FDA Unit 3 – Basics of Statistics

Points To Be Covered:

1. Introduction to Statistics: Importance of Statistics.
2. Types of Statistics: Descriptive and Inferential.
3. Basic Concepts: Population, Sample, Parameter, Statistics, and Variable.
4. Different types of variables: dependent and independent, extraneous, continuous and discrete, qualitative and quantitative.
5. Concept of noise.
6. Measures of Center: Mean, Mode, Median.
7. Measures of Variation: Variance, Standard Deviation, Range.

Introduction to Statistics – Importance of Statistics

Definition:

Statistics is the branch of mathematics that deals with the collection, classification, presentation, analysis, and interpretation of numerical data to draw meaningful conclusions and support decision-making.

It deals with:

1. Collecting data
2. Organizing data
3. Summarizing data
4. Analysing data
5. Interpreting results
6. Gathering information
7. Arranging it in tables, charts, or graphs
8. Finding averages (mean, median, mode)
9. Looking for trends, patterns, or differences
10. Drawing conclusions and making decisions

Steps in Statistics

1. **Collecting Data**
   - First step in statistics.
   - We gather raw facts and figures from different sources.
   - Methods: surveys, questionnaires, experiments, interviews, sensors, records.
   - Example: A teacher collects marks of all students in a test.
2. **Organizing Data**
   - Collected data is usually messy/unstructured.
   - Organization makes it neat and readable.
   - Methods: tables, frequency distributions, charts, graphs.
   - Example: Instead of writing 50 marks randomly, arrange them in a table or bar graph showing score ranges.
3. **Summarizing Data**
   - Large data sets are hard to compare directly.
   - Summarizing condenses data into key values.
   - Tools: mean, median, mode, range, percentages.
   - Example: Teacher calculates the class average instead of looking at 50 marks individually.
4. **Analysing Data**
   - Study the data carefully to understand what it tells us.
   - Purpose: identify patterns, relationships, differences.
   - Tools: graphs, correlation, regression, hypothesis testing.
   - Example: Comparing boys vs girls marks, or finding if more study hours = higher marks.

5. **Interpreting Results**
   - Final step where numbers turn into conclusions.

- Helps in: decision-making, planning.
- Example: If average marks are low, teacher concludes exam was tough → arranges extra classes.

Purpose of Statistics

1. To simplify complex data
2. To describe and understand
3. To support decision-making
4. To identify trends and relationships
5. To reduce uncertainty
6. To aid in planning and forecasting

To Simplify Complex Data:

- Large amounts of raw data are hard to understand.
- Statistics organizes and summarizes them into simple forms like averages, percentages, and graphs.

To Describe and Understand:

- Helps us describe characteristics of data (Eg: average height of student).
- Makes it easier to understand patterns and differences.

To Support Decision-Making:

- Converts data into evidence-based insights.
- Example: a company uses sales statistics to decide which product to promote.

To Identify Trends and Relationships:

- Finds patterns over time (Eg: rising prices, exam performance trends).
- Measures relationships between variables (Eg: hours studied vs marks) [how two or more factors are connected – whether they increase together, decrease together].

To Reduce Uncertainty:

- Using data (facts and evidence) to make decisions reduces uncertainty and increases accuracy.
- Example: weather forecasts use statistics to predict rain chances.

To Aid in Planning and Forecasting:

- Governments, businesses, and researchers use statistics to plan for the future.
- Example: forecasting population growth, predicting demand, budgeting.

Applications of Statistics:

1. **Business & Industry**
   - Uses: Market research, sales forecasting, quality control.
   - Example: A company studies customer buying habits to decide which product to launch next.

2. **Research & Experiments**
   - Uses: Designing experiments, analysing data, validating findings.
   - Example: Researchers use statistics to analyse survey results and test hypotheses in social science studies.

3. **Healthcare & Medicine**
   - Uses: Collecting, analysing, and interpreting clinical data.
   - Example: Scientists use statistics to test if a new medicine is more effective than an existing one.

4. **Education**
   - Uses: Evaluating student performance, designing fair assessments, improving teaching strategies.
   - Example: Teachers calculate the average marks of students to assess overall class performance.

