Statistical Inference

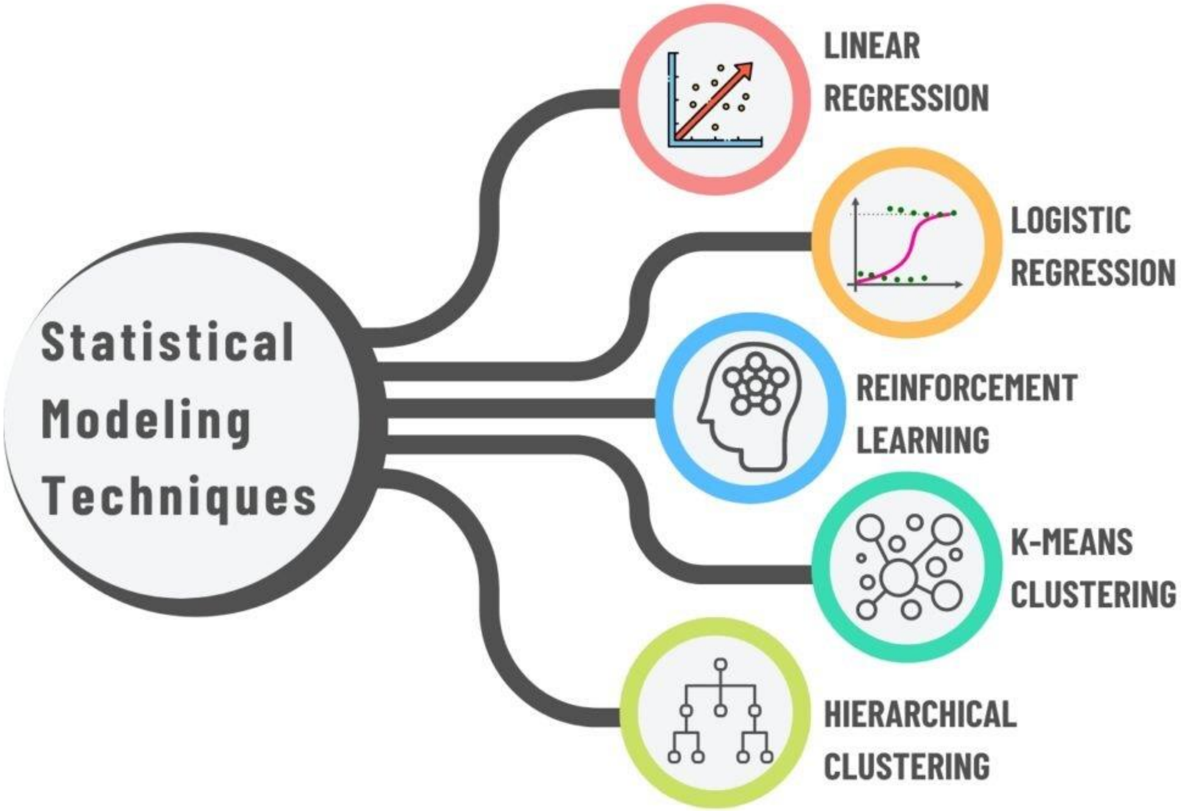
Populations and samples:

A population is the entire group that you want to draw conclusions about. A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population. In research, a population doesn't always refer to people.

Population
Sample

Statistical modeling

- Statistical modeling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment.



Probability Distribution

A probability distribution is an idealized frequency distribution. A frequency distribution describes a specific sample or dataset. It's the number of times each possible value of a variable occurs in the dataset. The number of times a value occurs in a sample is determined by its probability of occurrence.

