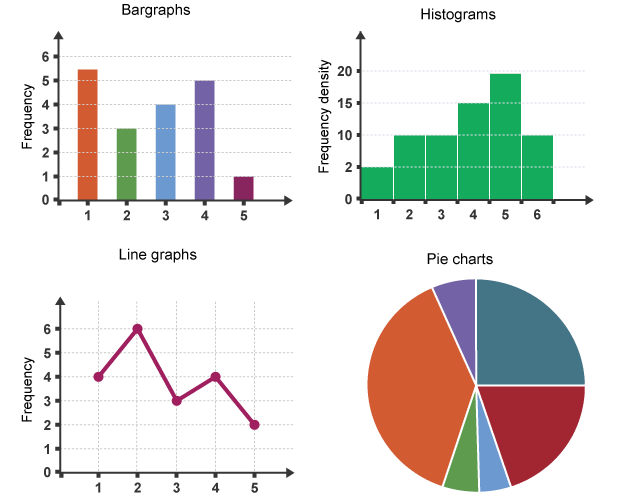
 Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation. The field of Statistics has an influence over all domains of life, the Stock market, life sciences, weather, retail, insurance and education are but to name a few. Statistical information is often shown using graphs and other diagrams. These can be easier to read than lists of numbers.

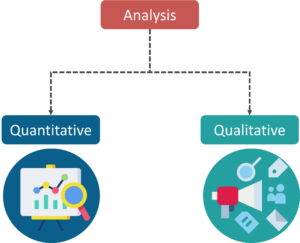


**Terminologies in Statistics – Statistics For Data Science**

* Population is the set of sources from which data has to be collected.
* A **Sample** is a subset of the Population
* A **Variable** is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item.

### ****Types Of Analysis****

An analysis of any event can be done in one of two ways:



1. **Quantitative** **Analysis**: Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identify patterns and trends.
2. **Qualitative** **Analysis**: Qualitative or Non-Statistical Analysis gives generic information and uses text, sound and other forms of media to do so.

For example, if I want a purchase a coffee from Starbucks, it is available in Short, Tall and Grande. This is an example of Qualitative Analysis. But if a store sells 70 regular coffees a week, it is Quantitative Analysis because we have a number representing the coffees sold per week.

For example, typing speed is quantitative. Favorite colors are qualitative/categorical. Some more examples are :

**Quantitative or numerical:**

1. Weight in pounds
2. Length in inches
3. Time in seconds

|  |  |  |
| --- | --- | --- |
| **Qualitative or Categorical:**   1. Model of car 2. Gender 3. Yes or No   **Fundamentals of Statistics-Basic Concepts :: Discrete and Continuous**  If you have quantitative data, like time to complete a task or number of questions correct on a quiz, then the data can be either continuous or discrete. Discrete data have finite values, or buckets. You can count them. Continuous data technically have an infinite number of steps, which form a continuum.  The number of questions correct would be discrete--there are a finite and countable number of questions. Time to complete a task is continuous since it could take 178.8977687 seconds. Time forms an interval from 0 to infinity.  **Discrete:**   1. Number of children in a household 2. Number of languages a person speaks 3. Number of people sleeping in stats class   **Continuous:**   1. Height of children 2. Weight of cars 3. Time to wake up in the morning 4. Speed of the train   Nominal basically refers to categorically [discrete data](http://www.usablestats.com/lessons/datatypes2) such as name of your school, type of car you drive or name of a book. This one is easy to remember because nominal sounds like name (they have the same Latin root).  Ordinal refers to quantities that have a natural ordering. The ranking of favorite sports, the order of people's place in a line, the order of runners finishing a race or more often the choice on a rating scale from 1 to 5. With ordinal data you cannot state with certainty whether the intervals between each value are equal. This is also an easy one to remember, ordinal sounds like order. ****Categories In Statistics****  1. Descriptive Statistics 2. Inferential Statistics  ****Descriptive Statistics****  Descriptive Statistics uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.  Suppose you want to study the average height of students in a classroom, in descriptive statistics you would record the heights of all students in the class and then you would find out the maximum, minimum and average height of the class. Descriptive Statistics Example - Math And Statistics For Data Science - Edureka ****Inferential Statistics****  Inferential Statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.  Inferential statistics generalizes a large data set and applies probability to arrive at a conclusion. It allows you to infer parameters of the population based on sample stats and build models on it.  Inferential Statistics - Math And Statistics For Data Science - Edureka   |  | | --- | |  | |  |  ****Measures Of The Center****  1. **Mean**: Measure of average of all the values in a sample is called Mean.   If we want to find out the mean or average horsepower of the cars among the population of cars, we will check and calculate the average of all values. In this case, we’ll take the sum of the Horse Power of each car, divided by the total number of cars.  Mean = (110+110+93+96+90+110+110+110)/8 = 103.625   1. **Median**: Measure of the central value of the sample set is called Median.   If we want to find out the center value of mpg among the population of cars, we will arrange the mpg values in ascending or descending order and choose the middle value. In this case, we have 8 values which is an even entry. Hence we must take the average of the two middle values.  The mpg for 8 cars: 21,21,21.3,22.8,23,23,23,23  Median = (22.8+23 )/2 = 22.9   1. **Mode**: The value most recurrent in the sample set is known as Mode.   The mode represents the most frequent value in a set of data. For example in the set of data: 3,5,6,7,7,9,8,7,5,6,4,5,3,1 the number 7 is the mode. The mode doesn't have to be the center of a set of data and there can be more than one mode. Measures of Variability The center of a distribution of data is helpful in telling you something about the most common values, but it doesn't tell you much about how spread out or variable the values are. To describe variability, we need an additional measure.  **Range:** the difference between the highest and lowest value in a dataset.  **Variance** (σ2): measures how spread out a set of data is relative to the mean.  **Standard Deviation (σ):**another measurement of how spread out numbers are in a data set; it is the square root of variance.  The shape of a distribution is described by the following characteristics.   * **Symmetry**. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other. * **Number of peaks**. Distributions can have few or many peaks. Distributions with one clear peak are called **unimodal**, and distributions with two clear peaks are called **bimodal**. When a symmetric distribution has a single peak at the center, it is referred to as **bell-shaped**. * **Skewness**. When they are displayed graphically, some distributions have many more observations on one side of the graph than the other. Distributions with fewer observations on the right (toward higher values) are said to be **skewed right**; and distributions with fewer observations on the left (toward lower values) are said to be **skewed left**. * **Uniform**. When the observations in a set of data are equally spread across the range of the distribution, the distribution is called a **uniform distribution**. A uniform distribution has no clear peaks.  Unusual Features Sometimes, statisticians refer to unusual features in a set of data. The two most common unusual features are gaps and outliers.  **Gaps**. Gaps refer to areas of a distribution where there are no observations.  **Outliers**. Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called outliers. Univariate, Bivariate data and its analysis  1. **Univariate data –**This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.   The example of a univariate data can be height.Suppose that the heights of seven students of a class is recorded,there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.   1. **Bivariate data –**This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the 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. 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. Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent. |
|  |