UNIT 1

1.1 Statistics and Types of Statistics

- Definition
- ***Statistics*** is the science of collecting, analyzing, presenting, and interpreting data, as well as of making decisions based on such analyses.

- Statistics is a mathematical science that includes methods for collecting, organizing, analyzing and visualizing data in such a way that meaningful conclusions can be drawn.

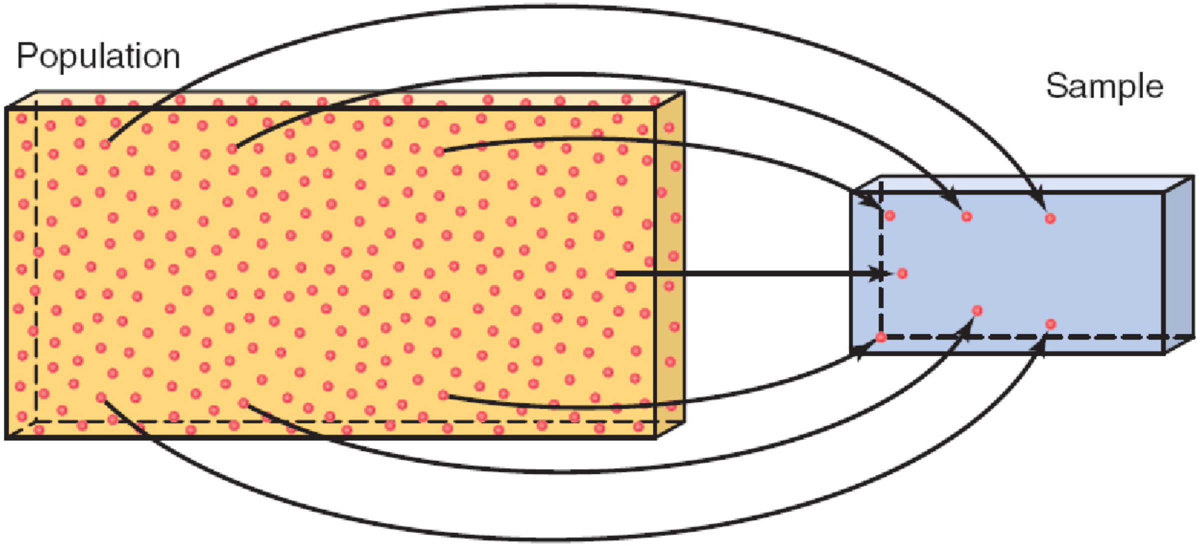
- Statistics is also a field of study that summarizes the data, interpret the data making decisions based on the data

Basic terminology of Statistics

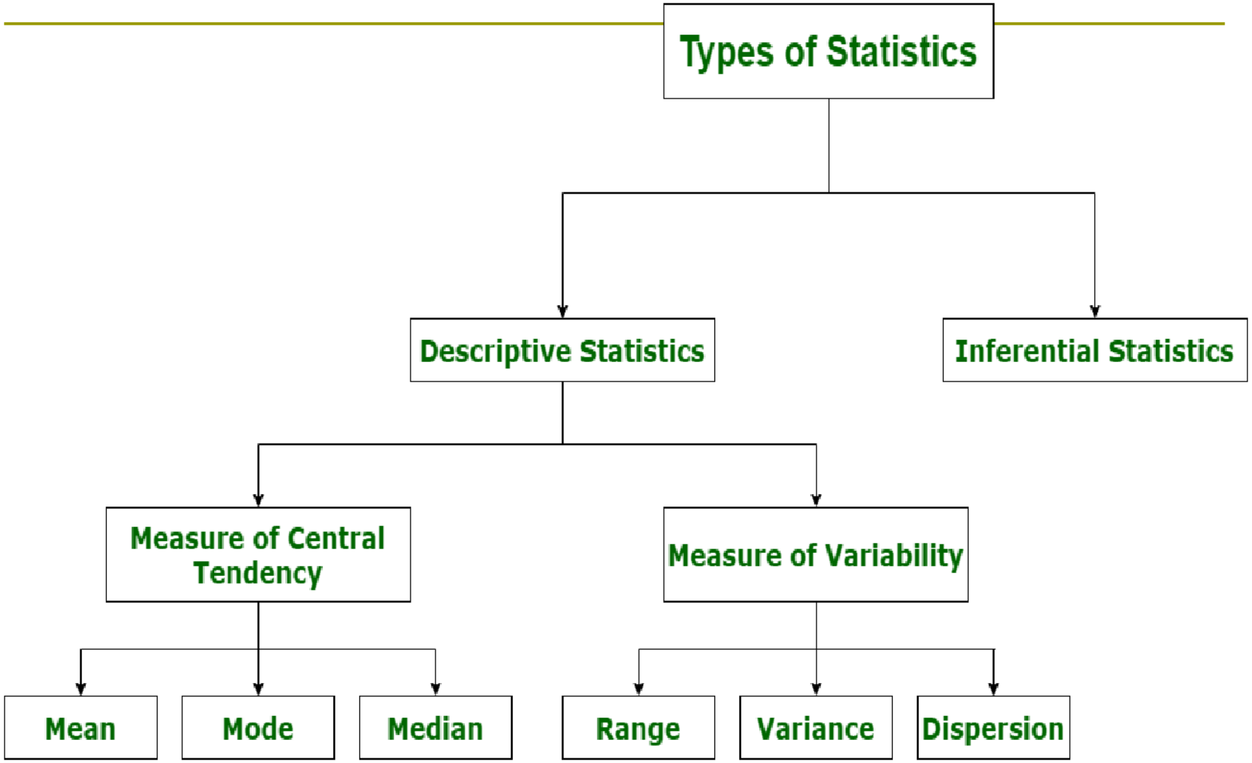
- **Population –**
  It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.