Fitting a model

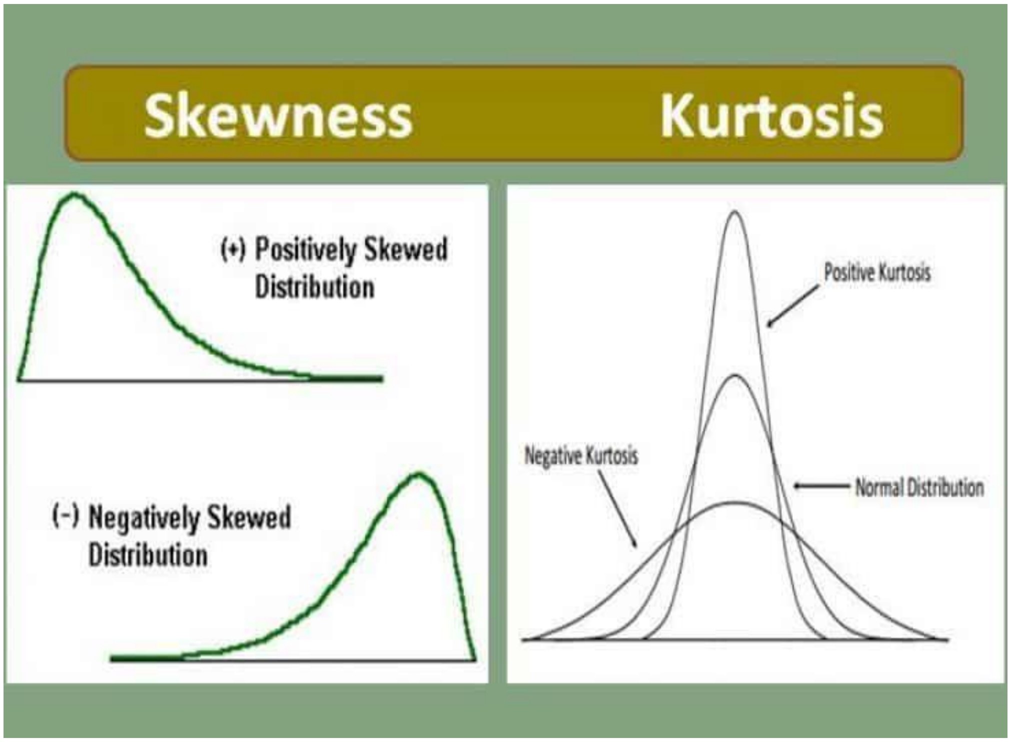
Fitting a model to data means choosing the statistical model that predicts values as close as possible to the ones observed in your population. While doing a statistical analysis, it's very important to make sure about the godness of fit of the model used.

What is Model Fitting?

- Model Fitting is a measurement of how well a machine learning model adapts to data that is similar to the data on which it was trained. The fitting process is generally built-in to models and is automatic. A well-fit model will accurately approximate the output when given new data, producing more precise results.
- A model is fitted by adjusting the parameters within the model, leading to improvements in accuracy.
- During the fitting process, the algorithm is run on test data, otherwise known as "labeled" data. Once the algorithm has finished running, the results need to be compared to real and observed values of the target variable, in order to identify the accuracy of the model.
- Using the results, the parameters of the algorithm can be further adjusted to better uncover relationships and patterns between the inputs, outputs, and targets. The process can be repeated until valid and accurate insights are given.