5. **Government & Planning**
   - Uses: Conducting population studies, unemployment surveys, and budget allocation.

- Example: Census data helps the government plan for schools, hospitals, public transport, and social welfare schemes.

6. **Weather & Environment**
   - Uses: Predicting climate trends, rainfall, and natural disasters like storms or droughts.
   - Example: Meteorologists use statistical models to forecast rain, temperature changes, and storm probabilities.

7. **Sports**
   - Uses: Analysing player performance, forming team strategies, predicting outcomes.
   - Example**:** A cricket team uses batting averages and strike rates to select players for an upcoming tournament.

Importance of Statistics:

1. **Simplifies complex data:** Organizes large amount of information into tables, charts, and graphs for easy understanding.
2. **Support decision-making:** Provides a factual basis for making reliable and less risky decision.
3. **Helps in understanding relationships:** Measures connections between variables (Eg: study time and exam scores).
4. **Aids in forecasting:** Predicts future using past data (Eg: sales, population growth, weather).
5. **Test hypothesis:** Verifies assumptions through statistical tests, making conclusions more scientific.
6. **Guides policy making:** Governments and organizations use statistics to plan budgets resources and social programs.
7. **Used in research:** Essential in scientific studies for collecting analyzing and interpreting data.
8. **Applicable in everyday life:** From sports and education to healthcare and business, statistics help us make sense of daily information.

Basic concepts in Statistics:

1. **Population:**
   - It refers to the individual set of individuals, objects, or data points that you want to study.
   - It can be large or small depending on the scope of research.
   - It is the complete data from which a sample may be drawn.
   - It provides a complete picture and is usually denoted by N.
   - It is used when you have access to data from every member of the population.
   - Example: All the students in the university.

   Key Features of Population:

   1) **Wholeness:** Population means the whole group.
   2) **Defined scope:** It is clearly set by where when and what we are studying.
   3) **Finite or infinite:** It can be countable or uncountable.
   4) **Homogeneity:** All members share something common.
   5) **Types:** Can be the whole target group or the part we can reach.
   6) **Parameters:** We describe populations with parameters like mean or standard deviation.
   7) **Basis for sampling:** We take a sample from the population when studying the whole group is difficult.

   Advantages & Disadvantages:

Advantages	Disadvantages
High accuracy.	Time consuming.
Complete information.	Expensive.
No sampling error.	Impractical for large population.
Better for small groups.	Data management issues.
Useful for policy decisions.	

2. **Sample:**
   - A subset of population that is selected for actual study/analysis.
   - It is meant to represent the population.
   - When your population is large, geographically dispersed, or difficult to contact, it is necessary to use a sample.
   - Example: 500 students chosen from all students in India.

Key Features of Sample:

1) **Part of population:** A sample is subset of the whole population.
2) **Representativeness:** A good sample should reflect the characteristics of the population.
3) **Size (sample size):** The no. of units taken; it should be large enough to give reliable results.
4) **Randomness:** Often selected randomly so that every member has a fair chance to be chosen.
5) **Variability:** Different samples may give slightly diff results, but close to the population value.
6) **Used for inference:** We study the sample to draw conclusions about the whole population.

Advantages & Disadvantages:

Advantages	**Disadvantages**
Time saving.	Sampling error.
Cost effective.	Less accurate.
Practical for large population.	
Quick decision-making.	
Manageable data.	

Difference between Population and Sample:

Aspect	**Population**	**Sample**
Meaning	Whole group under study.	Part of population.
Size	Large (often infinite).	Smaller, manageable.
Study	Difficult, costly, time-consuming.	Easier, cheaper.
Example	All voters in a company.	1000 voters surveyed.

3. **Parameter:**
   - It is a numerical value that describes some characteristics of a population.
   - It is usually fixed but unknown, because we cannot measure the entire population.
   - It is denoted by Greek letters.
   - Example: Population Mean, Proportion, Standard Deviation.

Key Features of Parameter:

1) **Relates to population:** A parameter always describes the whole population, not a sample.
   Example: Average height of all students in a school.
2) **Fixed values:** Since the population is fixed, the parameter is also fixed (but often unknown).
3) **Symbols:** Common symbols used: Mean (μ), Standard Deviation (σ), Proportion (P), Variance (σ^2).
4) **Estimated by statistics:** We take a sample and use statistics to estimate the parameter.
   Example: Sample mean ($\bar{x}$) is used to estimate population mean (μ).

