The term **univariate analysis**refers to the analysis of one variable. You can remember this because the prefix “uni” means “one.”

The purpose of univariate analysis is to understand the distribution of values for a single variable.

**Importance of univariate analysis**

Univariate analysis serves as an important first step in many research projects, as it provides essential insights and lays a foundation for further research. It offers researchers the following benefits:

**Data exploration**

Univariate analysis allows researchers to understand the distribution, central tendency, and variability of a variable.

**Identification of outliers**

By detecting anomalous values, univariate analysis helps identify outliers that require further investigation or treatment during the [**data analysis**](https://www.toolshero.com/information-technology/data-analytics/) phase.

**Data cleaning**

Univariate analysis helps identify missing data, inconsistencies or errors within a variable, allowing researchers to refine and optimize their data set before moving on to more complex analyses.

**Variable selection**

Researchers can use the univariate analysis to determine which variables are most promising for further research. This enables efficient allocation of resources and hypothesis testing.

**Reporting and visualization**

Summarizing and visualizing univariate statistics facilitates clear and concise reporting of research results. This makes complex data more accessible to a wider audience.

**Applications of univariate analysis**

Univariate analysis is used in various research areas and disciplines. It is often used in:

* Epidemiological studies to analyze risk factors
* Social science research to investigate attitudes, behaviors or socio-economic variables
* Market research to understand consumer preferences, buying patterns or market trends
* Environmental studies to investigate pollution, climate data or species distributions

By using univariate analysis, researchers can uncover valuable insights, detect trends, and lay the groundwork for more comprehensive statistical analysis.

**Example situation of an Univariate Analysis**

An example of univariate analysis might be examining the age of employees in a company.

Data is collected on the age of all employees and then a univariate analysis is performed to understand the characteristics and distribution of this single variable.

We can calculate summary statistics, such as the mean, median, and standard deviation, to get an idea of the central tendency and range of ages.

Histograms can also be used to visualize the frequency of different age groups and to identify any patterns or outliers.

This univariate analysis helps organizations to better understand the age structure of the workforce in the company.

Univariate data classifications are as follows:

* ID: This data has no statistical or aggregative properties, and they are used to identify a subject uniquely. For example, the column ‘**Employee Number**’.
* Numerical (Quantitative): This data has statistical properties. They can be of two types- Discrete and Continuous.
  + *Discrete*: This dataset has discrete values (i.e., cannot have decimals). For example- ‘No of Family Members.
  + *Continuous*: This dataset can have numbers with decimals. For example- ‘Income’.
* Categorical (Qualitative): Categorical data deals with descriptions or categories. They have aggregative properties and are of two types- Ordinal and Nominal. (Note- categorical univariate data can have numeric datatype)
  + *Ordinal:* These categories have an order. For example- ‘Designation’ where the order can be Manager, Sr Manager, CEO and cannot be any other.
  + *Nominal*: These categories do not have any order. For example- ‘Location’ has mutually exclusive categories.

Typically, a univariate is data that belongs to any of the types mentioned above. Now, to analyze such a dataset, different types of univariate analysis techniques are used depending on the type of variable in question.

The primary purpose of univariate analysis is to describe data. Using different techniques, these descriptions are found. These techniques can be categorized into the following groups:

1. Graphical
2. Tables
3. Descriptive statistics
4. Inferential statistics (i.e., use of frequency distributions)

Each of these techniques provides information about the data in a unique way. Typically, a data analyst uses more than one technique to form their opinion about the data they are dealing with, as this helps them make important decisions related to data preparation. Let’s understand each technique.

**Graphical analysis**

Various types of graphs can be used to understand data. The standard type of graphs include-

1. **Histograms:**A histogram displays the frequency of each value or group of values (bins) in numerical data. This helps in understanding how the values are distributed.
2. **Boxplot:**A boxplot provides several important information such as minimum, maximum, median, 1st, and 3rd quartiles. It is beneficial in identifying outliers in the data.
3. **Density Curve:**The density curve helps in understanding the shape of the data’s distribution. It helps answer questions such as if the data is bimodal, normally distributed, skewed, etc.
4. **Bar Chart:**Bar Charts, mainly frequency bar charts, is a univariate chart used to find the frequency of the different categories of categorical data.
5. **Pie Chart:**Frequency Pie charts convey similar information to bar charts. The difference is that they have a circular formation with each slice indicating the share of each category in the data.

