Introduction to statistics:

Statistics is the branch of mathematics that deals with the collection, analysis, interpretation, and presentation of data.

It provides methods for making decisions and predictions based on data.

DATA:

A data is a fact/piece of information that can be stored ,measured and re-accessed.

A data is used to bring insights to increase a company’s revenue by collecting ,organizing and analysing.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **1. Collecting Data** | **2. Organizing Data** | **3. Analyzing Data** |
| **Techniques:** | **Surveys & Questionnaires** – Google Forms, Typeform, Qualtrics  **Web Scraping** – Python (BeautifulSoup, Scrapy)  **IoT & Sensors** – Collecting real-time data from devices  **APIs & Databases** – Google Analytics, SQL databases  **Manual Data Entry** – Excel, Google Sheets | **Data Cleaning** – Handling missing values, removing duplicates  **Data Structuring** – Converting raw data into tables, CSVs, or databases  **Data Storage** – Using databases, spreadsheets, or cloud storage | **Descriptive Statistics** – Mean, Median, Mode, Standard Deviation  **Inferential Statistics** – Hypothesis testing, Regression analysis  **Machine Learning** – Predictive analytics, clustering, classification |
| **Tool** | **Google Forms, SurveyMonkey’ -**Online surveys & feedback collection  **Scrapy, BeautifulSoup -**Web scraping for data  **SQL, PostgreSQL, MySQL-**  Database management  **IoT Sensors, Raspberry Pi**  **-**Real-time data collection | **Microsoft Excel, Google Sheets**  -Sorting, filtering, structuring data  **SQL, NoSQL Databases**  **-**Storing and managing structured data  **Python (Pandas, NumPy)**  -Data manipulation and preprocessing  **Power BI, Tableau-**  Data visualization & organization | **Python (Pandas, NumPy, SciPy, Scikit-learn)**  **-** Statistical & predictive analysis  **R (ggplot2, dplyr, tidyr)**  **-S**tatistical modeling & visualization  **Excel (Pivot Tables, Data Analysis ToolPak)**  **-**Basic statistical analysis  **Tableau, Power BI**  **-** Data visualization & reporting |

**Example:**

**Real-World Case Study:**

**Analyzing Customer Satisfaction for an E-Commerce Business**

**Objective**

An **e-commerce company** wants to analyze **customer satisfaction** based on their shopping experience.

**1. Collecting Data**

✅ **Primary Data (Direct Collection):**

* **Customer Surveys** – After each purchase, customers fill out a survey rating their experience (1 to 5 stars).
* **Website Analytics** – Tracking how long users stay on the website and their interactions.
* **Customer Support Logs** – Recording complaints, issues, and feedback.

✅ **Secondary Data (Existing Data Sources):**

* **Sales Records** – Checking customer purchase history.
* **Competitor Analysis** – Using industry reports to compare with competitors.
* **Social Media Reviews** – Analyzing customer comments and ratings on social platforms.

**Tools Used for Data Collection:**

| **Method** | **Tool** |
| --- | --- |
| Surveys | Google Forms, Typeform |
| Web Analytics | Google Analytics |
| Customer Support Logs | Zendesk, Freshdesk |
| Sales Data | SQL Databases, Excel |
| Social Media Data | Web Scraping (Python, BeautifulSoup) |

**2. Organizing Data**

Once the data is collected, it needs to be cleaned and structured for analysis.

1. **Remove Duplicates** – If a customer filled the survey multiple times, only one entry is kept.
2. **Handle Missing Data** – If some customers skipped questions, missing values are handled using statistical methods.
3. **Categorization** –
   * Grouping customers by age, location, and shopping habits.
   * Sorting satisfaction ratings (1–5 stars).
4. **Visualizing Data** –
   * **Tables** – Showing average rating per month.
   * **Bar Charts** – Number of customers per satisfaction level.
   * **Pie Charts** – Percentage of satisfied vs. unsatisfied customers.

**Example of Organized Data (Table Format)**

| **Month** | **Avg. Satisfaction (1-5)** | **No. of Complaints** | **Avg. Delivery Time (days)** |
| --- | --- | --- | --- |
| Jan | 4.2 | 50 | 3.1 |
| Feb | 4.5 | 40 | 2.8 |
| Mar | 4.1 | 60 | 3.4 |
| Apr | 3.8 | 90 | 4.2 |

**Tools Used for Data Organization:**

| **Method** | **Tool** |
| --- | --- |
| Data Cleaning & Sorting | Excel, Python (Pandas) |
| Categorization | SQL, Python (NumPy) |
| Visualization | Tableau, Power BI, Matplotlib |

**3. Analyzing Data**

company applies **statistical techniques** to extract insights.

🔹 **Descriptive Analysis:**

* **Mean Satisfaction Score** → The company finds that the average rating is **4.1 out of 5**.
* **Complaint Rate** → More complaints were received in April, which aligns with an increase in delivery times.

🔹 **Inferential Analysis:**

* **Regression Analysis** → Shows that **faster deliveries** lead to **higher customer satisfaction**.
* **Hypothesis Testing** → Tests whether **offering discounts** significantly increases repeat purchases.