Advantages & Disadvantages:

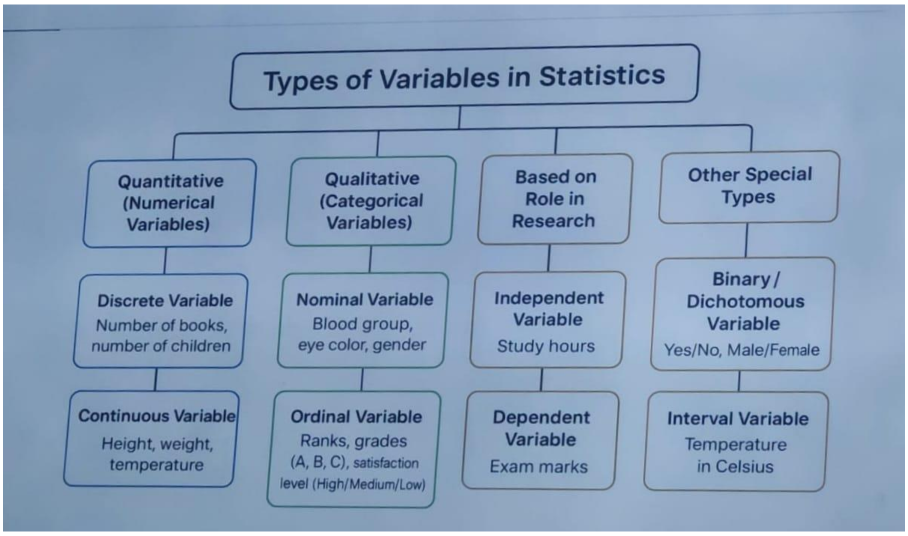
Advantages	**Disadvantages**
High accuracy.	Difficult to obtain.
Reliable for decision making.	Time consuming.
Comprehensive insights.	Expensive.
	Not always possible.

4. **Variable:**
   - A variable is any characteristic, number, or quantity that can be measured or counted and that can change from one observation to another.

Key Features of Variable:

1) **Changeable:** A variable does not stay the same; it may take different values.
   Example: A student’s marks can be 45, 60, or 80.
2) **Types of variables:**
   a) Quantitative (Numerical): Expressed in numbers.
      Discrete → Countable (Eg: No. of blocks).
      Continuous → Measurable (Eg: Height, weight).
   b) Qualitative (Categorical): Expresses qualities or categories.
      Nominal → No order (Eg: Eye colour).
      Ordinal → Has order (Eg: Rank, satisfaction level).
3) **Used for analysis:** Variables provide the data that is analysed to understand populations and samples.
   Example of Variables:
   a) Age of employees (Quantitative, Continuous).
   b) Number of children in a family (Quantitative, Discrete).
   c) Blood group (Qualitative, Nominal).
   d) Rank in a competition (Qualitative, Ordinal).

Different types of Variables in Statistics:


Types of Variables in Statistics
Quantitative (Numerical Variables)
Qualitative (Categorical Variables)
Based on Role in Research
Other Special Types
Discrete Variable
Number of books, number of children
Nominal Variable
Blood group, eye color, gender
Independent Variable
Study hours
Binary / Dichotomous Variable
Yes/No, Male/Female
Continuous Variable
Height, weight, temperature
Ordinal Variable
Ranks, grades (A, B, C), satisfaction level (High/Medium/Low)
Dependent Variable
Exam marks
Interval Variable
Temperature in Celsius

1. **Quantitative (Numerical Variables):**
   Values are numbers that can be measured or counted.
   - **Discrete:** Countable whole numbers.
     Examples: Number of books, number of children.
   - **Continuous:** Measurable values that can take decimals.
     Examples: Height, weight, temperature.

2. **Qualitative (Categorical Variables):**
   Values describe qualities or categories, not numbers.
   - **Nominal:** Categories without order.
     Examples: Blood group, eye colour.
   - **Ordinal:** Categories with natural order/ranking.
     Examples: Ranks, grades (A, B, C).