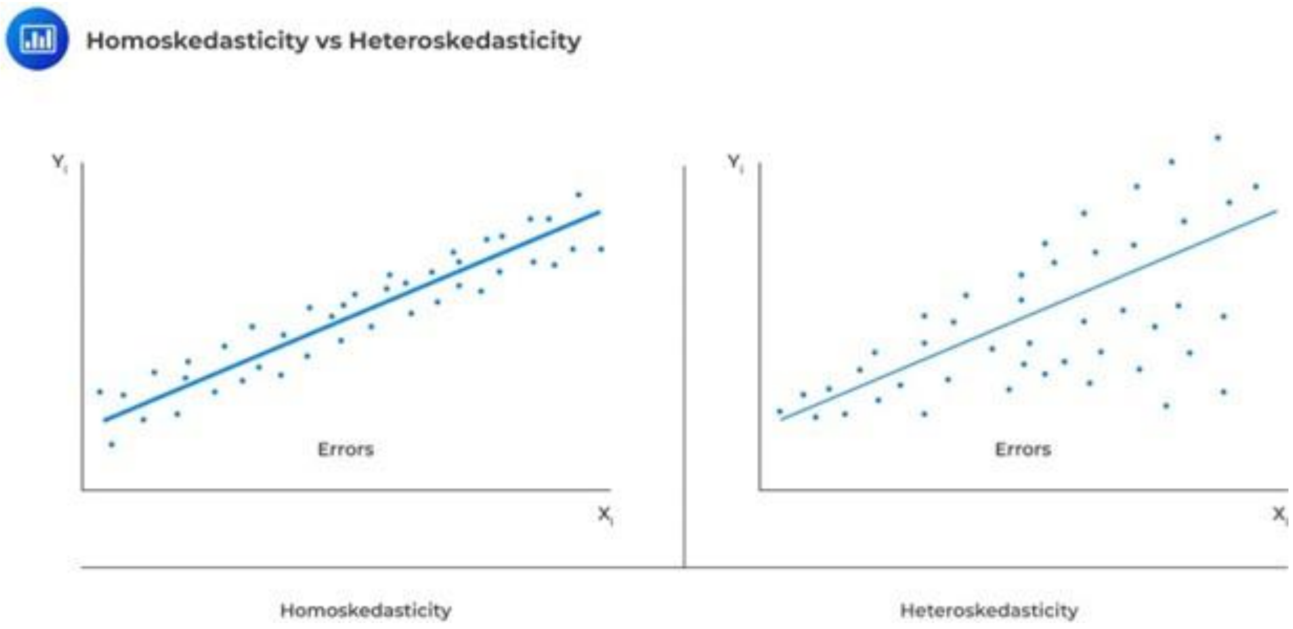
Skewness and Kurtosis

- Skewness and Kurtosis are used to describe the spread and height of your normal distribution.
- Skewness is used to denote the horizontal pull on the data. It tells you how spread out the data is, and Kurtosis is used to find the vertical pull or the peak's height.



Heteroskedasticity

- Heteroscedasticity is mainly due to the presence of outlier in the data. Outlier in Heteroscedasticity means that the observations that are either small or large with respect to the other observations are present in the sample. Heteroscedasticity is also caused due to omission of variables from the model.
- Heteroskedasticity is a statistical concept that refers to the non-constant variance of a dependent variable.
- In other words, it occurs when the variability of a dependent variable is unequal across different values of an independent variable. This can often be seen in financial data, where the volatility of stock prices tends to be greater during periods of economic uncertainty.
- While heteroskedasticity can present challenges for statistical analysis, it can also provide valuable insights into the relationships between variables. For instance, by taking heteroskedasticity into account, economists can better understand how changes in one variable may impact the variability of another.

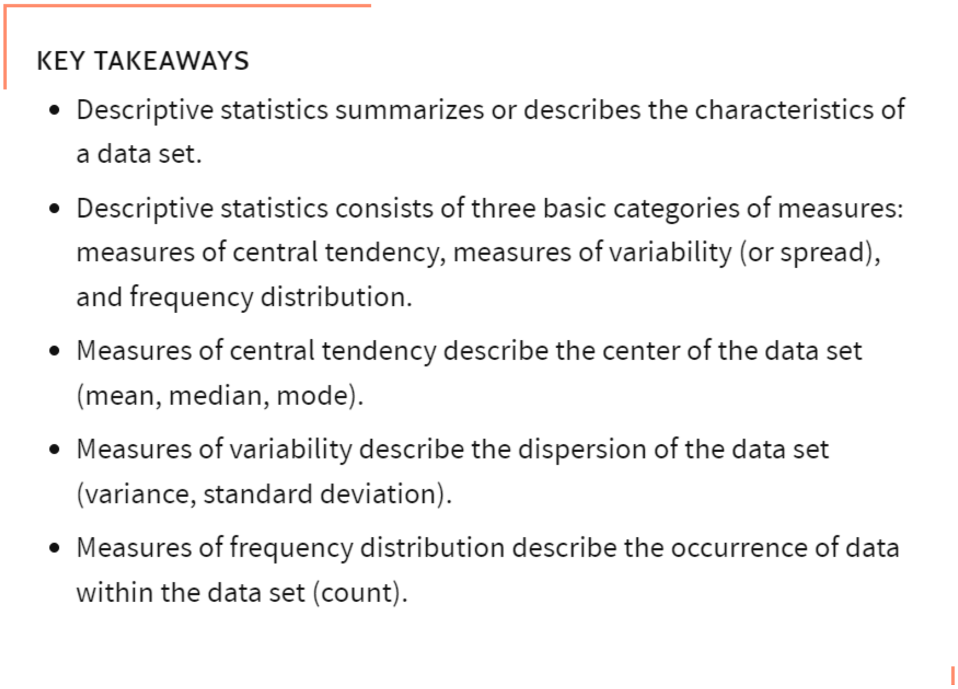


Descriptive Statistics

Descriptive statistics refers to a set of methods used to summarize and describe the main features of a dataset, such as its **central tendency, variability, and distribution**. These methods provide an overview of the data and help identify patterns and relationships.

