Unit 1: Data Analytics Overview

- Introduction
- Data Analytics Importance
- Types of Data Analytics
    - Descriptive Analytics
    - Diagnostic Analytics
    - Predictive Analytics
    - Prescriptive Analytics
    - Data Analytics Benefits
- Decision-making
- Cost Reduction and other benefits
- Applications of Data Analytics
- Examples of Data Analytics
- Difference between Data Analysis & Data Analytics
- Difference between Data Analyst and Data Scientist
- Difference between Business Analyst and Business Intelligence Analyst
- Role, responsibilities, and skillsets required to be a Data Analyst

Data Analytics Importance

- Definition: Data analytics is the process of collecting, transforming, and analyzing data to extract useful insights for informed decision-making.
- Key Points:
    - Improves decision quality by converting data into actionable information.
    - Enhances operational efficiency, optimizes processes, and reduces costs.
    - Helps understand customer needs for better product and service personalization.
    - Identifies trends for forecasting, risk management, and fraud detection.
    - Provides a competitive advantage by enabling faster and smarter strategies.
- Advantages:
    - Accurate, evidence-based decisions.
    - Cost reduction and resource optimization.
    - Better customer targeting and satisfaction.
    - Early identification of risks and opportunities.
- Disadvantages:
    - High initial investment in technology and skills.
    - Data quality and privacy concerns.
    - Possible bias or misinterpretation of results.

Types of Data Analytics

1. Descriptive Analytics

- Definition: Summarizes historical data to show what happened.
- Focus question: “What happened?”
- Methods: Aggregation, counts, averages, dashboards, trend lines, data visualization.
- Outputs: KPI reports, summaries, heatmaps, time-series trends.
- Example: Monthly sales by region; daily active users trend.
- Advantages: Simple, fast overview; reveals patterns for monitoring.
- Disadvantages: Past-focused only; no causal explanation.

2. Diagnostic Analytics

- Definition: Explores data to find why an event occurred.
- Focus question: “Why did it happen?”
- Methods: Drill-down/roll-up, segmentation, correlation analysis, root-cause analysis, A/B comparisons.
- Outputs: Variance analyses, cohort breakdowns, cause drivers.
- Example: Investigating a sales dip by channel, product, and campaign to find driver.
- Advantages: Identifies drivers of change; guides targeted fixes.
- Disadvantages: Correlation ≠ causation; needs richer, cleaner data.

3. Predictive Analytics

- Definition: Uses statistical/ML models on historical data to forecast future outcomes.
- Focus question: “What is likely to happen?”
- Methods: Regression, time-series (ARIMA/ETS), classification, survival analysis, ensemble models.
- Outputs: Forecasts, probability scores, risk rankings.
- Example: Predicting next-quarter demand or customer churn probability.
- Advantages: Proactive planning; quantifies risk/likelihood.
- Disadvantages: Model risk from drift/bias; accuracy depends on data quality and assumptions.

4. Prescriptive Analytics

- Definition: Recommends actions to achieve desired outcomes, often using predictions as inputs.
- Focus question: “What should we do?”
- Methods: Optimization (linear/integer programming), simulation, reinforcement learning, decision rules.
- Outputs: Action plans, recommended setpoints, optimal schedules.
- Example: Optimizing inventory reorder quantities given forecasted demand and constraints.
- Advantages: Actionable guidance; maximizes efficiency/ROI under constraints.
- Disadvantages: Complex, data- and compute-intensive; sensitive to model/prediction errors.

Quick comparison table

Type of Analytics	Description	Advantages	Disadvantages
Descriptive	Summarizes historical data to show what has happened.	Simple to understand, helps spot trends and patterns.	Limited to past data, cannot explain causes or predict future.
Diagnostic	Examines data to answer why something happened by finding causes or correlations.	Identifies root causes, helps improve processes.	May require complex data, risk of inaccurate cause identification.
Predictive	Uses statistical models and past data to forecast future outcomes.	Aids in planning, risk management, and decision support.	Predictions may be inaccurate if data or model is poor.
Prescriptive	Recommends actions based on predictive analysis and outcomes simulation.	Offers best possible solutions, maximizes efficiency.	Complex to implement, relies on accurate predictive models.

Decision-Making

- Definition:Decision-making is the process of selecting the best action from alternatives, guided by analysis of data and information.

