**Statistics:**

Statistics is the science of collecting organizing and analyzing data, There are two types of statistics

1. **Descriptive stats**
2. **Inferential Stats**

**Descriptive stats:**

It consists of organizing and summarising the data**.** They provide simple summaries of the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

With descriptive statistics, you are simply describing what is or what the data shows

1. **Measures of Central Tendency**
2. **Measures of Dispersion**
3. **Measures of Distribution Shape**
4. **Measures of Association**
5. **Frequency Distribution**
6. **Graphical Representation:**
7. **Estimation:**

**Measures of Central Tendency:**

**Mean:** The arithmetic average of a set of values. It is calculated by summing all values and dividing by the number of observations.

**Median:** The middle value in a dataset is when the values are arranged in ascending or descending order. It divides the dataset into two equal halves.

**Mode:** The most frequently occurring value in a dataset.

**Measures of Dispersion:**

**Range:** The difference between the maximum and minimum values in a dataset.

**Variance:** A measure of the spread or dispersion of values around the mean. It is the average of the squared differences from the mean.

**Standard Deviation:** The square root of the variance. It measures the average distance of values from the mean and provides information about the spread of data.

**Interquartile Range (IQR):** The range between the first quartile (Q1) and the third quartile (Q3). It represents the middle 50% of the data and is less sensitive to outliers compared to the range.

**Percentiles:** Percentiles divide a dataset into hundred equal parts. For instance, the 25th percentile is the value below which 25% of the data falls.

These components provide insights into the central tendency, variability, shape, and distribution of data, helping researchers and analysts understand and interpret the characteristics of the dataset.

**Measures of Distribution Shape:**

**Skewness:** A measure of the asymmetry of the distribution. Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side.

**Kurtosis:** A measure of the peakedness or flatness of the distribution. Higher kurtosis indicates sharper, more peaked distributions, while lower kurtosis indicates flatter distributions.

**Measures of Association:**

**Covariance:** A measure of how much two random variables change together. Positive covariance indicates that one variable increases as the other increases, while negative covariance indicates an inverse relationship.

**Correlation Coefficient:** A standardized measure of the strength and direction of the linear relationship between two variables.

* It ranges from -1 to +1, where +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

**Graphical Representation:**

**Histogram:** A graphical representation of the frequency distribution of a continuous variable. It consists of bars where the height represents the frequency of each interval.

**Box Plot:** A graphical summary of the distribution of a dataset. It displays the median, quartiles, and outliers of the data.

**Scatter Plot:** A graphical representation of the relationship between two variables. It helps visualize patterns, correlations, and outliers in the data.

Descriptive statistics provide insights into the central tendency, dispersion, shape, and distribution of a dataset, allowing researchers and analysts to understand and interpret the underlying characteristics of the data.

**Frequency Distribution:**

The number of times each value occurs in a dataset.

**Relative Frequency:** The proportion of times each value occurs relative to the total number of observations.

**Cumulative Frequency:** The sum of the frequencies up to a certain value in the dataset.

**Types of Distributions:**

1. **Normal Distribution (Gaussian Distribution):**
2. **Binomial Distribution:**
3. **Poisson Distribution:**
4. **Uniform Distribution:**
5. **Exponential Distribution:**
6. **Log-Normal Distribution:**
7. **Gamma Distribution:**

**Normal Distribution:**

Asymmetric bell-shaped distribution is characterized by its mean and standard deviation. It is often used to model natural phenomena and is central to many statistical tests.

**Inferential Stats:**

it consists of making some conclusions using some experiments based on collected data.

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think.

Inferential statistics involves making inferences and drawing conclusions about a population based on sample data. It uses probability theory to generalize from a sample to a larger population, and it is commonly used in hypothesis testing, estimation, and prediction. Here are the key components of inferential statistics:

**Basics of International Stats:**

1. Sampling Methods:
2. Central Limit Theorem (CLT):
3. Probability:
4. Estimation:

**Types of Inferential Statistics:**

1. Hypothesis Testing:
2. Regression Analysis:

**Sampling Methods:**

**Random Sampling:** Each member of the population has an equal chance of being selected for the sample.

**Stratified Sampling:** The population is divided into subgroups (strata), and random samples are taken from each subgroup.

**Cluster Sampling:** The population is divided into clusters, and a random sample of clusters is selected for the study.

**Systematic Sampling:** Every nth member of the population is selected for the sample.

**Convenience Sampling:** Individuals are selected for the sample based on their availability or accessibility.

**Central Limit Theorem (CLT):**

A fundamental concept states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution. The CLT is essential for making inferences about population means and proportions.