**Univariate tables**

Tables help in univariate analysis and are typically used with categorical data or numerical data with limited cardinality. Different types of tables include:

1. **Frequency Tables**: Each unique value and its respective frequency in the data is shown through a table. Thus, it summarizes the frequency the way a histogram, frequency bar, or pie chart does but in a tabular manner.
2. **Grouped Tables**: Rather than finding the count of each unique value, the values are binned or grouped, and the frequency of each group is reflected in the table. It is typically used for numerical data with high cardinality.
3. **Percentage (Proportion) Tables**: Rather than showing the frequency of the unique values (or groups), such a table shows their proportion in the data (in percentage).
4. **Cumulative Proportion Tables**: It is similar to the proportion table, with the difference being that the proportion is shown cumulatively. It is typically used with binned data having a distinct order (or with categorical ordinal data).

*In some instances, all such univariate tables can be used as an alternative to a more graph-based way of describing the analysis.*

**Descriptive Statistics**

As the name suggests, descriptive statistics are used to describe data. The statistics used here are commonly referred to as *summary statistics.*

For instance, if you have to describe a cube, you have to ‘measure’ it. By measuring its length, breadth, and height, you can describe it. Similarly, these descriptive or univariate statistics have specific measures that help us in describing the data. These measures are-

1. **Measure of Central Tendency**: Statistics such as mean, median, and mode are considered here. They help in summarizing all the data through a single central value.
2. **Measure of Variability**: Analysts also need to understand how the data varies from the central point. To understand this, specific univariate statistics can be calculated, such as range, interquartile range, variance, standard deviation, etc.
3. **Measure of Shape**: The shape of the data distribution can explain a great deal about the data as the shape can help in identifying the type of distribution followed by the data. Each of these distributions has specific properties that can be used to your advantage. By analyzing the shapes, you will know if the data is symmetrical, non-symmetrical, left or right-skewed, is suffering from positive or negative kurtosis, among other things.

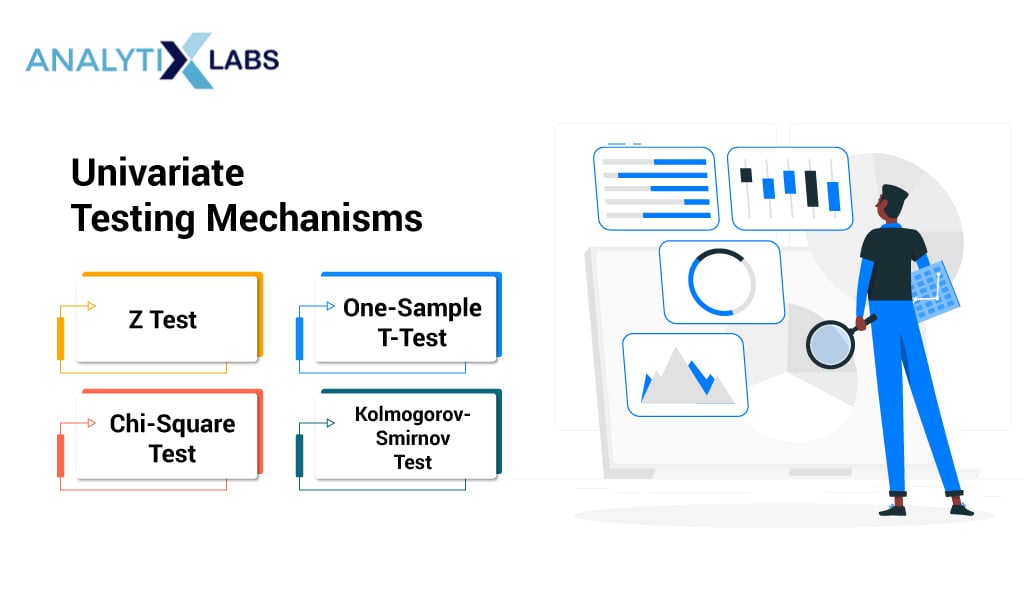
These descriptive statistics can be used for calculating things like missing value proportions, upper and lower limits for outliers, level of variance through the coefficient of variance, etc.