- Key Points:
    - Uses insights derived from data (data-driven decision-making).
    - Steps include:
        - Gathering data
        - Analyzing
        - Generating solutions
        - Evaluating options
        - Choosing action.
    - Applied at all management levels—strategic, tactical, operational.
    - Improves accuracy, speed, and objectivity of choices.

Benefits of Data Analytics (5 Marks)

- **Improved Decision Making: Data** analytics provides accurate, evidence-based insights, enabling informed and timely decisions.

- **Cost Reduction:** Identifies inefficiencies and optimizes resource use, lowering operational and production costs.

- **Enhanced Customer Experience:** Analyzes customer behavior to personalize products and services, increasing satisfaction and loyalty.

- **Risk Management:** Helps forecast potential risks and detect fraud early, reducing financial and operational losses.

- **Increased Revenue:** Enables better marketing, sales strategies, and product innovations that drive business growth.

- **Operational Efficiency:** Streamlines processes and automates tasks, leading to faster and more efficient workflows.

- **Competitive Advantage:** Provides market trends and business intelligence to stay ahead of competitors.

Applications of Data Analytics

Healthcare: Improves patient care by identifying high-risk patients, monitoring treatment effectiveness, and optimizing resource use. Helps in disease outbreak detection and personalized treatment plans.

- **Finance and Banking:** Detects fraud, assesses credit risk, predicts client financial behavior, and helps optimize investment portfolios. Enables personalized financial services and regulatory compliance.

- **E-commerce:** Analyzes customer behavior to personalize shopping experiences, optimize marketing campaigns, and improve customer retention with targeted product recommendations.

- **Cybersecurity:** Detects anomalies and potential cyberattacks by analyzing user behavior and system logs, enabling real-time threat response and security enhancement.

- **Supply Chain and Logistics:** Optimizes inventory management, demand forecasting, transportation routes, and delivery schedules to reduce costs and improve efficiency.

- **Retail**: Enhances pricing strategies, product placement, inventory management, and customer segmentation to boost sales and customer loyalty.

Examples of Data Analytics

- **Analyzing Sales Data:** To find monthly sales trends and identify the best-selling products.

- **Customer Segmentation:** Grouping customers based on purchase behavior to target marketing campaigns.

- **Website Traffic Analysis:** Studying visitor patterns to improve website navigation and user experience.

- **Inventory Analysis:** Monitoring stock levels to reduce overstocking and stockouts.

- **Employee Performance Tracking:** Analyzing productivity data to identify training needs.

Difference between Data Analysis & Data Analytics

Aspect	Data Analysis	Data Analytics
Definition	The process of inspecting, cleaning, transforming, and modeling data to find useful information.	A broader process involving collecting, analyzing, and interpreting data to make informed decisions and solve problems.
Purpose	To understand past data and generate insights.	To discover patterns, predict outcomes, and support decision-making.
Scope	Focuses on examining existing datasets.	Encompasses the entire data lifecycle from collection to actionable insights.
Techniques	Includes descriptive and diagnostic analysis, data cleaning, and visualization.	Uses advanced methods like machine learning, predictive modeling, and prescriptive analytics.
Tools	Excel, SQL, R, Python (Pandas, Matplotlib)	Tableau, Power BI, Python (ML libraries), cloud platforms.
Outcome	Provides reports explaining “what” happened.	Provides insights on “what”, “why”, “what next”, and “what to do”.
Relation	Data analysis is a subset of data analytics.	Data analytics uses data analysis as one of its components.

Difference between Data Analyst and Data Scientist

Aspect	**Data Analyst**	**Data Scientist**
Primary Focus	Analyzes and interprets existing structured data to help make business decisions.	Uses advanced statistical, machine learning, and programming techniques to uncover insights and predict future trends.
Typical Tasks	Data cleaning, reporting, visualization, descriptive and diagnostic analytics.	Predictive modeling, algorithm development, handling big and unstructured data, and creating data products.
Skills Required	SQL, Excel, data visualization tools (Tableau, Power BI), statistical basics.	Strong programming (Python, R), machine learning, statistics, big data technologies (Hadoop, Spark).
Education	Usually a bachelor's degree in statistics, math, or business-related fields.	Often requires advanced degrees (Master's or PhD) in data science, computer science, or related areas.
Outcome	Provides insights and reports to support decision-making.	Builds models and solutions to automate decision-making and solve complex problems.
Tools	Excel, SQL, Tableau, Power BI	Python, R, Jupyter, TensorFlow, Hadoop