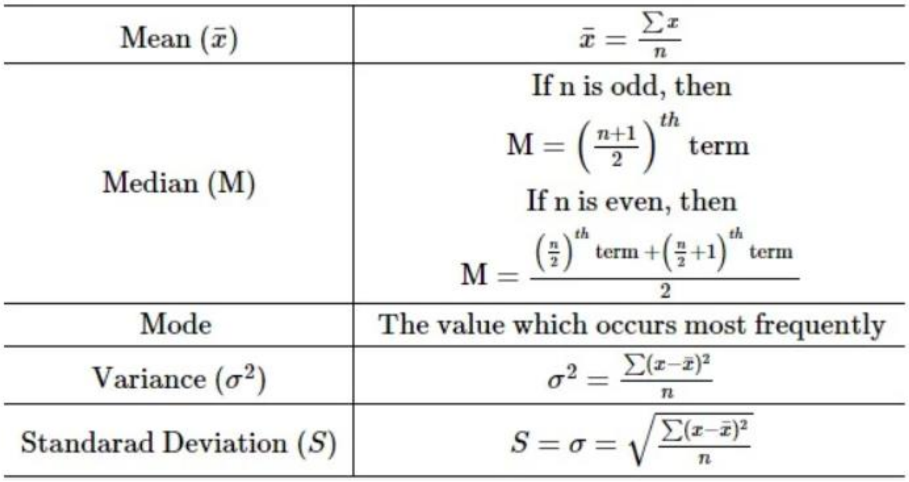
What Are Descriptive Statistics?

- Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population.
- Descriptive statistics are broken down into measures of central tendency and measures of variability (spread), frequency distribution.
- Measures of central tendency include the mean, median, and mode,
- while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and skewness.



