Data science is the science which uses Computer science, Statistics and Machine Learning, Visualization and Human computer interactions to collect , clear, integrate, analyze, visualize interact with data to create data products.

Data science stands for study of processes and systems that extract knowledge or insights from data in various forms either structured or un-structured.

Data science is a combination of source of the fields such as statistics, data mining and predictive analysis.

Goal of data science is to turn data into data products. Data science quarrying the past, present, future.

Data science explore many models, build and tune many hybrids, understand empirical properties of models.

- ***Statistics:***

Now a days in order to learn something, you must first collect data. Statistics is the art of learning from data.

Statistics is branch of mathematics that deals with the collection , analysis and interpretation of data.

Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as with valid drawing conclusions and making reasonable decisions on the basis of such analysis.

Statistics may be defined as the science of collection, presentation, analysis and interpretation of numerical data.

Statistics is classified into two categories:

1. Descriptive statistics 2. Inferential statistics

1. Descriptive statistics:

1.Data should be described and summarized is known as Descriptive data. Such type of statistics is known as **Descriptive Statistics.**

2.Descriptive statistics are the numbers that are used to summarize and describe data.

3.Descriptive statistics is just descriptive; it is not generalizing beyond the data at hand.

4.Descriptive statistics is a collection of methods for summarizing data.

Example :Mean, Median, Mode, Range, Variance, Graphs, etc.

2. Inferential statistics:

A statistics concerned with drawing conclusions is called **Inferential statistics.**

i.e Inferential statistics is a collection of methods for using sample data to make conclusions about a population.

Population Vs Sample:

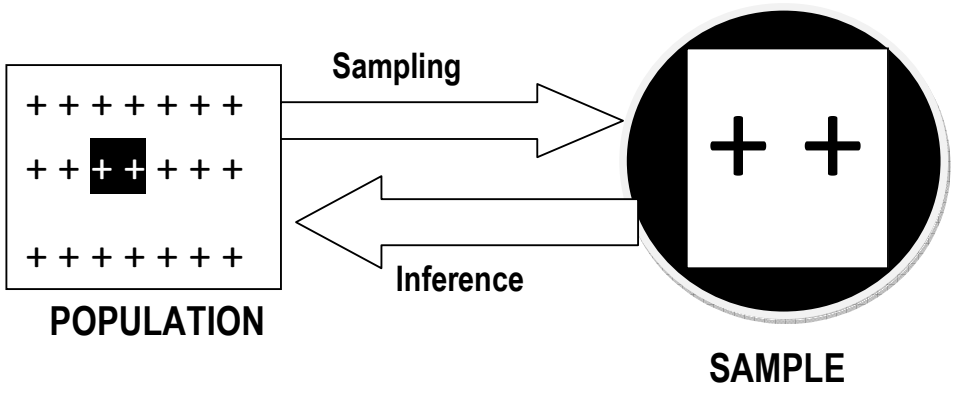
To draw logical conclusions from data, we make some assumptions about chances (probability) of obtaining different data values.

In statistics, we are interested in obtaining the information about a total collection of objects, it is referred to as the population.

The entire set of all possible observations in which we are interested is known as **population**.

If the population is too large for us to examine each of its members and it is not easy, we choose some subset of its elements. This subset of populatio

n is called a **Sample.**



❖ What is Data

Look around you, there is data everywhere

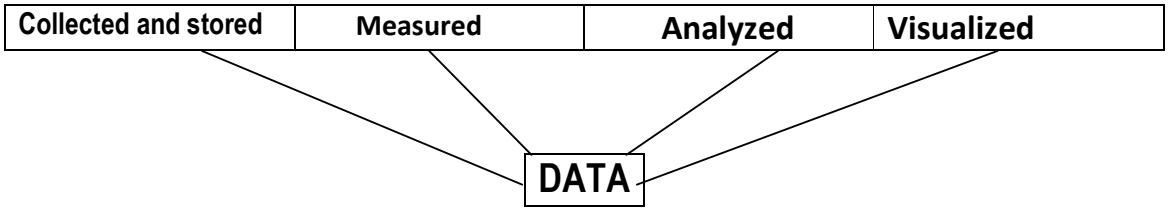
Each click on your phone generates more data than you know.

Data is the collected information (observations) we have about something or facts and statistics collected for reference or analysis.

Data can be defined as groups of information that represents qualitative or quantitative attributes of variables or set of variables.

A collection of facts(Numbers, words, measurements, observations etc) that has been translated into a form that computer can process.

