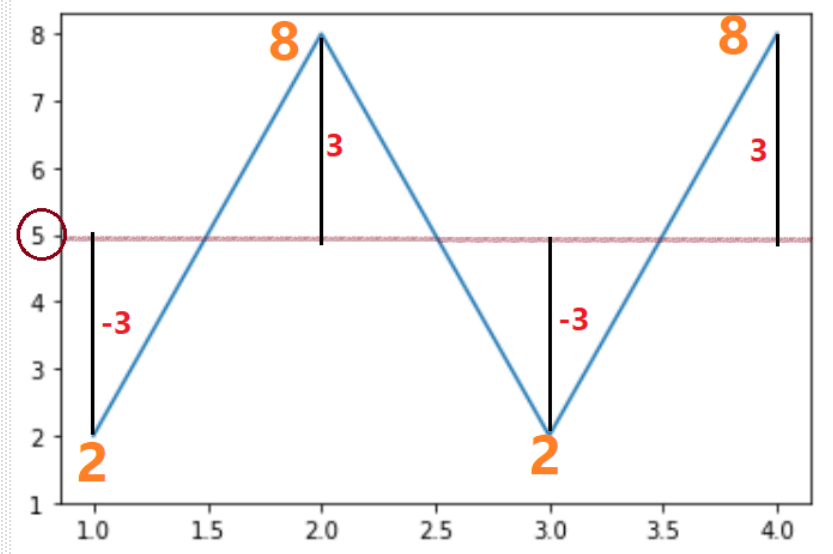
Variance

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

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

Variance

2, 8, 2, 8 Mean = 5

Variance

$$\frac{(2-5)^2 + (8-5)^2 + (2-5)^2 + (8-5)^2}{4}$$

= 9

Variance

Population Data **x1, x2,** x_n

Mean = μ

Variance

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2}{n}$$

n = Number of elements

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

Sample Data

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

Standard Deviation (আদর্শ বিচ্যুতি)

			(Mean)				
Result	38	44	45	50	55	44	74

From which number we can say it is Good or Bad result ?

Mean plus or minus standard deviation is optimum result
but beyond this is either good or bad

Suppose $SD = 2$

Then optimum number is $50 + 2 = 52$ or $50 - 2 = 48$

Standard Deviation (আদর্শ বিচ্যুতি)

A	B		A	B		A	B
6	7			7		9	7
7	9		7	4		7	4
8	3			3		1	5
9	2		9	2		2	6
2	8			3		8	4
5	4			4		3	4
37	33		16	23		30	30

Mean= 8 3.83

Standard Deviation (আদর্শ বিচ্যুতি)

How to calculate the Standard Deviation

$$\begin{array}{c} 2, 8, 2, 8 \quad \text{Mean} = 5 \\ \hline \sigma^2 \text{ Variance} = \frac{(2-5)^2 + (8-5)^2 + (2-5)^2 + (8-5)^2}{4} = 9 \end{array}$$

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

$$\sigma = \sqrt{9} = 3$$