➢ Indroduction Of Statistics

- **Statistics**

Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life. It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation. Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Basic terminology of Statistics is Population and Sample

- **Population –**
  It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.
- **Sample –**
  It is the subset of a population.

Types of Statistics :

There are two types of Statistics Inferential statistics and descriptive Statistics

➢ Differentiate between descriptive statistics and inferential Statistics

- Descriptive Statistics gives information about raw data regarding its description or features. Inferential statistics, on the other hand, draw inferences about the population by using data extracted from the population.
- We use descriptive statistics to describe a situation, while we use inferential statistics to explain the probability of occurrence of an event.
- As for descriptive statistics, it helps to organize, analyze and present data in a meaningful manner. Inferential statistics helps to compare data, make hypotheses and predictions.
- Descriptive statistics explains already known data related to a particular sample or population of a small size. Inferential statistics, however, aims to draw inferences or conclusions about a whole population.

- We use charts, graphs, and tables to represent descriptive statistics, while we use probability methods for inferential statistics.

➢ Differentiate between Population and Sample

The measurable characteristic of the population like the mean or standard deviation is known as the parameter. The measurable characteristic of the sample is called a statistic.

Population data is a whole and complete set. The sample is a subset of the population that is derived using sampling.

A survey done of an entire population is accurate and more precise with no margin of error except human inaccuracy in responses. However, this may not be possible always. A survey done using a sample of the population bears accurate results, only after further factoring the margin of error and confidence interval.

The parameter of the population is a numerical or measurable element that defines the system of the set. The statistic is the descriptive component of the sample found by using sample mean or sample proportion.

➢ Hypothesis

A hypothesis is an assumption, an idea that is proposed for the sake of argument so that it can be tested to see if it might be true. In the scientific method, the hypothesis is constructed before any applicable research has been done, apart from a basic background review. You ask a question, read up on what has been studied before, and then form a hypothesis. A hypothesis is usually tentative; it's an assumption or suggestion made strictly for the objective of being tested.

Differentiate betwwen null hypothesis and alternative hypothesis

- A null hypothesis is a statement, in which there is no relationship between two variables. An alternative hypothesis is a statement; that is simply the inverse of the null hypothesis, i.e. there is some statistical significance between two measured phenomenon.

- A null hypothesis is what, the researcher tries to disprove whereas an alternative hypothesis is what the researcher wants to prove.

- A null hypothesis represents, no observed effect whereas an alternative hypothesis reflects, some observed effect.

- If the null hypothesis is accepted, no changes will be made in the opinions or actions. Conversely, if the alternative hypothesis is accepted, it will result in the changes in the opinions or actions.

- As null hypothesis refers to population parameter, the testing is indirect and implicit. On the other hand, the alternative hypothesis indicates sample statistic, wherein, the testing is direct and explicit.

➢ Differentiate between type I and type II eroors

- Type I error is an error that takes place when the outcome is a rejection of null hypothesis which is, in fact, true. Type II error occurs when the sample results in the acceptance of null hypothesis, which is actually false.
- Type I error or otherwise known as false positives, in essence, the positive result is equivalent to the refusal of the null hypothesis. In contrast, Type II error is also known as false negatives, i.e. negative result, leads to the acceptance of the null hypothesis.
- When the null hypothesis is true but mistakenly rejected, it is type I error. As against this, when the null hypothesis is false but erroneously accepted, it is type II error.
- Type I error tends to assert something that is not really present, i.e. it is a false hit. On the contrary, type II error fails in identifying something, that is present, i.e. it is a miss.

- Greek letter ‘α’ indicates type I error. Unlike, type II error which is denoted by Greek letter ‘β’.

➢ Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

➢ Central Limit Theorem

The central limit theorem, the mean of a sample of data will be closer to the mean of the overall population in question, as the sample size increases, notwithstanding the actual distribution of the data. In other words, the data is accurate whether the distribution is normal or aberrant.

As a general rule, sample sizes of around 30-50 are deemed sufficient for the CLT to hold, meaning that the distribution of the sample means is fairly normally distributed. Therefore, the more samples one takes, the more the graphed results take the shape of a normal distribution. Note, however, that the central limit theorem will still be approximated in many cases for much smaller sample sizes, such as n=8 or n=5.

The central limit theorem is comprised of several key characteristics. These characteristics largely revolve around samples, sample sizes, and the population of data.

Statistical significance

Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the null hypothesis and the alternate hypothesis are stated, and the p-value is calculated. After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-

tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant.

Mean

In statistics, the mean is one of the measures of central tendency, apart from the mode and median. Mean is nothing but the average of the given set of values. It denotes the equal distribution of values for a given data set. To calculate the mean, we need to add the total values given in a datasheet and divide the sum by the total number of values.

Standard Deviation

Standard Deviation is a measure which shows how much variation such as spread, dispersion, spread, from the mean exists. The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set. Like the variance, if the data points are close to the mean, there is a small variation whereas the data points are highly spread out from the mean, then it has a high variance. Standard deviation calculates the extent to which the values differ from the average. Standard Deviation, the most widely used measure of dispersion, is based on all values. Therefore a change in even one value affects the value of standard deviation. It is independent of origin but not of scale. It is also useful in certain advanced statistical problems.

Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related meaning they change together at a constant rate. It’s a common tool for describing simple relationships without making a statement about cause and effect. The sample correlation coefficient, *r*, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

Covariance

Covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

Inferential Statistics

Inferential statistics is a branch of statistics that makes the use of various analytical tools to draw inferences about the population data from sample data. Inferential statistics is used for comparing the parameters of two or more samples and makes generalizations about the larger population based on these samples. There are two main types of inferential statistics - hypothesis testing and regression analysis. The samples chosen in inferential statistics need to be representative of the entire population.

One Sample t-test

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value. The one sample *t* test differs from most statistical hypothesis tests because it does not compare two separate groups or look at a relationship between two variables. It is a straightforward comparison between data gathered on a single variable from one population and a specified value defined by the researcher.

Relationship between Standard deviation and Standard variance

The relationship between the variance and the standard deviation for a sample data set -

Variance represents the average squared deviations from the mean value of data, while standard deviation represents the square root of that number.Both, the variance and the standard deviation measures variability in a distribution.Both have different units like the standard

deviation has the same units as the original values like minutes or meters while the variance has much larger units like meters squared.The variance is equal to the square of standard deviation or the standard deviation is the square root of the variance.
"The variance is equal to the square of the standard deviation" is widely used as the relationship between the variance and the standard deviation for a sample data set.

One-way ANOVA test

One-Way ANOVA test compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. Some of the assumptions followed are the adherence to the samples selected from a normally distributed population, independence of the samples, homogeneity of variance, etc.The dependent variable should be continuous, and the one categorical independent variable selected should have three levels or groups. It is commonly used in research fields like science, biology, business economics, and psychology to analyze datasets.