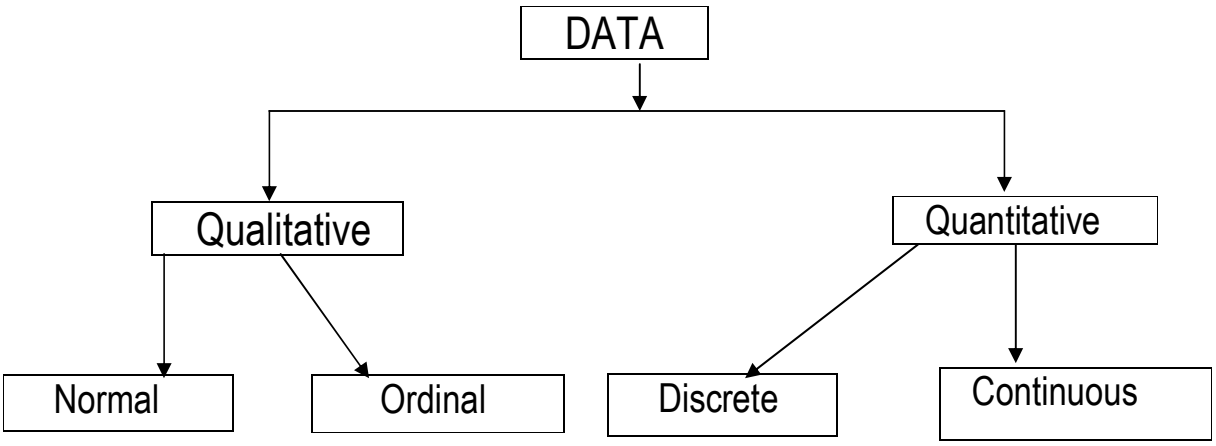
This generated data provides insights for analysis and helps us make better business decisions.



1.Data refers to facts and statistics collected together for analysis or reference.
2.Data can be collected, measured and analyzed.
3.It can also be visualized by using statistical models and graphs.

Categories of data:

1.Qualitative data 2. Quantitative data



Qualitative data:

It deals with characteristics and descriptors that cannot be easily measured, but can be observed subjectively.

Qualitative data is further divided into two types

- ***Nominal data:***

  Data with no inherent order or ranking such as gender or race.

  Example:- male ,female ⋯

- ***Ordinal data:***

  Data with an ordered series of information is called **ordinal data.**

  Example:-

Customer id	Rating
1	average
2	bad
3	good

Quantitative data:

Quantitative data deals with numbers and things you can measure objectively.

Quantitative data is further divided into two types.

Discrete data(also known as categorical data)

It can hold a finite number of values.

Example:- No.of students in a class.

Continuous data

Data that can hold an infinite number of values.

Example :-weight of a person.

There are two major approaches for gathering information about a situation, person, problem or phenomenon.

*Data collection plays a very crucial role in the statistical analysis. In research, there are different methods used to gather information. Based on the data can be categorized as

1. Primary data. **2.Secondary data.**

Data collection:

Data collection is the process of acquiring information from different sources about the topic under research.

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of technique for collecting data .if data is not properly collected an investigator may not be able to answer the questions under consideration with reasonable degree of confidence.

There are two major approaches for gathering information about a situation, person, problem or phenomenon.

Data collection plays a very crucial role in the statistical analysis. In research , there are different methods used to gather information. All of which falls into two categories.

1.Primary data. 2.Secondary data.

Primary data is one which is collected for the first time by researcher.
Primary data is the actual and original where as the **secondary data** is just analysis and interpretation of the primary data.

Secondary data is the data already collected or produced by the others 19 Differences between primary and secondary data.

Basis of comparison	Primary data	Secondary data
Meaning	Primary data refers to the first hand data gathered by researcher himself	Secondary data means data collected by someone else earlier
Data	Real time data	Past data
process	Time consuming	Quick and easy
source	Survey, experiment, interview, observation, questionnaire	Books, journals, newspapers, internal records, Government publications, websites, etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Available in	Raw form	Refined form
Specific	Always specific to researcher needs	May or may not be specific to the researcher
Accuracy and reliability	More	Relatively less
Specific	Always specific to researcher needs	May or may not be specific to the researcher
Information	First hand data	Second hand data
Form	Pure & Raw	Refined
Collector and user	Same person	Different person

Variables:

Variables are properties or characteristics of some event, object, or person that can take an different values or amounts.

Variables may be independent or dependent
qualitative or quantitative
discrete or continuous

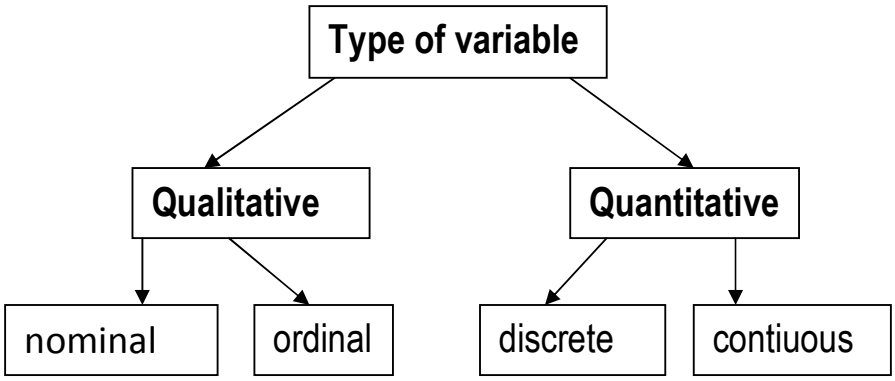
Independent and dependent variables

A variable can serve as independent in one study and dependent in another.