Difference between Business Analyst and Business Intelligence Analyst

Aspect	Business Analyst (BA)	Business Intelligence Analyst (BIA)
Primary Focus	Understands business needs, defines requirements, and suggests improvements.	Collects, processes, and analyzes data to provide actionable insights.
Role	Works closely with stakeholders to identify problems and recommend solutions.	Focuses on data extraction, reporting, and creating dashboards.
Skills	Business process understanding, requirement gathering, communication.	Data querying (SQL), visualization tools (Tableau, Power BI), analytical skills.
Outcome	Provides business-case solutions and process improvements.	Delivers reports, trends, and performance analytics for decision support.
Tools Used	JIRA, Confluence, process modeling software.	Power BI, Tableau, SQL, data warehousing tools.
Interaction	Bridges between business teams and IT/development.	Works mainly with data teams and management for reporting.
Scope	Broader business process and strategy improvement focus.	More specialized in data interpretation and visualization.

Role, responsibilities, and skillsets required to be a Data Analyst

- **Role:**
  A Data Analyst collects, processes, and analyzes data to extract meaningful insights that support business decision-making and strategy.

- **Responsibilities:**

  - **Data Collection and Preparation:** Gather data from multiple sources, clean it by handling missing or incorrect values, and prepare it for analysis.

  - **Data Analysis and Interpretation:** Use statistical and analytical techniques to identify patterns, trends, and correlations in the data.

  - **Data Visualization:** Create charts, graphs, and dashboards to present insights in an understandable and actionable way.

  - **Reporting:** Prepare clear reports and presentations to communicate findings to stakeholders for informed decisions.

  - **Database Management:** Maintain data systems and ensure data accuracy and integrity.

  - **Collaboration:** Work with other teams to understand data needs and support business goals through data-driven insights.

- **Skillsets Required:**

  - Proficiency in **SQL** for data querying.

  - Knowledge of **Excel and data visualization tools** like Tableau or Power BI.

  - Understanding of **statistics and data cleaning techniques.**

  - Strong **analytical thinking** and problem-solving skills.

  - Basic **programming** skills in languages such as Python or R (optional but preferred).

  - Good **communication skills** to explain complex data insights clearly.

Unit 2: Dealing with Different Types of Data

- Introduction
- Terminologies in Data Analytics
- Text analytics
- Web analytics
- Marketing analytics
- Types of Data
- Data, numerical data and categorical data
- Structured and unstructured data
- Ordinal data and nominal data
- Data types and data format
- Types of data repositories such as:
    - Databases
    - Data Warehouses
    - Data Marts
    - Data Lakes
    - Data Pipelines

Terminologies in Data Analytics

- **Algorithm:** A set of step-by-step instructions or procedures for solving a problem or performing a task on data.
- **Classification Model:** A model that categorizes data into distinct classes, e.g., spam or non-spam emails.
- **Completeness:** A measure of data quality that checks if all required data is present with no missing values.
- **Conformity:** Data quality dimension assessing if data adheres to standards and formats properly.
- **Machine Learning:** AI branch where models learn from data to identify patterns and make decisions with minimal human input.
- **Metadata:** Data that provides information about other data, like data descriptions and types.
- **Neural Networks:** Algorithms inspired by the human brain that detect complex relationships in data.
- **Online Analytical Processing (OLAP):** Software for fast multidimensional analysis of large volumes of data.
- **Predictive Analytics:** Using past data to predict future outcomes.
- **Prescriptive Analytics:** Using data-driven insights to recommend actions for optimal results.
- **Regression Analysis:** Statistical methods to estimate relationships between variables.
- **Structured Data:** Data neatly organized in rows and columns, easy to process.
- **Timeliness:** Data quality measure reflecting how current and available data is when needed.

Text analytics

- Text Analytics is the **process** of **extracting meaningful insights** from **large volumes** of **unstructured text data** using **computational techniques**.

- It combines **natural language processing (NLP)** and **machine learning** to analyze text such as **customer reviews**, **social media posts**, and **emails**.

- Key steps include **data collection**, **preprocessing** (like **tokenization** and **stopword removal**), **feature extraction**, and **analysis**.

- Common techniques are **sentiment analysis** (detecting **positive or negative feelings**), **topic modeling** (discovering **themes**), and **keyword extraction**.

- Text analytics helps in **business intelligence**, **customer feedback analysis**, and **trend detection** by converting text into **actionable quantitative data insights**.

