What is Statistics

Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, presenting, and organizing data. It is a crucial tool in a wide range of fields, including science, economics, medicine, engineering, social sciences, and more. The main objectives of statistics are to make inferences about populations based on samples, to describe data, and to provide a basis for decision making.

Here are some key concepts and components in statistics:

There are two Major types of Statistics **Descriptive Statistics and Inferential Statistics**

A. **Descriptive Statistics**: These methods summarize and describe the features of a dataset. Common measures include:

- **Mean**: The average of a set of numbers.
- **Median**: The middle value in a set of numbers.
- **Mode**: The most frequently occurring value(s) in a set of numbers.
- **Others include standard deviation, variance etc**
- **Graphs/Plots**

Graphs are primarily part of descriptive statistics. Descriptive statistics are concerned with summarizing and describing the features of a dataset. Graphs are visual tools that help in this summarization by presenting data in an easily understandable and interpretable way.

Types of Graphs in Descriptive Statistics

1. **Histograms**: Show the distribution of a single numerical variable by dividing the data into bins and counting the number of observations in each bin.
2. **Bar Charts**: Used to display and compare the number, frequency, or other measure (e.g., mean) for different discrete categories or groups.
3. **Pie Charts**: Show the proportions of a whole for different categories, useful for categorical data.
4. **Box Plots** (**Box-and-Whisker Plots**): Summarize the distribution of a dataset, showing the median, quartiles, and potential outliers.
5. **Scatter Plots**: Display values for typically two numerical variables for a set of data, showing the relationship or correlation between the variables.
6. **Line Graphs**: Show trends over time or another continuous variable, useful for time series data.

B. Inferential Statistics: These methods allow statisticians to make predictions or inferences about a population based on a sample of data from that population. Key concepts include:

- **Population**: The entire group that you want to draw conclusions about.

- **Sample**: A subset of the population that is used to represent the population.
- **Hypothesis Testing**: A method for testing a hypothesis about a population parameter.
- **Confidence Level**: This indicates the probability that the confidence interval contains the true population parameter. Common confidence levels are 90%, 95%, and 99%. A 95% confidence level, for example, means that if we were to take 100 different samples and compute a confidence interval for each sample, approximately 95 of the 100 confidence intervals would contain the true population parameter.

2. **Regression Analysis**: A set of statistical processes for estimating the relationships among variables. It includes linear regression, multiple regression, and more.
3. **Correlation**: A measure of the strength and direction of association between two variables.
4. **Data Collection Methods**: Techniques for gathering data, such as surveys, experiments, observational studies, and simulations.
5. **Statistical Tests**: Procedures for making decisions about hypotheses, including t-tests, chi-square tests, ANOVA, and others.

Statistics involves various tools and methodologies to handle data effectively, ensuring that conclusions and decisions are based on rigorous analysis and evidence. It is fundamental in research, business decision-making, policy formulation, and many other areas.

In this program we will do both Descriptive and Inferential Statistics.

First, We start with Descriptive Statistics

Exploratory Data Analysis (EDA) within the context of descriptive statistics focuses on summarizing and visualizing the main characteristics of a dataset. This helps in understanding the data's structure, identifying patterns, detecting anomalies, and forming hypotheses for further analysis.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential step in the data analysis process. It involves using statistical and graphical techniques to summarize and visualize the main characteristics of a dataset, often with the aim of uncovering patterns, spotting anomalies, testing hypotheses, and checking assumptions before formal modeling or hypothesis testing.

Key Steps in EDA

1. **Data Collection**: Gather the dataset you intend to analyze. This can come from various sources such as databases, CSV files, APIs, etc.
2. **Data Cleaning**: Prepare the data for analysis by handling missing values, correcting errors, dealing with outliers, and ensuring consistency.
3. **Data Summarization**: Calculate summary statistics to understand the basic features of the data.

- **Measures of Central Tendency**: Mean, median, mode.
- **Measures of Dispersion**: Range, variance, standard deviation, quartile range.

4. **Data Visualization**: Create graphical representations to explore the data visually.

- **Histograms**: To understand the distribution of a single numerical variable.
- **Bar Charts**: To compare frequencies or proportions of categorical variables.
- **Box Plots**: To visualize the spread and identify potential outliers.
- **Scatter Plots**: To examine relationships between two numerical variables.
- **Correlation Heatmaps**: To show correlation coefficients between multiple variables.

Tools and Techniques for EDA

1. **Python Libraries**:

- **Pandas**: For data manipulation and analysis.
- **Matplotlib** and **Seaborn**: For data visualization.
- **NumPy**: For numerical operations.

2. **Statistical Techniques**:

   - **Univariate Analysis**: Examining each variable individually (e.g., histograms, box plots).
   - **Bivariate Analysis**: Exploring relationships between two variables (e.g., scatter plots, correlation analysis).
   - **Multivariate Analysis**: Analyzing more than two variables simultaneously (e.g., pair plots, correlation matrices).

Descriptive Statistics Practicals

Secondly, We will do Inferential Statistics

Inferential statistics is a branch of statistics that involves using sample data to make generalizations (inferences) about a larger population. It is used to draw conclusions and make predictions or decisions about a population based on the analysis of a representative sample.

Key concepts and techniques in inferential statistics include:

1. **Sampling**: The process of selecting a subset of individuals or items from a larger population.
2. **Hypothesis Testing**: Making decisions or drawing conclusions about a population based on sample data.
3. In statistical hypothesis testing, the **null hypothesis** (denoted as $H0$) and the **alternative hypothesis** (denoted as $H1$ or Ha) are two statements about a population. These hypotheses are tested using sample data to determine which statement is supported by the evidence

4. This involves:

- Formulating a hypothesis (null hypothesis and alternative hypothesis).
- Selecting an appropriate test statistic and significance level.
- Calculating the p value from the sample data.
- The p-value is a measure that helps determine the significance of results in statistical hypothesis testing. It quantifies the evidence against the null hypothesis and is used to decide whether to reject the null hypothesis or not
- Making a decision to reject or fail to reject the null hypothesis based on the test statistic and the significance level.

5. **Confidence Intervals**: Estimating the range within which a population parameter lies with a certain level of confidence.

6. Tools and Techniques for Inferential Statistics

1. **Python Libraries**:

- **Pandas**: For data manipulation and analysis.
- **Matplotlib** and **Seaborn**: For data visualization.
- **NumPy**: For numerical operations.
- **SciPy**: For scientific and technical computing.

Inferential Statistics Practicals