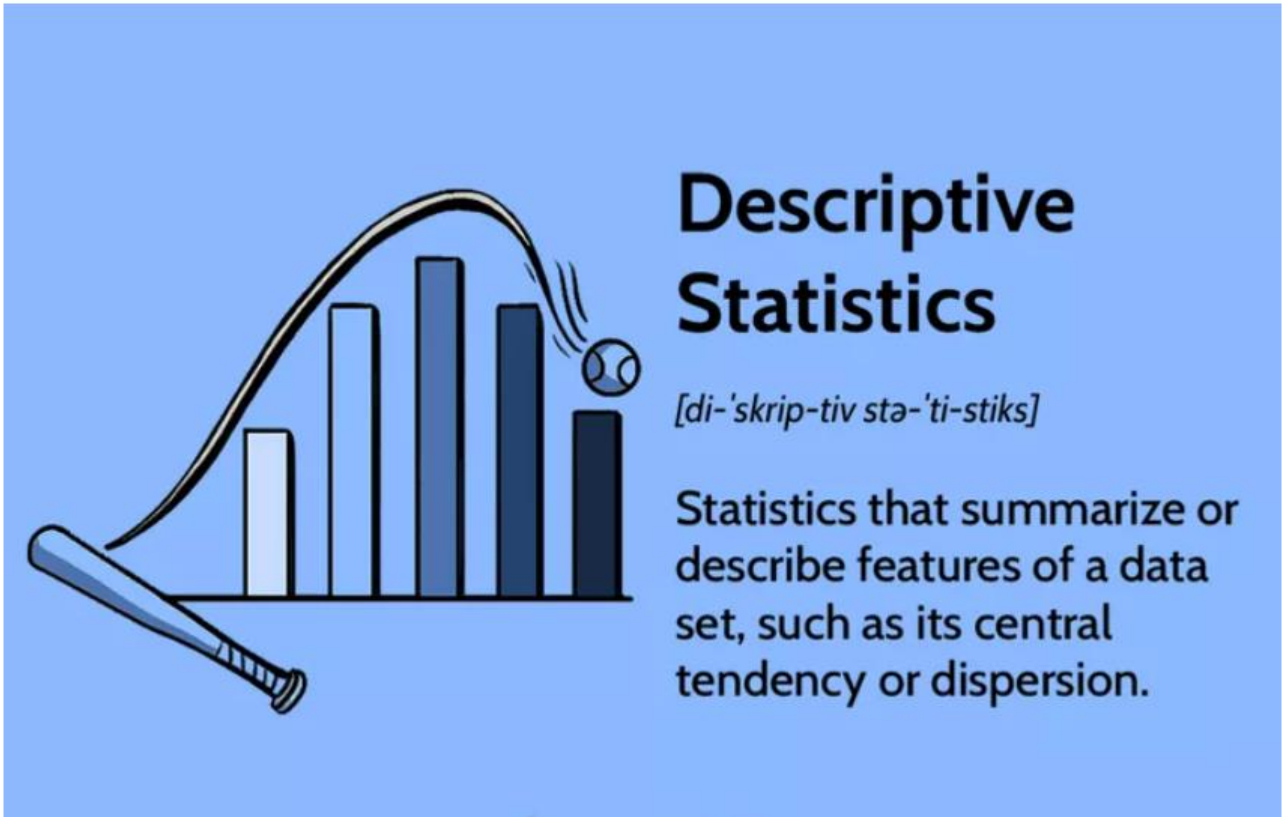
KEY TAKEAWAYS

- Descriptive statistics summarizes or describes the characteristics of a data set.
- Descriptive statistics consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution.
- Measures of central tendency describe the center of the data set (mean, median, mode).
- Measures of variability describe the dispersion of the data set (variance, standard deviation).
- Measures of frequency distribution describe the occurrence of data within the data set (count).



Mean ($\bar{x}$)	$\bar{x} = \frac{\sum x}{n}$
Median (M)	If n is odd, then $\mathrm{M} = \left(\frac{n+1}{2}\right)^{th}$ term If n is even, then $\mathrm{M} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$
Mode	The value which occurs most frequently
Variance (σ^2)	$\sigma^2 = \frac{\sum (x-\bar{x})^2}{n}$
Standarad Deviation (S)	$S = \sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$

where,
x = Observations given
x(bar)= Mean
n = Total number of observations



Descriptive Statistics

[di-'skrip-tiv stə-'ti-stiks]

Statistics that summarize or describe features of a data set, such as its central tendency or dispersion.

Understanding Descriptive Statistics

- Descriptive statistics, in short, help describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.
- The most recognized types of descriptive statistics are measures of center: the mean, median, and mode, which are used at almost all levels of math and statistics. The mean, or the average, is calculated by adding all the figures within the data set and then dividing by the number of figures within the set.

Types of Descriptive Statistics

All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion and frequency distribution.

Central Tendency

- Measures of central tendency focus on the average or middle values of data sets, whereas measures of variability focus on the dispersion of data. These two measures use graphs, tables and general discussions to help people understand the meaning of the analyzed data.
- Measures of central tendency describe the center position of a distribution for a data set. A person analyzes the frequency of each data point in the distribution and describes it using the mean, median, or mode, which measures the most common patterns of the analyzed data set.

Measures of Variability

- Measures of variability (or the measures of spread) aid in analyzing how dispersed the distribution is for a set of data. For example, while the measures of central tendency may give a person the average of a data set, it does not describe how the data is distributed within the set.
- So while the average of the data maybe 65 out of 100, there can still be data points at both 1 and 100. Measures of variability help communicate this by describing the shape and spread of the data set. Range, quartiles, absolute deviation, and variance are all examples of measures of variability.

Consider the following data set: 5, 19, 24, 62, 91, 100. The range of that data set is 95, which is calculated by subtracting the lowest number (5) in the data set from the highest (100).

Distribution

Distribution (or frequency distribution) refers to the quantity of times a data point occurs. Alternatively, it is the measurement of a data point failing to occur. Consider a data set: male, male, female, female, female, other. The distribution of this data can be classified as:

The number of males in the data set is 2.