Web analytics

- Web Analytics is the **process** of website **data collecting, measuring, analyzing, and reporting** to understand and optimize **web usage and performance**.
- It tracks visitor behavior, traffic sources, page views, bounce rates, session duration, and conversion rates to help improve user experience.
- The process begins by setting business goals and key performance indicators (KPIs), followed by data collection, processing, analysis, and implementation of insights.
- Popular tools like Google Analytics provide detailed reports on audience demographics, traffic patterns, user engagement, and conversion tracking.
- Web analytics aids in enhancing website content, increasing visitor retention, optimizing marketing strategies, and ultimately boosting return on investment (ROI).

Marketing analytics

- Marketing Analytics is the process of collecting, analyzing, and interpreting data from various marketing channels to measure the effectiveness of marketing campaigns.
- It helps businesses understand customer behavior, campaign performance, and ROI (Return on Investment) by evaluating metrics like web traffic, conversion rates, social media engagement, and ad performance.
- Different types include web analytics, social media analytics, email marketing analytics, SEO analytics, PPC analytics, and content analytics.
- Using these insights, companies can make data-driven decisions to optimize marketing strategies, improve targeting, allocate budgets efficiently, and enhance customer experience.
- Marketing analytics also enables real-time monitoring, competitive analysis, and campaign optimization, which are essential for staying competitive in dynamic markets.

Data Types

Data can be broadly classified into different types based on its characteristics and how it is analyzed.

- **Numerical Data:** consists of quantifiable values that represent measurable quantities, such as height, weight, or temperature. It supports mathematical operations and is further divided into discrete (countable) and continuous (measurable) data.
- **Categorical Data:** represents distinct groups or categories without a numerical meaning, such as colors, gender, or types of products. It is used for labeling or classifying data.
- **Structured Data:** is organized and stored in predefined formats like databases or spreadsheets, making it easy to search, process, and analyze (e.g., customer records, sales data).
- **Unstructured Data:** lacks a fixed format and includes data types such as emails, videos, social media posts, and documents, which are more complex to analyze due to variability and ambiguity.

- **Ordinal Data:** is a type of categorical data with a defined order or ranking among its categories (e.g., education level: high school, bachelor's, master's). The order matters but the difference between categories is not uniform.
- **Nominal Data:** is categorical data with no intrinsic order or ranking (e.g., blood groups, car brands). It is used only for labeling without any quantitative value.

Data Format

It refers to the physical representation or encoding of data for storage or transmission, such as:

- Text formats (CSV, TXT, JSON, XML) for readable data storage.
- Binary formats for compact data storage such as images (JPEG, PNG) and videos (MP4).
- Database formats like SQL tables for structured data.
- File formats often depend on data type and intended use and affect the accessibility and processing of data.

Data repositories

A data repository is a centralized storage location where data is collected, managed, and maintained for analysis, sharing, and reporting. Data repositories ensure a single source of truth and support data governance, security, and consistency across organizations. Choosing the right type of data repository is essential for efficient data management, accessibility, and decision-making.
Types of data repositories include:

- **Data Warehouse:** Stores large volumes of structured data collected from multiple sources, optimized for analysis and reporting.
- **Data Lake:** Stores raw data in various formats—structured, semi-structured, and unstructured—used for big data analytics and flexible storage.
- **Data Mart:** A subset of a data warehouse focused on specific business areas or departments for quick data access.
- **Data Cube:** Multidimensional data storage used for complex analysis such as OLAP, representing data beyond traditional rows and columns.

Unit 3: Basics of Statistics

- Introduction
- Types of statistics
- Descriptive statistics
- Inferential statistics
- Basic concepts in statistics:
  - Population
  - Sample
  - Parameter
  - Statistic
- Various different types of variables (dependent and independent, extraneous variable, continuous and discrete)
- Qualitative and Quantitative
- Concept of noise
- Measures of Centre:
  - Mean
  - Mode
  - Median
- Measures of variation:
  - Variance
  - Standard deviation
  - Range
- Importance of statistics

Introduction

Types of statistics

Descriptive Statistics:

- Summarizes and organizes existing data to describe its main features.
- Uses measures like mean, median, mode (central tendency) and range, variance, standard deviation (dispersion).
- Visualizes data using charts, graphs, tables for easy understanding.
- Does not make predictions or conclusions beyond the data collected.
- Example: Calculating average marks of students in a class.

