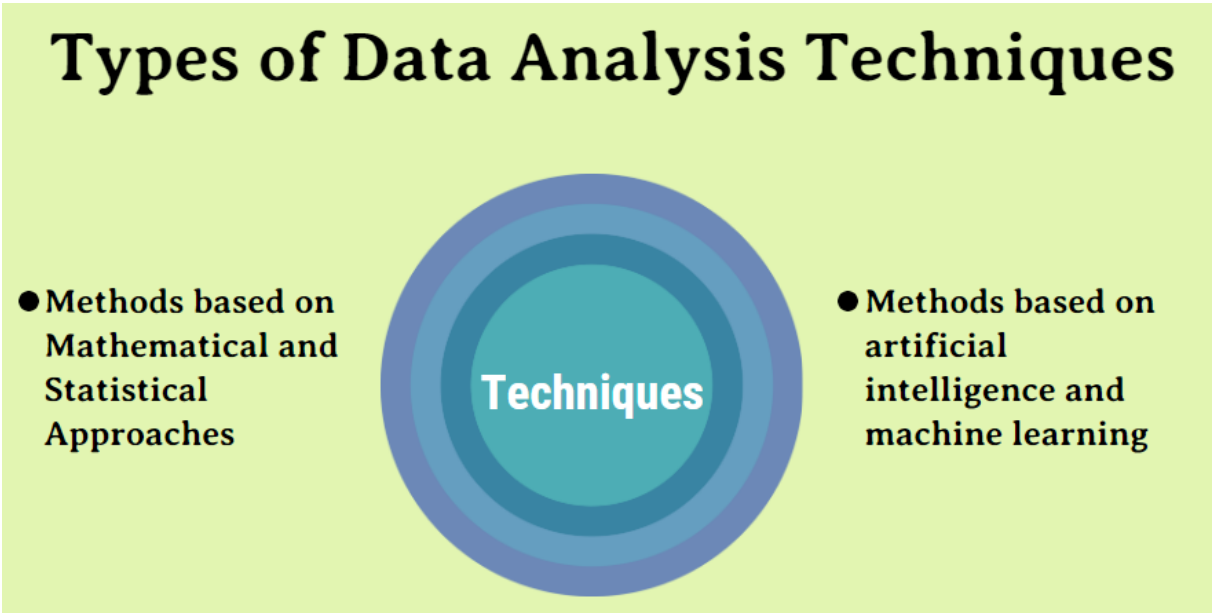


Data analysis method in AI

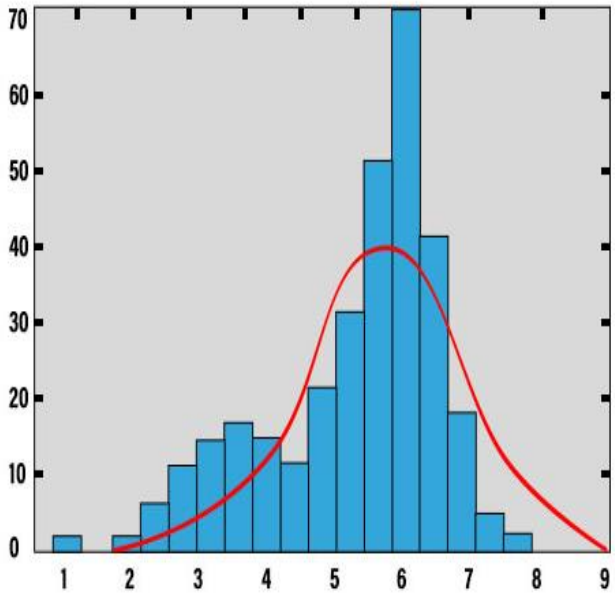
Introduction to Data Analysis Techniques

Data analysis is the process of capturing the useful information by inspecting, cleansing, transforming and modeling data using one of its types that are descriptive analysis, regression analysis, dispersion analysis, factor analysis(independent variable to find the pattern) and time series that are part of the methods based on mathematical and statistical approach or decision trees (tree-like structure for choosing different paths), neural network (set of algorithms), fuzzy logic(a decision that is not true or false) that are part of the methods based on machine learning and artificial intelligence.

