▾ What is Univariate

While 'uni' means one, variate indicates a variable. Therefore, univariate analysis is a form of analysis that only involves a single variable. In a practical setting, a univariate analysis means the analysis of a single variable (or column) in a dataset (data table).

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)

Graphical analysis

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

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.

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.

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.

Bar Chart: Bar Charts, mainly frequency bar charts, is a univariate chart used to find the frequency of the different categories of categorical data.

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:

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.

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.

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).

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).

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-

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.

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.

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.
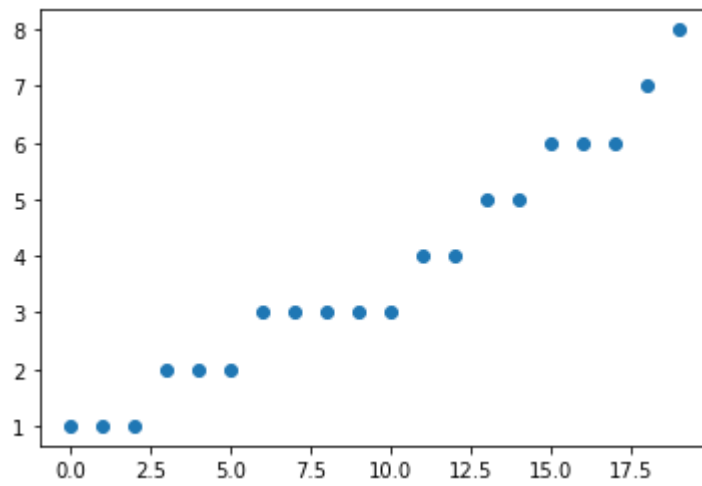
Double-click (or enter) to edit

```
import pandas as pd
#create DataFrame
df = pd.DataFrame({'hours': [1, 1, 1, 2, 2, 2, 3, 3, 3, 3,
                             3, 4, 4, 5, 5, 6, 6, 6, 7, 8]})
#view first five rows of DataFrame
df.head()
```

Double-click (or enter) to edit

```
import matplotlib.pyplot as plt
```

```
plt.scatter(df.index,df['hours'])
plt.show()
```

✓ 1s   completed at 20:33   ● ✕