**Probability Theory:**

Probability theory forms the foundation of inferential statistics, allowing for the quantification of uncertainty and the calculation of probabilities associated with different events or outcomes. Probability is used throughout inferential statistics, from sampling methods to hypothesis testing and estimation.

**Estimation:**

Estimation in inferential statistics involves using sample data to estimate unknown population parameters. The goal is to make educated guesses or predictions about the characteristics of a population based on information gathered from a sample.

**Point Estimation:**

Estimating a population parameter (e.g., mean, proportion) using a single value based on sample data.

Point estimation involves using a single value, typically derived from sample data, to estimate the value of an unknown population parameter. For example, if we want to estimate the population mean, we might use the sample mean as our point estimate. Similarly, if we want to estimate the population proportion, we might use the sample proportion.

Example: If we calculate the mean height of a sample of 100 individuals and find it to be 170 cm, we might use this value (170 cm) as our point estimate for the population's mean height.

**Interval Estimation:**

Estimating a range of values within which the population parameter is likely to fall, along with a level of confidence.

Interval estimation involves providing a range of values, known as a confidence interval, within which the true value of the population parameter is likely to lie. Unlike point estimation, which provides a single estimate, interval estimation acknowledges the uncertainty inherent in estimation by providing a range of plausible values.

Example: If we calculate a 95% confidence interval for the population mean height based on our sample data, we might find that the interval ranges from 165 cm to 175 cm. This means that we are 95% confident that the true population mean height falls within this interval.

**Hypothesis Testing:**

**Null Hypothesis (H0):**

A statement that there is no significant difference or effect.

**Alternative Hypothesis (H1 or Ha):**

A statement that contradicts the null hypothesis and suggests there is a significant difference or effect.

Test Statistic:

A measure calculated from sample data to assess the likelihood of observing the results if the null hypothesis is true.

**P-value:**

The probability of observing the test statistic or more extreme results if the null hypothesis is true. It indicates the strength of evidence against the null hypothesis.

**Significance Level (α):**

The predetermined threshold for rejecting the null hypothesis. Common values include 0.05 or 0.01.

**Types of Errors:**

**Type I Error:** it happens when the null hypothesis of an experiment is true but rejected often called a false positive.

**Type II Error:** it occurs when the null hypothesis is false but still not rejected, also known as a false negative.

**Which error is considered to be more dangerous?**

Type 1 error is considered to be worse or more dangerous than type 2 because rejecting what is true is more harmful than keeping the data that is not true.

**Confidence Intervals:**

**Confidence Level:**

The probability that the confidence interval contains the true population parameter.

**Margin of Error:**

The range of values added to and subtracted from the point estimate to construct the confidence interval.

**ANOVA (Analysis of Variance):**

A statistical technique used to compare means across multiple groups to determine if there are statistically significant differences between them.

**One-Way ANOVA:**

Compares means across two or more independent groups.

**Two-Way ANOVA:**

Considers two independent variables simultaneously to analyze their effects on a dependent variable.

**Nonparametric Methods:**

Statistical methods that do not rely on assumptions about the distribution of the data or the parameters of the population.

Examples include the Wilcoxon rank-sum test, Kruskal-Wallis test, and Mann-Whitney U test.

Inferential statistics allows researchers to make inferences about populations based on sample data, providing insights into relationships, differences, and trends that may exist within the data.

**Regression Analysis:**

**Linear Regression:**

Modeling the relationship between a dependent variable and one or more independent variables using a linear equation.

**Logistic Regression:**

Modeling the relationship between a binary dependent variable and one or more independent variables using a logistic function.

**Multiple Regression:**

Modeling the relationship between a dependent variable and two or more independent variables.

Variables:

There are two types of variables i.e Quantitative Variables and Qualitative Variables(Categorical Variables)

Quantitative Variables: Age, weight, No of Student, Rainfall, Temp, etc.

Quantitative Variables can be divided into two types

- Discrete Variables: usually a whole number and Discrete Variables are countable in a finite amount of time. For example, you can count the change in your pocket. A number of bank accounts. but the point is—it’s still countable.

it us No of bank Accounts, No of children.

- Continuous Variable: A variable that can take an uncountable set of values or an infinite set of values.

Example: Height or weight, Rainfall

Qualitative Variables(Categorical Variables): Gender, Different types of flowers, types of cars, Names of departments, Genre of movies.