Important Types of Data Analysis Techniques

Data analysis techniques are broadly classified into two types they are

- Methods based on Mathematical and Statistical Approaches
- Methods based on artificial intelligence and machine learning



Mathematical and Statistical Approaches

1. Descriptive Analysis

Descriptive analysis is an important first step for conducting statistical analysis. It provides us with an idea of the distribution of data, helps detect outliers, and enables us to identify associations among variables, thus preparing the data for conducting further statistical analysis. Descriptive analysis of a huge data set can be made easy by breaking down it into two categories, they are descriptive analysis for each individual variable and descriptive analysis for combinations of variables.

2. Regression Analysis

Regression analysis is one of the dominant data analysis techniques that is being used in the industry right now. In this kind of technique, we can see the relationship between two or more variables of interest and at the core, they all study the influence of one or more independent variables on the dependent variable. To see if there is any relationship between the variables or not we first need to plot the data on a chart and it will be evident if there is any relation. For example, consider the graph plotted below to have a clear understanding.

In data mining, this technique is used to predict the values of a variable, in that particular dataset. There are different types of regression models in usage. A few of them are Linear regression, logistic regression, and multiple regression.

3. Dispersion Analysis

Dispersion is the extent to which a distribution is stretched or squeezed. In the mathematical approach, the dispersion can be defined in two ways, fundamentally the difference of values among themselves and secondly the difference between the average value. If the difference between the value and average is very low, then we can say that dispersion is less in this case. And some of the common measures of dispersion are variance, standard deviation, and interquartile range.

4. Factor Analysis

Factor analysis is a kind of data analysis technique, which helps in finding the underlying structure in a set of variables. It helps with finding independent variables in the data set that describes the patterns and models of relationships. It is the first step towards clustering and classification procedures. Factor analysis is also related to Principal Component Analysis(PCA) but both of them are not identical we can call PCA as the more basic version of exploratory factor analysis

5. Time Series

Time series analysis is a data analysis technique, that deals with the time-series data or trend analysis. Now, let us understand what is time-series data? Time series data is data in a series of particular time intervals or periods. If we see scientifically, most of the measurements are executed over time.

Methods based on Machine Learning and Artificial Intelligence

1. Decision Trees

Decision tree analysis is a graphical representation, similar to a tree-like structure in which the problems in decision making can be seen in the form of a flow chart, each with branches for alternative answers. Decision trees are a top-down approach type, with the first decision node at the top, based on the answer at first decision node it will be divided into branches, and it will continue until the tree arrives at a final decision. The branches which do not divide any more are known as leaves.

2. Neural Networks

Neural networks are a set of algorithms, which are designed to mimic the human brain. It is also known as the “Network of Artificial neurons”. The applications of neural network in data mining are very broad. They have a high acceptance ability for noisy data and high accuracy results. Based on the necessity many types of neural networks are currently being used, few of them are recurrent neural networks and convolutional neural networks. Convolutional neural networks are mostly used in Image processing, natural language processing, and recommender systems. Recurrent neural networks are mainly used for handwriting and speech recognition.

3. Evolutionary Algorithms

Evolutionary algorithms use the mechanisms inspired by recombination and selection. These types of algorithms are independent of the domain and they have the ability to explore large data sets, discovering patterns and solutions. They are insensitive to noise compared with other data techniques.

4. Fuzzy logic

It is an approach in computing based on “Degree of truth” rather than the common “Boolean logic” (truth/false or 0/1). As discussed above in decision trees at decision node we either have yes or no as an answer, what if we have a situation where we can’t decide absolute yes or absolute no? In these cases, fuzzy logic plays an important role. It is a diverse valued logic in which the truth value can be between completely true and completely false, that is it can take any real value between 0 and 1. Fuzzy logic is applicable when there is a significant amount of noise in the values.

Conclusion

The tough question that all corporates or companies face is which type of data analysis technique is the best for them? We cannot define any technique as the best instead what we can do is try multiple techniques and see which one best fits our data set and use it. The above-mentioned techniques are some of the important techniques that are currently being used in the industry.

What is Statistical Analysis?

Statistical Analysis is the scientific way to collect, preprocess and apply a set of statistical methods to discover the insights or underlying pattern of the data. With the increase in cheap data and incremental bandwidth, we are now sitting on a ton of structured and unstructured data. Along with the need for acquiring and maintaining this huge data, one main challenge is to deal with the noise and convert the data into a meaningful way. The statistical analysis comes up with a set of statistical methodologies and tools to address the problem.