**Inferential Statistics**

Often, the data you are dealing with is a subset (sample) of the complete data (population). Thus, the common question here is –

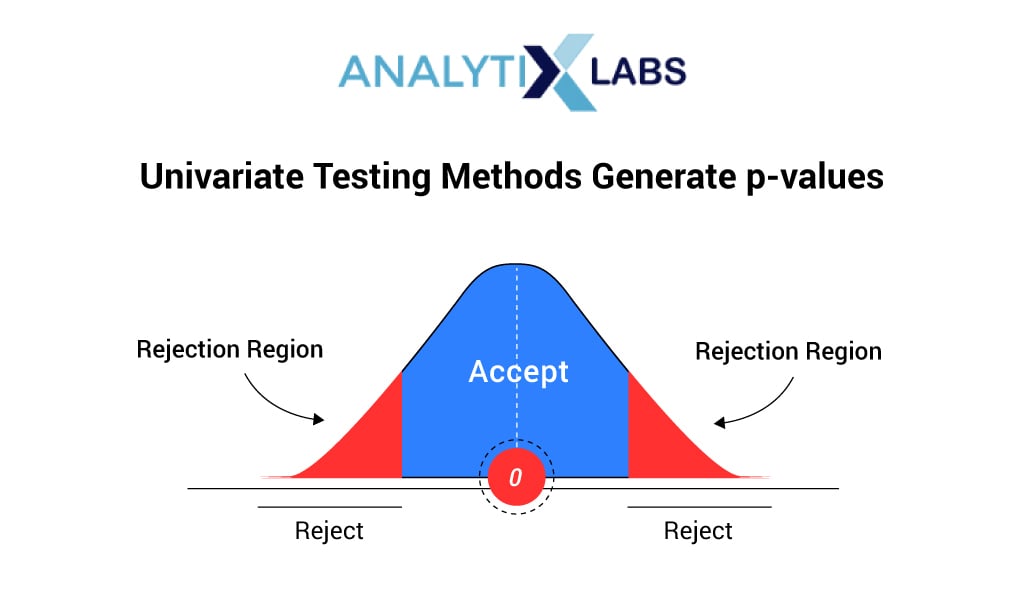
Can the findings of the sample be extrapolated to the population? i.e., Is the sample representative of the population, or has the population changed? Such questions are answered using specific hypothesis tests designed to deal with such univariate data-based problems.

***Hypothesis tests*** help us answer crucial questions about the data and their relation with the population from where they are drawn. Several hypotheses or univariate testing mechanisms come in handy here, such as-



1. **Z Test**: Used for numerical (quantitative) data where the sample size is greater than 30 and the population’s standard deviation is known.
2. **One-Sample t-Test**: Used for numerical (quantitative) data where the sample size is less than 30 or the population’s standard deviation is unknown.
3. **Chi-Square Test**: Used with ordinal categorical data
4. **Kolmogorov-Smirnov Test**: Used with nominal categorical data

All such univariate testing methods generate p-values that can be used to accept or reject different types of hypotheses.



In real-time, all these techniques are used depending upon the situation, type of data, and problem statement.

**Univariate Analysis Examples**

While there can be hundreds of univariate analysis examples where univariate analysis is used, some of them are-

1. Finding the average height of a country’s men from a sample.
2. Calculate how reliable a batsman is by calculating the variance of their runs.
3. Finding which country is the most frequent in winning Olympic Gold Medal by creating a frequency bar chart or frequency table.
4. Understanding the income distribution of a county by analyzing the distribution’s shape. A right-skewed distribution can indicate an unequal society.
5. Checking if the price of sugar has statistically significantly risen from the generally accepted price by using sample survey data. Hypothesis tests such as the Z or t-test solve such questions.
6. Assessing the predictive capability of a variable by calculating the coefficient of variance.

[A Deeper Dive into Univariate Analysis | by Dorjey Sherpa | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/intro-to-univariate-analysis-de75454b4719)

BIVARIATE ANALYSIS

* **Bivariate Analysis**: Bivariate analysis is performed when two variables are involved. Here, you typically try to understand how two variables affect each other, how they are related, or how they compare to each other. Like univariate analysis that is performed through graphs, tables, and statistics, bivariate analysis can also be performed somewhat similarly. For example- scatterplots, bar charts, pie charts, multi-line charts, cross-frequency tables, and tests such as dependent t-test, independent t-test, and one-way ANOVA are used for bivariate analysis.