Inferential Statistics:

- Uses a sample of data to make generalizations or predictions about a larger population.
- Involves techniques like hypothesis testing, regression, confidence intervals, ANOVA.
- Enables drawing conclusions and making decisions in the face of uncertainty.
- Results include a margin of error or confidence level indicating prediction reliability.
- Example: Predicting election results by analyzing survey samples.

Basic concepts in statistics:

- **Population:** The entire group or complete set of individuals, objects, or events that share a common characteristic and are being studied. It represents the whole from which data is to be collected or analyzed.
    - Populations can be finite (countable members like employees) or infinite (uncountable like bacteria).
    - It is usually denoted by N
    - Examples: All students in a school, all voters in a country, or all plants in a garden.
- **Sample:** A smaller subset or portion selected from the population to represent it in a study. Sampling is used because studying the entire population is often impractical or impossible.
    - Examples: 100 students randomly selected from a school, or 500 voters surveyed in a city.
    - A sample should be random and representative to avoid bias.
    - Sample size is denoted by n, where $n<N$
- **Parameter:** A numerical value that describes a specific characteristic of the population (e.g., population mean μ, population variance σ^2). Parameters are generally unknown and are estimated using sample data.
    - It reflects the true value for the entire population.
- **Statistic:** A numerical value computed from the sample data used to estimate the corresponding population parameter (e.g., sample mean $\bar{x}$ sample variance s^2).
    - Statistics are known values calculated from collected data and help infer or predict population parameters.

Types of variables

1. **Independent Variable:**
    - The variable that is manipulated or controlled by the researcher in an experiment to observe its effect on another variable.
    - It is the cause or input factor.
    - Example: In a study on sleep effects on memory, sleep duration (4, 8, 12 hours) is the independent variable.

2. **Dependent Variable:**
    - The variable that is measured or observed to assess the effect of the independent variable.
    - It is the effect or outcome that depends on changes in the independent variable.
    - Example: The number of words recalled in the memory study is the dependent variable.

3. **Extraneous Variable:**
    - Any variable other than the independent variable that may affect the dependent variable if not controlled.
    - These variables can introduce bias or confounding effects in the experiment.
    - Example: Noise level in the testing room or time of day when conducting the memory test.
    - Controlling extraneous variables is crucial for valid results.

4. **Continuous Variable:**
    - A numerical variable that can take any value within a range, including fractions or decimals.
    - Example: Height, weight, temperature.

5. **Discrete Variable:**
    - A numerical variable that takes countable, distinct values, often integers.
    - Example: Number of students, number of cars.

Qualitative and Quantitative

Feature	**Qualitative Variables**	**Quantitative Variables**
Definition	Variables that describe attributes or qualities	Variables that represent numerical measurements
Data Type	Categorical (labels or names)	Numerical (numbers, counts, measurements)
Examples	Gender, eye color, nationality	Age, height, weight, number of children
Subtypes	Nominal (no order), Ordinal (ordered categories)	Discrete (countable values), Continuous (any value in range)
Purpose	Describe what type or category	Describe how much or how many
Operations Allowed	Grouping, frequency counts, mode	Addition, subtraction, mean, median, standard deviation
Visualization Tools	Bar charts, pie charts	Histograms, line graphs, scatter plots

Concept of noise

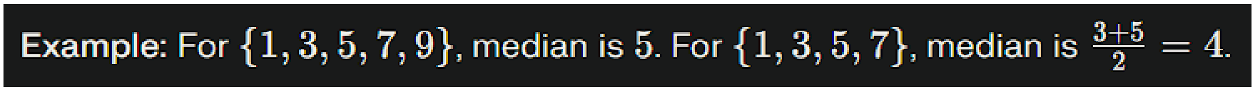
- Noise refers to random irregularities or unwanted variations in data that obscure true signals.
- It is caused by measurement errors, environmental factors, or intrinsic randomness in the data.
- Noise makes data less reliable, complicating analysis and interpretation.
- Recognizing and reducing noise is essential for accurate results, often using statistical techniques like filtering or smoothing.