When conducting research experiments after manipulate variables.

For example, an experiment might compare the effectiveness of four types of drugs. In this case the variable is "type of drug". When a variable is manipulated by an experimenter, then it is called an "independent variables".

1. The experiment seeks to determine the effect of the **independent variable** on relief from depression.
2. In this example, relief from depression is called a **dependent variable**. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.



QUALITATIVE AND QUANTITATIVE VARIABLES

Qualitative variables are those that express a quality attributes such as hair, color, eye color, religion, favorite movie, gender and so on..... The values of qualitative variable do not imply a numerical ordering.

Qualitative variables are also referred to as categorical variables. Qualitative variables are the variables which takes the values that cannot be ordered in a logical or natural way.

For example the color of eye, the name of political party, type of transport used to travel to work all are Qualitative variables.

- It is common to assign numbers to qualitative variables, for practical purpose in data analysis.
- For intake, if we consider the variable 'gender', then each observation can take either the value 'male' or 'female'. We may decide to assign 1 to male and 0 to female. Then this variable gender remains a qualitative variable.

Quantitative variables

Quantitative variables represent measurable quantities. The values which these variables can take can be ordered in a logical and natural way.

Quantitative variables are those variables that are measured in terms of numbers.

Example for quantitative variables are weight, height, size, Size of shoes, price for houses, no. of semesters studied, weight of a person.

Discrete variables are variables which can only take a finite number of values.

All qualitative variables are discrete such as the color of eye, or the region of country.

But quantitative variables can be discrete.

Number of shoes ,No. of semesters in a class

Continuous variables

A variable which can take an infinite number of values are called **Continuous variables.**

The time it takes to travel to university. The length of an antelope.

The distance between two planets

Sometimes, continuous variables are variables which are measured rather than counted.

The thoughts and considerations from above indicates that different variables contain different amount of information. A useful classification of scale of a variable by the concept of scale of a variable.

- ***Nominal scale :-***

The values of a nominal variable cannot be ordered. Ex: gender of a person (male – female),The status of the application (pending – not pending)

- ***Ordinal scale :-***

The values of an ordinal variable can be ordered. However the difference between these values cannot be interpreted in a meaningful way. Ex: The possible values of education level (none – primary education – secondary education – university degree) can be ordered meaningfully but the differences between these values cannot be interpreted. The satisfaction with a product (unsatisfied – satisfied – very satisfied) is an ordinal variable because the values of the variable can be ordered but cannot be compared in a numerical way.

- ***Continuous scale :-***

The values of a continuous variable can be ordered. Also the differences between these values can be interpreted in a meaningful way.

The height of a person refers to a continuous variable because the values can be ordered and the difference of these values can be compared.

Interval scale :-

Only differences between the values but not ratios can interpreted.

Ex : Scale should be temperature (measured in ºC).

The difference between -2ºC & 4ºC is 6ºC.

But ratio of 4/-2 is '-2' doesn't meaning that -4ºC is twice as cold as 2ºC.

- ***Ratio scale :-***

Both differences and ratios can be interpreted.

Ex : Speed 60kmph more than 20kmph. 60kmph is three times faster than 20kmph because ratio between them is 3.

- ***Absolute scale :-***

The absolute scale is the same as the ratio scale with the exception that the values are measured in 'natural' units.

Ex : No. of semesters studied where no artificial unit such as kmph or ºC is needed. The values are simply 1,2,3,…

- ***Grouped data :-***

Sometimes, data maybe available in a summarized form instead of original value. One may know only category or group the values belongs to.

If data is available in grouped form, we call the respective variable capturing their information a grouped variable. Sometimes, these variables are also known as categorical variables.

Any grouped or categorical variables which can only take two values is called a binary variable.

Key points :-

- Qualitative data is always discrete but quantitative data can be both discrete (size of shoes or a grouped variable) and continuous (Ex. temp)
- Nominal variables are always qualitative and discrete (Ex. Color of the eye)
- Whereas continuous variables are always quantitative (Ex. temp)

- Categorical variables can be both qualitative (Ex. Clor of the eye) and quantitative (Ex. Satisfaction level on scale) Categorical variables are never continuous.

Classification of variables