3. **Independent Variables:**
   Variables that can be manipulated or varied in an experimental study to explore their effects.
   They are called **independent** because they are not influenced by other variables in the study.

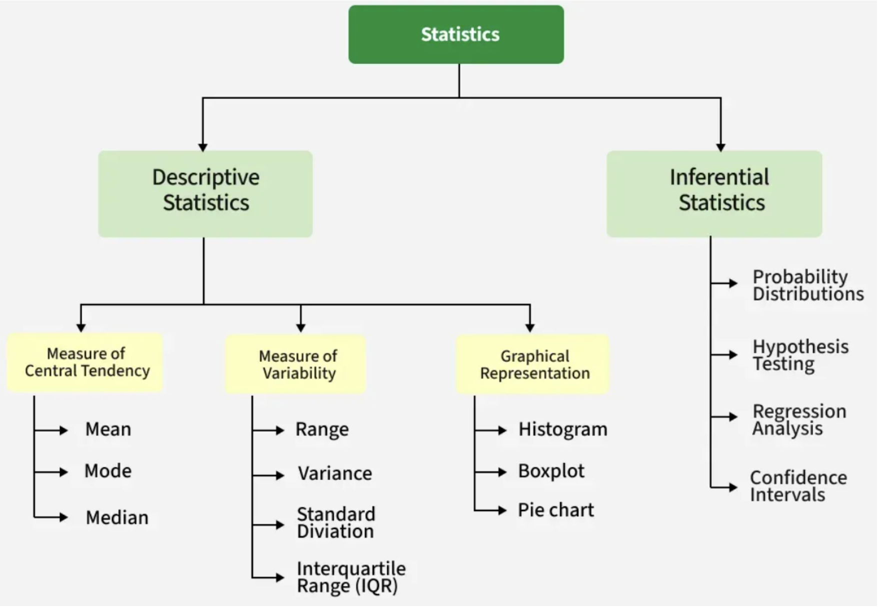
4. **Dependent Variables:**
Variables that are measured or observed in an experiment. They change due to the independent variable and represent the outcome of the experiment.
Examples:
   - Study hours vs. exam marks → Dependent: Exam marks (depends on study hours).
   - Fertilizer vs. crop yield → Dependent: Crop yield.

5. **Binary Variables:**
A type of categorical variable with only two possible values. These typically represent presence/absence, yes/no, or true/false conditions.
   - **Values:** Typically coded as 0 and 1, but can also be Yes/No, True/False, Male/Female.
   - **Type:** Discrete variable, since it can take only two distinct values.

6. **Interval Variables:**
Variables with ordered values and equal intervals between them.
Examples: Temperature in Celsius or Fahrenheit, calendar years (e.g., 2000, 2010).

7. **Extraneous Variables:**
Any factor that is not the independent or dependent variable but has the potential to influence the outcome of a study.
If not controlled, they may create confounding effects (influence both independent and dependent variables).
Example: Testing whether study hours affect exam scores → Extraneous variables could be sleep, IQ, or class attendance, which may also influence exam scores.

Types of Statistics:



1. **Descriptive Statistics:**
The branch of statistics that deals with **collecting, summarizing, and presenting data** in a simple and understandable way without making predictions or generalizations.
   - Descriptive statistics = "Describe the data."
   - Helps in summarizing large datasets into simple numbers and visuals.
   - Answers the question: "What does the data look like?".

Types/Methods of Descriptive Statistics
1) **Measures of Central Tendency** (shows the center or average of the data):
   - **Mean:** Arithmetic average.
   Example: Marks = 10, 20, 30 → Mean = (10+20+30)/3 = 20.

- **Median:** Middle value when data is arranged in order.
  Example: Marks = 10, 20, 30, 40, 50 → Median = 30.
- **Mode:** Most frequently occurring value.
  Example: 10, 20, 20, 30 → Mode = 20.