The number of females in the data set is 3.

The number of individuals identifying as other is 1.

The number of non-males is 4.

Univariate vs. Bivariate

In descriptive statistics, **univariate data analyzes only one variable**. It is used to identify characteristics of a single trait and is not used to analyze any relationships or causations.

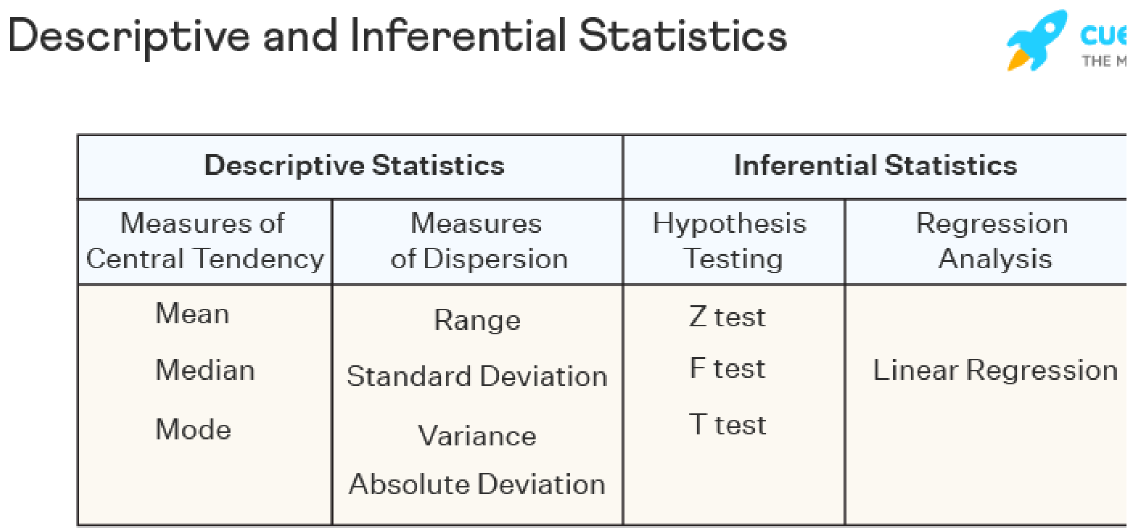
For example, imagine a room full of high school students. Say you wanted to gather the average age of the individuals in the room. This univariate data is only dependent on one factor: each person's age. By gathering this one piece of information from each person and dividing by the total number of people, you can determine the average age.

Bivariate data, on the other hand, attempts to link two variables by searching for correlation. Two types of data are collected, and the relationship between the two pieces of information is analyzed together. Because multiple variables are analyzed, this approach may also be referred to as multivariate.

Let's say each high school student in the example above takes a college assessment test, and we want to see whether older students are testing better than younger students. In addition to gathering the age of the students, we need to gather each student's test score. Then, using data analytics, we mathematically or graphically depict whether there is a relationship between student age and test scores.

The preparation and reporting of financial statements is an example of descriptive statistics Analyzing that financial information to make decisions on the future is inferential statistics.

Descriptive Statistics vs. Inferential Statistics



Descriptive and Inferential Statistics

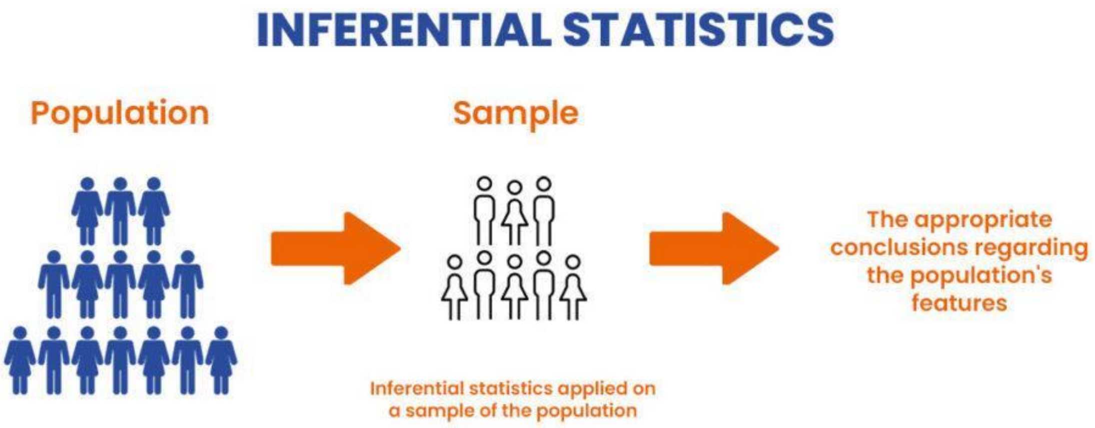
Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean Median Mode	Range Standard Deviation Variance Absolute Deviation	Z test F test T test	Linear Regression