How Statistical Analysis is Performed?

Statistical analysis is a vast literature of data analysis itself. Let us discuss the most common approaches of statistical data analysis:

Searching for Central Tendency

While working with structural data it is often the preliminary step to get an idea on the central tendency of the data set. Suppose you are analyzing the salary data of an organization. Then you may be interested in the following questions like what is the average salary of a manager working in the organization for 3 years with so and so qualification? The following are used as a measurement of central tendency.

Mean: Mean is basically the average of all the data points. Mean is the total salary divided by the number of data points.

$$md = x_{\frac{(n-1)}{2}} \quad \text{for n is odd}$$

$$md = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \quad \text{for n is even}$$

Median: Median is the 50th percentile of the data. When we are seeking information like average salary, the median will be a more robust measure. It is less sensitive to outliers.

Mode: Mode is the most frequent value in the list of numbers. Suppose we are dealing with a list of numbers [12, 33, 44, 55, 67, 55, 8, 55], here the mode with be 55.

Searching for Dispersion

Dispersion is the measurement of variability in the data. Dispersion helps us to find out how a data point is different from its central tendency. Finding the proper

$$SD = \sqrt{\frac{\sum \left(r_i - r_{avg}\right)^2}{n-1}}$$

distribution is important to decide which machine learning algorithm to use based on the use case.

Standard Deviation: Standard Deviation quantifies how much the data point varies from its central tendency (dispersion). The lower the value, the more the data points are identical with its central value.

Variance: Variance is the square of standard deviation. The variance gives us the spread (variability) of the data. While working with high dimensional data we often come up with a situation where we need to reduce the dimensionality or analyze the important variables of the data set. In such situations, we convert the axis in such a way that maximum variability is preserved. This new rotating axis is called the principal components. We choose N important components (an axis with high variance) from the rotating components.

Interquartile Range (IQR): Interquartile range is the range of data between the 25th and 75th percentile values of the data set. We use box plot, violin plot, etc. to analyze the IQR in graphical ways.

Regression Problems

Regression is a set of problems where the independent variable is a continuous variable. For example, we have the historical sales data of car manufactures and various factors that affect the car manufacturing and sales process and we need to predict the sales of a particular brand. Now we will formulate the regression problem as ‘find the sales of a car brand ABC based on the factors x1, x2, x3, etc.’

Advantages of Using Statistical Analysis

Below are the points that explain the advantages of using Statistical Analysis:

- In the era of Big Data, while implementing any machine learning use case it is the utmost importance of how we choose the sample from the huge data lake. Statistical analysis helps us to determine the proper sampling methodology (i.e random, random without substitution, stratified sampling, etc) and reduce the sampling bias.
- For example, we are dealing with binary classification problem where 80% of data points belong to the class A and only 20% belong to class B. Now if we want to perform any statistical test with samples from the population, we must ensure the samples are also in 80:20 ratio (80% class A: 20% class B).
- Be it sampling or decision making the basis of statistical analysis is historical data. This makes statistical data analysis more acceptable as an industry-standard than another manual process of data analysis.

Why Do We Need Statistical Analysis?

The main goal of statistical analysis is to find valuable insights from the data which may be used to discover Industry trends, customer rate of attrition to a product or service, making a valuable business decision, etc.

From the collection of data to find the underlying patterns of the data, statistical analysis is the base of all data-driven methodologies and classical machine learning.

Scope of Statistical Analysis

The following are the points that explain the scope of Statistical Analysis:

- In today's world, more and more Industries are switching to data-based decision-making systems instead of classical deterministic rule-based approaches.
- Statistical analysis is being used dominantly to solve various business problems across domains like Manufacturing, Insurance, Banking and Finances, Automobile, etc. from the industry point of view.

- From a technical perspective statistical analysis helps to solve linear regress, time series forecasting, predictive analysis, etc.

Conclusion

In this article, we have discussed the various aspects of statistical data analysis like methodologies, the need, and scope of use cases, etc. Statistical analysis is a very old area of study which lays out the base for modern machine learning and data-driven business models. The practical implementation of statistical analysis methodologies differs based on the type of use case and industry.