2) **Measures of Dispersion** (shows how the data varies):
   - **Range:** Difference between largest and smallest values.
   - **Variance:** Average of squared differences from the mean.
   - **Standard Deviation (SD):** Square root of variance; shows variation from the mean.
   - **Coefficient of Variation:** Ratio of SD to mean.
3) **Measures of Position** (shows the relative standing of a value):
   - **Percentiles:** Divide data into 100 equal parts.
     Example: 90th percentile = better than 90% of values.
   - **Quartiles:** Divide data into 4 parts (Q1, Q2 = Median, Q3).
   - **Deciles:** Divide data into 10 equal parts.
4) **Data Presentation** (organizes data for easier understanding):
   - **Tabular form:** Frequency tables, cross-tabulations.
   - **Graphical form:** Bar chart, histogram, pie chart, line graph, frequency polygon/ogive.

2. **Inferential Statistics:**
   The branch of statistics that helps us draw conclusions, make decisions, or predict outcomes about the population based on data collected from a sample.
   Answers the question: “What can we infer or conclude from the data?”

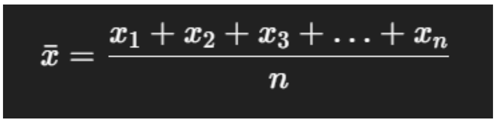
   Common Techniques
   1) **Estimation:** Using sample statistics to estimate population parameters.
      Example: Estimating average height of all students using a sample of 50.
   2) **Hypothesis Testing:** Checking if a claim about a population is likely true.
      Example: Testing whether a new teaching method improves exam scores.
   3) **Confidence Intervals:** Giving a range where the true population parameter likely lies.
      Example: Average weight of students is 60 ± 2 kg with 95% confidence.
   4) **Correlation & Regression:** Studying relationships between variables.

Difference Between Descriptive & Inferential Statistics:

Aspect	Descriptive Statistics	Inferential Statistics
Purpose	Summarize and present existing data.	Draw conclusions, make predictions about population.
Question	What does the data show?	What can we conclude/predict?
Methods	Mean, Median, Mode, Range, Charts, Tables.	Hypothesis testing, Confidence intervals, Regression.
Use	Organizes and presents data clearly.	Tests, predicts, and compares data using samples.
Examples	Average marks of a class, % of boys and girls, graphs.	Predicting election results, testing drug effectiveness.

Mean, Median, Mode:

1. **Mean (Arithmetic Average):**
   - Sum of all values ÷ Number of values.
   - Population mean (μ), Sample mean (x̄).
   - Formula:



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

Examples:

1) Ages of 10 family members: 5, 12, 18, 25, 30, 35, 45, 50, 60, 65.
   → Mean = 34.5.
2) Student's marks: Maths – 80, English – 60, Science – 70, History – 90, Hindi – 85.
   → Mean = 77.
3) Monthly expenses: Food – 2500, Rent – 5000, Travel – 1500, Entertainment – 1000.
   → Mean expense = 2500.

2. **Median (Middle Value):**

- The middle value of sorted data.
- If odd number of values → (n+1)/2.
- If even number of values → Average of (n/2)th and (n/2 + 1)th values.

Example 1: The marks of 5 students are: 12, 18, 10, 15, 20. Find the median.
Data: 12, 18, 10, 15, 20
Step 1: Sort
Sorted data = 10, 12, 15, 18, 20
Step 2: Count
n=5 (odd)
Step 3: Median position
Position = (n + 1)/2 = (5 + 1)/2 = 6/2 = 3 → 3rd value
Step 4: Median
3rd value = 15 → Median = 15

Example 2: The daily wages of 6 workers are: Rs. 200, 250, 300, 400, 500, 600. Find the median wage.
Data: 200, 250, 300, 400, 500, 600
Step 1: Sort
Already arranged → 200, 250, 300, 400, 500, 600
Step 2: Count
n=6 (even)
Step 3: Median position
Positions = n/2 = 6/2 =3 and 3 + 1 = 4 → 3rd & 4th values
Step 4: Median
3rd value = 300, 4th value = 400
Median = (300 + 400)/2 = 700/2 = 350 → Median = 350

Example 3: Ages of children in a park are: 5, 8, 7, 6, 10, 9, 11, 7. Find the median age.
Data: 5, 8, 7, 6, 10, 9, 11, 7
Step 1: Sort
Sorted data = 5, 6, 7, 7, 8, 9, 10, 11
Step 2: Count
n=8 (even)
Step 3: Median position
Positions = n/2 = 8/2 =4 and 4 + 1 = 5 → 4th & 5th values
Step 4: Median
4th value = 7, 5th value = 8
Median = (7 + 8)/2 = 15/2 =7.5 → Median = 7.5
Extra calculations:

- Mean = 63/8=7.87563/8 = 7.87563/8=7.875
- Mode = 7

Example 4: Weekly travel time for 7 days is: 2, 3, 4, 8, 6, 9, 15. Find the median.
Data: 2, 3, 4, 8, 6, 9, 15
Step 1: Sort
Sorted data = 2, 3, 4, 6, 8, 9, 15

Step 2: Count

n=7 (odd)

Step 3: Median position

Position = (n + 1)/2 = (7 + 1)/2 = 8/2 = 4 → 4th value

Step 4: Median

4th value = 6 → Median = 6

Extra calculations:

- Mean = 47/7 = 6.714
- Mode = No Mode

3. **Mode (Most Frequent Value)**
   - Value(s) that occur most often in the dataset.
   - Symbol: Z.
   - Types: Unimodal (1 mode), Bimodal (2 modes), Multimodal (many modes).

Example 1: Simple case

Data: 1, 2, 2, 2, 3, 3, 4, 5

Step 1: Count frequencies

- 1 → 1 time
- 2 → 3 times
- 3 → 2 times
- 4 → 1 time
- 5 → 1 time

Step 2: Find highest frequency

Highest frequency = 3 (value = 2)

Step 3: Mode

→ Mode = 2

Example 2: A cricketer's runs in IPL-20 matches

Data: 45, 60, 20, 55, 70, 10, 40, 30, 90

Step 1: Count frequencies

All values occur only once

Step 2: Check repetition

No value repeats

Step 3: Mode

→ No mode

Other values (given):

- Mean = 46.66
- Median = 45

Example 3: Weekly travel time

Data: 2, 3, 4, 5, 6, 5, 4, 3

Step 1: Count frequencies

- 2 → 1 time
- 3 → 2 times
- 4 → 2 times
- 5 → 2 times
- 6 → 1 time

Step 2: Find highest frequency

Highest frequency = 2 (values = 3, 4, 5)

Step 3: Mode

→ Mode = 3, 4, 5 (multimodal)

Other values (given):

- Mean = 4
- Median = 4

Example 4: Monthly wages of a family
Data: 2500, 5000, 1500, 1000, 2500, 5000
Step 1: Count frequencies

- 1000 → 1 time
- 1500 → 1 time
- 2500 → 2 times
- 5000 → 2 times

Step 2: Find highest frequency
Highest frequency = 2 (values = 2500, 5000)
Step 3: Mode
→ Mode = 2500, 5000 (bimodal)
Other values (given):

- Mean = 2916.66
- Median = 2500

Difference between Mean, Median and Mode:

Aspect	Mean	Median	Mode
Definition	Average of all values	Middle value in an ordered list	Most frequent value
Formula	Sum of values ÷ Number of values	Position = (n+1)/2 (odd), average of middle two (even)	Value(s) with highest frequency
Data Type	Works with numerical data	Works with numerical and ordinal data	Works with numerical, categorical, ordinal
Uniqueness	Always one unique value	Always one unique value	Can be more than one mode (bimodal, multimodal)
Example (3, 4, 4, 6, 100)	Mean = 23.4	Median = 4	Mode = 4

Range:
It is the difference between the highest and the lowest value in the dataset.
Formula: Range = Highest value – Lowest value.

Example 1:
Find the range of the dataset: 12, 19, 6, 2, 15, 4

- Highest value = 19
- Lowest value = 2
- Range = 19 – 2 = 17

Example 2:
Find the range of the dataset: 10, 25, 19, 39, 35

- Highest value = 39
- Lowest value = 10
- Range = 39 – 10 = 29

Example 3:
Marks of 2 sections of students are:

- Section A: 12, 18, 25, 30, 42
- Section B: 15, 20, 28, 33, 39, 50
  Find the combined range.
- Highest value = 50
- Lowest value = 12
- Range = 50 – 12 = 38

Example 4:
Monthly salaries of 5 employees are: 25, 28, 30, 35, 40.
If a new employee joins with a salary of 50, find the range.