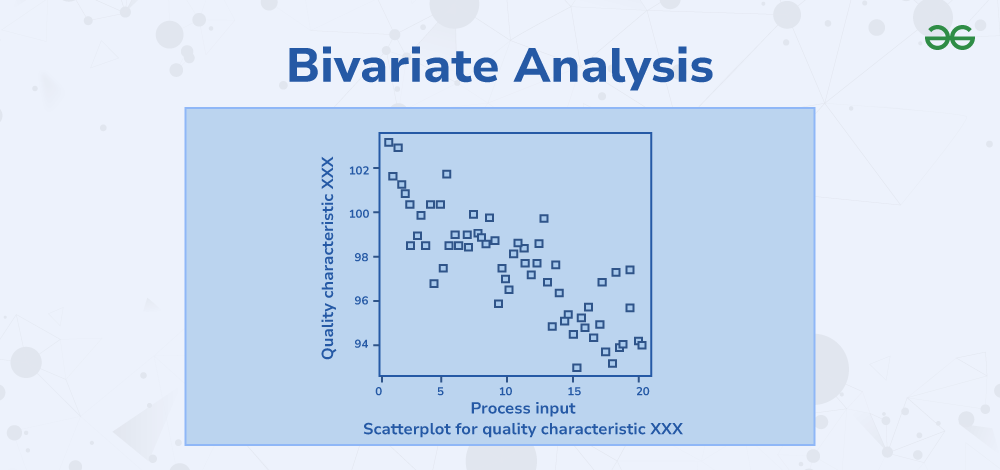
Bivariate data involves two different variables, and the analysis of this type of data focuses on understanding the relationship or association between these two variables. **Example**of bivariate data can be temperature and ice cream sales in summer season.

suppose the temperature and ice cream sales are the two variables of a bivariate data(table 2). Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.

**Key points in Bivariate analysis:**

1. **Relationship Analysis:** The primary goal of analyzing bivariate data is to understand the relationship between the two variables. This relationship could be positive (both variables increase together), negative (one variable increases while the other decreases), or show no clear pattern.
2. **Scatterplots:** A common visualization tool for bivariate data is a scatterplot, where each data point represents a pair of values for the two variables. Scatterplots help visualize patterns and trends in the data.
3. **Correlation Coefficient:** A quantitative measure called the correlation coefficient is often used to quantify the strength and direction of the linear relationship between two variables. The correlation coefficient ranges from -1 to 1.

**Bivariate analysis**examines the relationship between two variables. It is often denoted as X and Y. It helps uncover correlations and associations between different factors in data analysis.



*Bivariate Analysis*

In this article, we will understand the meaning of bivariate analysis and its definition, as well as the types of bivariate analysis and applications of bivariate analysis.

**Bivariate Analysis**

Bivariate analysis is a statistical method used to investigate the relationship between two variables. It is often used in quality-of-life research. It’s a straightforward form of [quantitative analysis](https://www.geeksforgeeks.org/what-is-quantitative-analysis/) which examines two variables denoted as X and Y. For instance, consider a study examining the relationship between exercise duration (X) and heart rate (Y) during physical activity. By analyzing this bivariate data, researchers can determine if there’s a [correlation](https://www.geeksforgeeks.org/mathematics-covariance-and-correlation/) between the duration of exercise and heart rate.

In bivariate analysis, researchers aim to understand to what extent one variable (possibly independent, like exercise duration) predicts the value of another (possibly dependent, like heart rate). It’s a basic yet valuable tool for testing hypotheses of association. Bivariate analysis contrasts with univariate analysis, which examines only one variable at a time. Results from bivariate analysis are typically organized in a two-column data table, making it easy to visualize the relationship between the variables.

In a study analyzing the relationship between study hours (X) and exam scores (Y), bivariate data can be collected from a sample of students. Each student’s study hours and corresponding exam score form a pair of variables denoted as (Xi, Yi), where i represents the individual student. The data can be represented in a table format as follows:

| **Observations** | **Study Hours** | **Exam Scores** |
| --- | --- | --- |
| 1 | 10 | 85 |
| 2 | 5 | 70 |
| 3 | 6 | 75 |
| 4 | 8 | 80 |
| 5 | 4 | 65 |

This bivariate data allows researchers to analyze the relationship between study hours and exam scores to understand if there’s a correlation between the two variables.