Measures of Centre:

Mean

- Definition: The mean, or arithmetic mean, is the sum of all values in a dataset divided by the total number of values. It represents the average value and is a measure of central tendency.
- Formula:
    - Sum of all values/total values
- Example: For the data set $\{4,36,45,50,75\}$, the mean is
    - $4+36+45+50+755/5$
    - $=210/5$
    - $=42$

Median

- Definition: The median is the middle value of a data set when the numbers are arranged in ascending or descending order. It divides the dataset into two equal halves.
- Formula:
    - Odd number of values: The median is the middle value, found by taking the ((n + 1) / 2)-th term after sorting the data in ascending order.
    - Even number of values: The median is the average of the two middle values, calculated as ((n/2)th term + ((n/2) + 1)th term) / 2 after sorting the data.
- 

  Example: For $\{1,3,5,7,9\}$, median is 5. For $\{1,3,5,7\}$, median is $\frac{3+5}{2} = 4$.

Mode

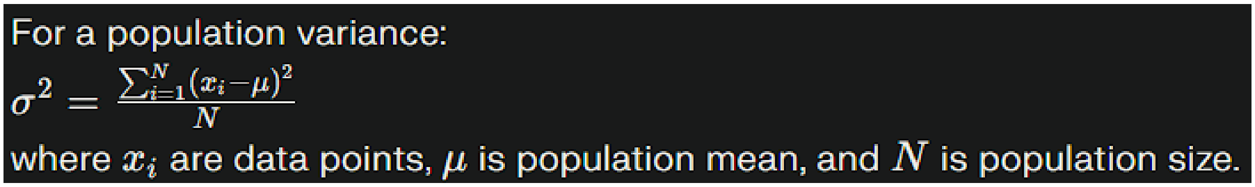
- Definition: The mode is the most frequently occurring value(s) in a dataset. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode.
- Formula: No formula; it is the value(s) with the highest frequency.
- Example: Consider the data set 4, 6, 8, 16, 22, 24, 41, 24, 42, 24, 15, 13, 61, 24, 29. The mode of this data set is 24 because it occurs most frequently (4 times).

Measures of variation:

Range

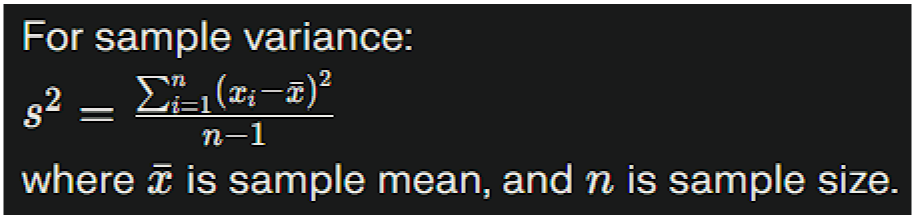
- Range is the simplest measure of variation.
- It is the difference between the maximum and minimum values in a dataset.
- Formula: Range= max val - min val
- It gives a quick idea of how widely data points are spread.
- Example: For data {3,7,10,15,18}, range = 18−3=15

Variance

- Variance measures the average squared deviation of each data value from the mean.
- It shows how much data values spread around the mean on average.
- Larger variance means data points are more spread out; smaller variance means they are closer to the mean.
- 

  For a population variance:

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

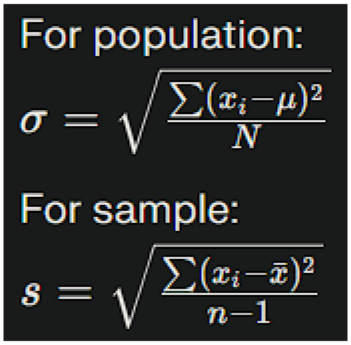
  where x_i are data points, μ is population mean, and N is population size.
- 

  For sample variance:

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

  where $\bar{x}$ is sample mean, and n is sample size.
- Example: If sample data are {2,4,6,8,10}, variance measures average squared distance from sample mean.

Standard Deviation

- Standard deviation is the positive square root of variance.
- It is in the same units as the data, making interpretation easier.
- 

  For population:

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

  For sample:

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

- Smaller standard deviation indicates data points are close to the mean; larger means more spread out.
- Example: If variance is 25, standard deviation is rt25 = 5

Importance of statistics

- Statistics is essential for collecting, analyzing, and interpreting data to understand and solve real-world problems.

- It helps classify and organize data, making it easier to analyze and draw meaningful insights.

- Statistics detects trends, patterns, and anomalies in data, facilitating better decision-making.

- It aids in predicting future outcomes based on historical data, thus supporting strategic planning.

- In various fields like industry, economics, medicine, and technology, it provides critical information for growth, efficiency, and innovation.

- Overall, statistics is fundamental for making informed, data-driven decisions that impact business, scientific, and social activities.