Statistics in Data Analysis

By Younus Chilakala

Introduction to Statistics

Statistics is the science of collecting, analyzing, presenting, and interpreting data. Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation.

Data may be classified as either quantitative or qualitative. Quantitative data measure either how much or how many of something, and qualitative data provide labels, or names, for categories of like items.

The two major areas of statistics are **descriptive** and **inferential** statistics.

Descriptive vs Inferential Statistics

Descriptive statistics:

Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data.

- Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median and mode.
- Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as variance, range and standard deviation.
- The distribution refers to the overall "shape" of the data, which can be depicted on a chart such as a histogram or dot plot, and includes properties such as the probability distribution function, skewness, and kurtosis.

Inferential Statistics:

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Methods and tools involve:

- Regression Analysis: Regression models show the relationship between a set of independent variables and a dependent variable. This statistical method lets you predict the value of the dependent variable based on different values of the independent variables. Hypothesis tests are incorporated to determine whether the relationships observed in sample data actually exist in the data set.
- Hypothesis Tests: Hypothesis testing is used to compare entire populations or assess relationships between variables using samples. Hypotheses or predictions are tested using statistical tests so as to draw valid inferences.

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).

Assumptions of Central Limit Theorem

There are four assumptions associated with a linear regression model:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

Hypothesis

In Statistics, a hypothesis is defined as a formal statement, which gives the explanation about the relationship between the two or more variables of the specified population. It helps the researcher to translate the given problem to a clear explanation for the outcome of the study.

Null Hypothesis: The null hypothesis is the claim that there’s no effect in the population. If the sample provides enough evidence against the claim that there’s no effect in the population ($p \leq \alpha$), then we can reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

- Null hypothesis (H_o): There’s **no effect** in the population.

Alternative Hypothesis: The alternative hypothesis claims that there’s an effect in the population. Often, your alternative hypothesis is the same as your research hypothesis. In other words, it’s the claim that you expect or hope will be true.

- Alternative hypothesis (H_a or H_1): There’s an **effect** in the population.

Type 1 and Type 2 errors

In Statistics , a **Type I error** is a false positive conclusion, while a **Type II error** is a false negative conclusion.

Using hypothesis testing, you can make decisions about whether your data support or refute your research predictions with null and alternative hypotheses. For example, you test whether a new drug intervention can alleviate symptoms of an autoimmune disease.

In this case:

- The null hypothesis (H_0) is that the new drug has no effect on symptoms of the disease.
- The alternative hypothesis (H_1) is that the drug is effective for alleviating symptoms of the disease.

A **Type I error** happens when you get false positive results: you conclude that the drug intervention improved symptoms when it actually didn’t. These improvements could have arisen from other random factors or measurement errors.

A **Type II error** happens when you get false negative results: you conclude that the drug intervention didn’t improve symptoms when it actually did. Your study may have missed key indicators of improvements or attributed any improvements to other factors instead.

Central limit theorem

In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population's actual distribution shape.

The central limit theorem is useful when analyzing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases.

THANK YOU.