What Is Descriptive Statistics?

Descriptive statistics is a means of describing features of a data set by generating summaries about data samples. It's often depicted as a summary of data shown that explains the contents of data. For example, a population census may include descriptive statistics regarding the ratio of men and women in a specific city.

Inferential statistics

Inferential statistics is a branch of statistics that deals with making inferences about a population based on a sample of data from that population. The goal of inferential statistics is to make generalizations about a larger group of individuals or objects based on the characteristics observed in a smaller sample.


INFERENTIAL STATISTICS
Population
Sample
The appropriate conclusions regarding the population's features
Inferential statistics applied on a sample of the population

What Are Examples of Descriptive Statistics?

Descriptive statistics are informational and meant to describe the actual characteristics of a data set. When analyzing numbers regarding the prior Major League Baseball season, descriptive statistics including the highest batting average for a single player, the number of runs allowed per team, and the average wins per division.

What Is the Main Purpose of Descriptive Statistics?

The main purpose of descriptive statistics is to provide information about a data set. In the example above, there are hundreds of baseballs players that engage in thousands of games. Descriptive statistics summarizes the large amount of data into several useful bits of information.

What Are the Types of Descriptive Statistics?

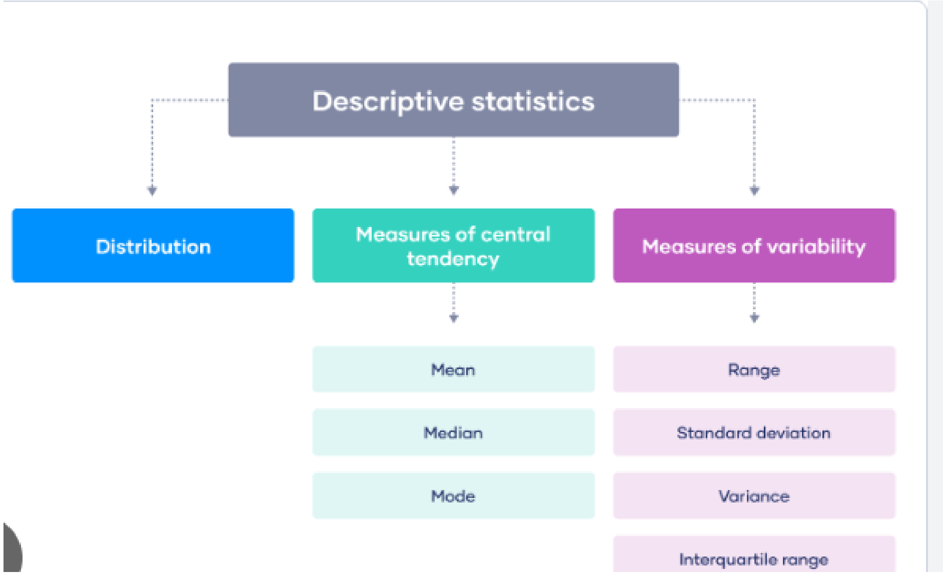
The three main types of descriptive statistics are frequency distribution, central tendency, and variability of a data set. The frequency distribution records how often data occurs, central tendency records the data's center point of distribution, and variability of a data set records its degree of dispersion.

Can Descriptive Statistics Be Used to Make Inference or Predictions?

No. While these descriptives help understand data attributes, inferential statistical techniques—a separate branch of statistics—are required to understand how variables interact with one another in a data set.