**Definition of Bivariate Analysis**

Bivariate analysis is when we look at two things together to see how they’re related. For example, if we want to know if how much someone studies (one thing) affects their test scores (other thing), we would use bivariate analysis.

**Types of Bivariate Analysis**

The various types of bivariate analysis are:

* [Scatter Plots](https://www.geeksforgeeks.org/scatter-plot-using-plotly-in-python/)
* [Correlation Analysis](https://www.geeksforgeeks.org/what-is-correlation-analysis/)
* [Regression Analysis](https://www.geeksforgeeks.org/what-is-correlation-analysis/)
* [Chi-Square Test](https://www.geeksforgeeks.org/chi-square-test/)
* [T-tests](https://www.geeksforgeeks.org/t-test-in-statistics/) and [ANOVA](https://www.geeksforgeeks.org/anova-formula/)

**Scatter Plots**

Scatter plots visually display the relationship between two variables. Each dot on the plot represents a single observation, with one variable plotted on the x-axis and the other on the y-axis. The pattern formed by the dots can reveal the nature of the relationship between the variables—whether it’s positive, negative, or no correlation.

**Correlation Analysis**

Correlation analysis quantifies the strength and direction of the relationship between two continuous variables. The correlation coefficient, typically denoted by “r,” ranges from -1 to 1. A positive value indicates a positive correlation (as one variable increases, the other tends to increase), while a negative value suggests a negative correlation (as one variable increases, the other tends to decrease). A value close to zero indicates little to no correlation.

**Regression Analysis**

Regression analysis explores the relationship between two or more variables, typically by predicting one variable (the dependent variable) based on the values of one or more other variables (the independent variables). Simple linear regression involves predicting a dependent variable from a single independent variable, while multiple linear regression involves predicting the dependent variable from multiple independent variables.

**Chi-Square Test**

The chi-square test examines the association between two categorical variables by comparing the observed frequencies in a contingency table to the frequencies that would be expected if the variables were independent. It determines whether the observed association between the variables is statistically significant or due to random chance.

**T-tests and ANOVA**

T-tests and analysis of variance (ANOVA) are used to compare means between groups for one or more independent variables. In bivariate analysis, they can be applied to examine whether there are significant differences in the mean values of a continuous variable across different categories of another variable. T-tests are suitable for comparing means between two groups, while ANOVA is used for comparing means among three or more groups.

**Advantages and Disadvantages of Bivariate Analysis**

Bivariate analysis offers several advantages and disadvantages, depending on the context and the specific goals of the analysis.

**Advantages:**

* **Simplicity**: Bivariate analysis is straightforward and easy to understand, making it accessible to a wide range of users, including those with limited statistical expertise.
* **Identification of Relationships**: Bivariate analysis helps identify relationships between two variables, allowing researchers to explore associations, correlations, or dependencies between them.
* **Visualization**: Techniques like scatterplots and contingency tables provide visual representations of the relationship between variables, making it easier to interpret the data and identify patterns.
* **Foundation for Further Analysis**: Bivariate analysis serves as a foundation for more advanced multivariate analyses. Understanding the relationship between two variables can inform subsequent analyses involving multiple variables.
* **Hypothesis Testing**: Bivariate analysis allows researchers to test hypotheses about the relationship between variables, such as whether there is a significant correlation or association.
* **Practicality**: In many situations, particularly in exploratory data analysis or initial stages of research, focusing on the relationship between two variables is practical and sufficient for drawing meaningful insights.