- Highest value = 50
- Lowest value = 25
- Range = 50 – 25 = 25

Example 5:
The temperature of a hill station during a week: –5, –2, 0, 4, 6, –3, 2, –5.

- Highest value = 6
- Lowest value = –5
- Range = 6 – (–5) = 11

Example 6:
The height of 2 basketball teams are:

- Team A: 160, 162, 165, 170, 172
- Team B: 168, 170, 175, 180, 185
  Find the range and mean of the combined data.
- Highest value = 185
- Lowest value = 160
- Range = 185 – 160 = **25**
- Mean = (160+162+165+170+172+168+170+175+180+185) ÷ 10 = **170.7**

Example 7:
Find the missing value and range of the dataset: 12, 15, 18, 20, x, 25.
Given mean = 18.
(12+15+18+20+x+25) ÷ 6 = 18
90 + x = 108
x = 18
Range = 25 – 12 = 13

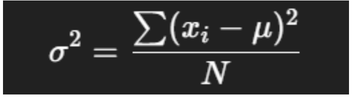
Example 8:
In a classroom, the test scores are: 56, 78, 67, 45, 56, 56, 90, 56, 67, 78, 82.
Find the mean, median, and mode.

- Mean = (sum ÷ n) = 731 ÷ 11 = 66.45
- Median = middle value = 67
- Mode = most frequent = 56, 67

Variance:
Variance measures how far the data values spread out from the mean.

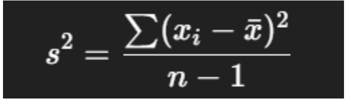
- **Small variance** → values are close to the mean.
- **Large variance** → values are widely spread.
- **Symbols**: Population variance = σ^2, Sample variance = s^2.

Formula:
Population Variance:



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

Sample Variance:



$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Example 1 (Population Variance):
Find the population variance of the dataset: 5, 7, 9, 10, 14, 15.

- Mean = 10

- Squared differences = 25, 9, 1, 0, 16, 25 → Sum = 76
- Variance = 76 ÷ 6 = 12.67 (Low variance)

Example 2 (Sample Variance):
Find the sample variance of the dataset: 4, 6, 8, 10.

- Mean = 7
- Squared differences = 9, 1, 1, 9 → Sum = 20
- Variance = 20 ÷ (4 – 1) = 6.67

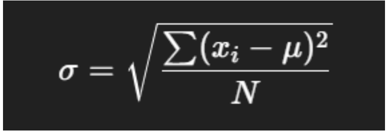
Example 3:
Find the sample variance of the dataset: 7, 11, 15, 19, 24.

- Mean = 15.2
- Squared differences = 67.24, 17.64, 0.04, 14.44, 77.44 → Sum = 176.8
- Variance = 176.8 ÷ (5 – 1) = 44.2

Example 4 (Sample Variance):
If variance of a dataset is 12 and the sum of squared differences is 156, find the number of observations.

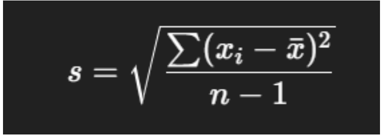
12 = 156/n-1
n-1 = 156/12
n = 13+1
n = 14
The dataset has 14 observations.

Standard Deviation (SD):
Standard Deviation is the square root of variance. It shows how much data deviates from the mean.
Formula:
Population SD:



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

Sample SD:



$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Steps to Calculate SD:

1. Calculate the mean of the data.
2. Subtract mean from each value and square it.
3. Find the average of squared differences (variance).
4. Take square root of variance = SD.

Example 1:
Find variance and SD of {2, 4, 6}.

- Mean = 4
- Squared differences = 4, 0, 4 → Sum = 8
- Variance = 8 ÷ 3 = **2.67**
- SD = √2.67 ≈ **1.63**

Example 2:
Find mean, median, mode, variance, and SD of dataset: 45, 60, 20, 55, 70, 10, 40, 30.

- Mean = (330 ÷ 8) = **41.25**
- Median = average of 4th & 5th = (40 + 45) ÷ 2 = **42.5**
- Mode = none (all unique)

- Variance = **335.36**
- SD = √335.36 ≈ **18.32**