Summary

Descriptive statistics refers to the analysis, summary, and communication of findings that describe a data set. Often not useful for decision-making, descriptive statistics still hold value in explaining high-level summaries of a set of information such as the mean, median, mode, variance, range, and count of information.



Higher-Order Moments:

High-order moments are moments beyond 4th-order moments. As with variance, skewness, and kurtosis, these are higher-order statistics, involving non-linear combinations of the data, and can be used for description or estimation of further shape parameters.

What is Moment in Statistics?

In Statistics, Moments are popularly used to describe the characteristic of a distribution. Let's say the random variable of our interest is X then, moments are defined as the X's expected values. For Example, E(X), E(X^2), E(X^3), E(X^4),..., etc.

Matrices

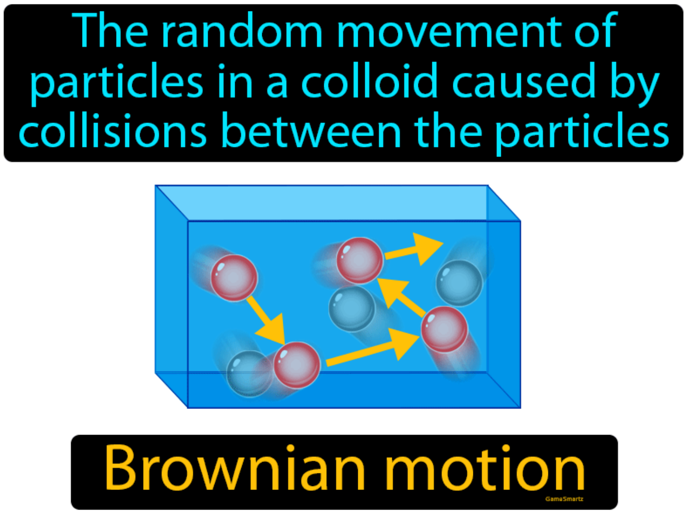
The matrices are a two-dimensional set of numbers or symbols distributed in a rectangular shape in vertical and horizontal lines so that their elements are arranged in rows and columns. They are useful for describing systems of linear or differential equations, as well as representing a linear application.

Maximum-likelihood

In statistics, maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

Introduction to Brownian Motions

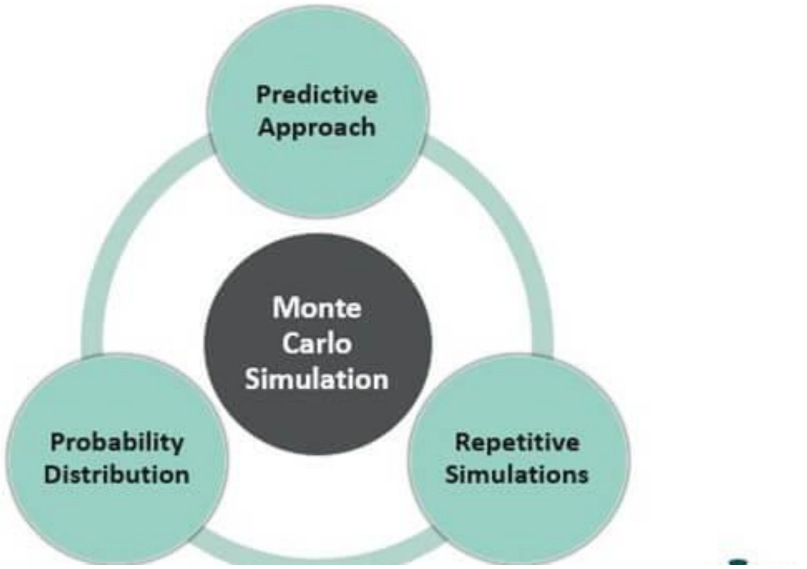
Brownian motion is the random motion of particles suspended in a medium


The random movement of particles in a colloid caused by collisions between the particles
Brownian motion
GameSmartz

Monte Carlo

A Monte Carlo simulation is a mathematical technique that simulates the range of possible outcomes for an uncertain event. These predictions are based on an estimated range of values instead of a fixed set of values and evolve randomly. Computers use Monte Carlo simulations to analyze data and predict a future outcome based on a course of action.

Monte Carlo Simulation Methods



https://www.investopedia.com/terms/d/descriptive_statistics.asp