**Disadvantages:**

* **Limited Scope**: Bivariate analysis examines the relationship between only two variables, which may oversimplify complex phenomena influenced by multiple factors. It may not capture the full complexity of real-world relationships.
* **Confounding Variables**: Bivariate analysis does not account for the influence of confounding variables—factors that may affect both variables being studied—potentially leading to biased or misleading conclusions.
* **Causality**: While bivariate analysis can identify associations between variables, it cannot establish causality. Correlation does not imply causation, and additional research is needed to determine causal relationships.
* **Missed Patterns**: Focusing solely on the relationship between two variables may overlook important patterns or trends that emerge when considering additional variables (multivariate analysis).
* **Assumptions**: Some bivariate analysis techniques, such as correlation and regression, rely on specific assumptions (e.g., linearity, normality) that may not hold true in all cases. Violating these assumptions can lead to inaccurate results.
* **Data Limitations**: Bivariate analysis may be limited by the availability or quality of data. If the dataset is small, incomplete, or biased, the results of the analysis may not be representative or reliable.
* **Interpretation Challenges**: While bivariate analysis provides insights into the relationship between variables, interpreting the results requires careful consideration of context, potential biases, and alternative explanations.

**Applications of Bivariate Analysis**

Bivariate analysis finds applications in various fields, including:

* It helps researchers understand relationships between variables like income and education level, crime rates and unemployment, or happiness and marital status.
* Bivariate analysis is used to study the relationship between factors like supply and demand, interest rates and inflation, or GDP and unemployment.
* It helps in analyzing the correlation between factors such as diet and health outcomes, exercise and disease risk, or medication adherence and treatment effectiveness.
* Bivariate analysis assists marketers in understanding relationships between variables like advertising expenditure and sales revenue, customer demographics and purchasing behavior, or product features and consumer satisfaction.
* It helps in studying correlations between factors such as pollution levels and respiratory illnesses, climate variables and agricultural productivity, or habitat loss and species diversity.
* Bivariate analysis is used to explore relationships between factors like study habits and academic performance, class size and student engagement, or teacher qualifications and student achievement.
* It helps in analyzing relationships between variables like stock prices and company earnings, interest rates and bond yields, or asset allocation and investment returns.
* Bivariate analysis helps psychologists understand correlations between factors such as stress levels and mental health, personality traits and behavior patterns, or therapy outcomes and treatment adherence.

**Difference between Univariate, Bivariate and Multivariate Analysis**

The basic difference between univariate, bivariate, and multivariate analysis is explained in the table added below:

| **Univariate Analysis** | **Bivariate Analysis** | **Multivariate Analysis** |
| --- | --- | --- |
| Univariate analysis involves the analysis of a single variable. This helps to describe and summarize the characteristics and distribution of that variable. | Bivariate analysis involves the analysis of the relationship between two variables. This helps to examines how one variable behaves in relation to another. | Multivariate analysis involves the simultaneous analysis of three or more variables. This helps to examines the complex relationships between multiple variables. |
| Techniques used in univariate analysis include:   * Measures of Central Tendency (Mean, Median, Mode) * Measures of Dispersion (Range, Variance, Standard Deviation) * Frequency Distributions, etc. | Techniques used in bivariate analysis include   * Correlation Analysis * Scatter Plots * Cross-Tabulations (Contingency Tables) * Bivariate Regression Analysis, etc. | Common techniques used in multivariate analysis include:   * Multiple Regression Analysis * Factor Analysis * Principal Component Analysis (PCA) * Cluster Analysis, etc. |
| Univariate analysis is useful for understanding the basic properties of a variable and identifying any outliers or patterns within it. | Bivariate analysis helps in understanding the strength and direction of the relationship between two variables. It is particularly useful for identifying associations or dependencies between variables. | Multivariate analysis allows for the exploration of interactions and dependencies among multiple variables. It helps in understanding the combined effect of several variables on an outcome or in identifying latent structures within the data. |

**Examples of Bivariate Analysis**

**Example 1: A teacher collect data on total hours studied by students and total marks scored by them is shown in the table below:**

| **Total Hours Studied in a Week** | **Marks scored out of 700** |
| --- | --- |
| 14 | 450 |
| 12 | 413 |
| 20 | 490 |
| 22 | 566 |
| 24 | 576 |
| 29 | 640 |
| 13 | 340 |

**Solution:**

*The teacher might decide to determine the correlation between the two variables, which she or he may find to be 0.926.*

*This shows that the two variables have a very strong positive association.*

*In other words, the marks scored by a student in a test is directly correlated to number of hours stuied by him/her.*

[Bivariate Analysis - GeeksforGeeks](https://www.geeksforgeeks.org/bivariate-analysis/)

[Bivariate Analysis: What is it, Types + Examples | QuestionPro](https://www.questionpro.com/blog/bivariate-analysis/)