🔹 **Predictive Analysis (Machine Learning):**

* A **classification model** predicts whether a customer is likely to return based on their shopping history and satisfaction score.

**Analysis Findings:**

✅ **Customers who received deliveries in 3 days or less rated the service 4.5+ on average.**  
✅ **Customers who had a complaint were 60% less likely to shop again.**  
✅ **Offering a 10% discount increased repeat purchases by 15%.**

**Tools Used for Data Analysis:**

| **Method** | **Tool** |
| --- | --- |
| Descriptive Statistics | Excel, Python (Pandas, NumPy) |
| Inferential Statistics | SPSS, Python (Statsmodels, SciPy) |
| Machine Learning | Scikit-learn, TensorFlow |

**4. Business Decision & Outcome**

🔹 **Problem Identified:** Customers were dissatisfied with **longer delivery times** in April, leading to more complaints.  
🔹 **Solution Implemented:** The company **partnered with a faster delivery service** and introduced **free shipping for orders above $50**.  
🔹 **Result:**  
✅ **Customer satisfaction increased from 3.8 to 4.5 in the following months.**  
✅ **Complaints dropped by 30%.**  
✅ **Sales improved by 20% due to better customer retention.**

**Types of Statistics**

|  |  |
| --- | --- |
| **1. Descriptive Statistics** | **2. Inferential Statistics** |
| Deals with **summarizing and presenting data** in a meaningful way.  Orgainzing and summarizing the complete data/population.(ex:average delay of flights/train,  Height/weight of students in class)  ✅ **Measures of Central Tendency** – Find the "center" of the data.   * **Mean (Average)** – Sum of values divided by total count. * **Median** – Middle value in an ordered dataset. * **Mode** – Most frequently occurring value.   ✅ **Measures of Dispersion (Spread of Data)**   * **Range** – Difference between the highest and lowest value. * **Variance** – How far data points are spread from the mean. * **Standard Deviation (SD)** – Measures data variability.   ✅ **Measures of Shape & Symmetry**   * **Skewness** – Measures if data is asymmetrical. * **Kurtosis** – Measures whether data has heavy or light tails.   ✅ **Graphical Representation of Data**   * **Bar Charts, Histograms, Pie Charts** – Used for categorical data. * **Box Plots, Scatter Plots** – Used for numerical data. | **draws conclusions** about a population based on a **sample**.  Using data,has been measured to form conclusion about population(ex- no of trees in forest)  (height/weight of people in india)  Why?  Population is large ,because of time and resource constraints.  Within given sample , data can be concluded something about population  ✅ **Probability Distributions** – Used to predict outcomes.   * Normal Distribution * Binomial Distribution * Poisson Distribution * Pmf * Pdf * Cdf * Ctl * Statisticaltest * **Normal Distribution (Bell Curve)** – Many natural datasets follow this pattern. * **Binomial & Poisson Distributions** – Used in probability-based events.   ✅ **Hypothesis Testing** – Determines if a result is significant.   * **Null Hypothesis (H₀)** – No difference or effect. * **Alternative Hypothesis (H₁)** – There is a significant effect. * **p-value** – If p < 0.05, reject H₀.   ✅ **Confidence Intervals** – Gives a range of values where a population parameter is likely to be.  ✅ **Regression Analysis** – Identifies relationships between variables.  **Linear Regression** – Predicts continuous outcomes.  **Logistic Regression** – Predicts categorical outcomes. |
| **Scope-**  Entire dataset  Graphs used -  Bar charts, Histograms, | **Scope-**  Uses a sample to infer about population  Graphs used-  Confidence Intervals, Probability Distributions |
| **Example :**  📌 **Dataset:** Exam scores → 65, 75, 80, 85, 90 ✅ **Mean =** (65+75+80+85+90) ÷ 5 = 79 ✅ **Median =** 80 (Middle value) ✅ **Range =** 90 - 65 = 25  📊 **Visualization:** A histogram of the scores shows the distribution. | **Example :**  📌 A company surveys **500 customers** to estimate satisfaction for **all customers**. ✅ **Hypothesis:** "Discounts increase repeat purchases." ✅ **p-value < 0.05**, so the effect is significant. ✅ **Regression:** More discounts → Higher retention. |
| Method - ------- Tools Used  Central Tendency ----Excel, Python (NumPy, Pandas)  Dispersion Measures -----R, SPSS, Python (SciPy)  Graphs & Charts ----Tableau, Power BI, Matplotlib | Method ---Tools Used  Probability Distributions-- Python (SciPy, Statsmodels)  Hypothesis Testing--- SPSS, R, Python (t-tests, ANOVA)  Regression Analysis ---Excel, Python (Scikit-learn) |

**Conclusion:**

📌 **Descriptive Statistics** helps summarize **what happened** in the data.  
📌 **Inferential Statistics** helps predict **what will happen** in the larger population.

### ****Why is Statistics Important in Data Science & Analytics?****

Statistics is the **foundation** of Data Science (DS) and Analytics because it helps in **data collection, processing, analysis, and interpretation** to make informed decisions.

It ensures that data-driven insights are reliable and accurate.

**Few applications of statistics in data science/data analytics.**