- **Sample –**
  It is the subset of a population.

Figure 1.1 Population and Sample



TYPES OF STATISTICS



1. Descriptive Statistics :

- Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table.

- It provides a graphical summary of data. It is simply used for summarizing objects, etc..

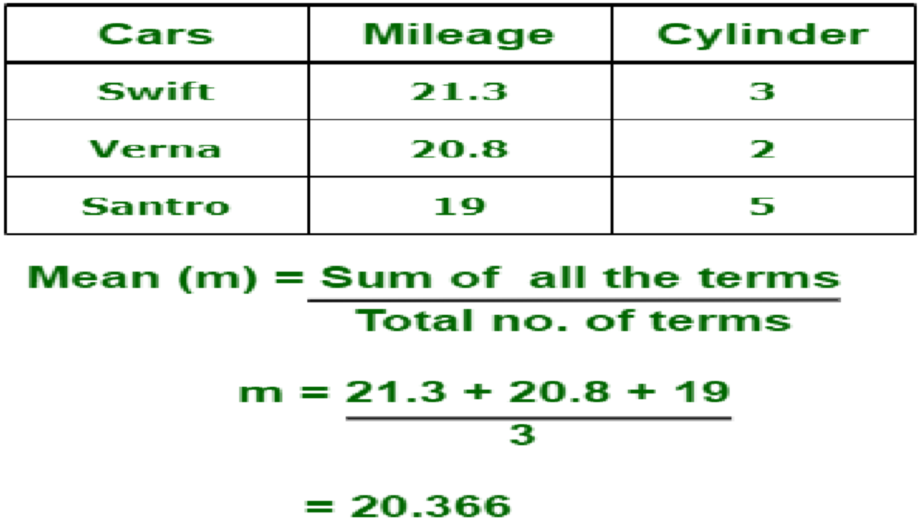
Measure of central tendency

- Measure of central tendency is also known as summary statistics that is used to represents the center point or a particular value of a data set or sample set.
- In statistics, there are three common measures of central tendency as shown below:

1. MEAN
2. MODE
3. MEDIAN

Mean :

- It is measure of average of all value in a sample set.



Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

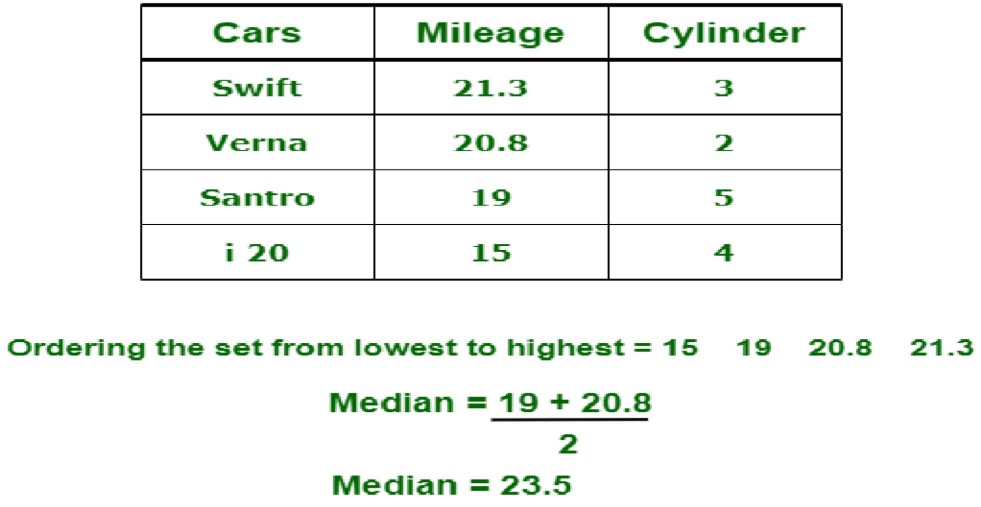
$$\text{Mean (m)} = \frac{\text{Sum of all the terms}}{\text{Total no. of terms}}$$

$$m = \frac{21.3 + 20.8 + 19}{3}$$

$$= 20.366$$

Median

- It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.



Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i 20	15	4

Ordering the set from lowest to highest = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2}$$

Median = 23.5

Mode

- The mode is the most frequent score in our data set
- The value repeated most of time in central set is actually mode.

2 3 4 2 4 6 4 7 7 4 2 4

Mode = 4

Measure of Variability:

- Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below

1. **Range**
2. **Variance**
3. **Dispersion :**

Range:

- It is given measure of how to spread apart values in sample set or data set
- Range = Maximum value - Minimum value
- Find the range of given observations: 32, 41, 28, 54, 35, 26, 23, 33, 38, 40.
- arrange the given values in ascending order. 23, 26, 28, 32, 33, 35, 38, 40, 41, 54
- Range (X) = Max (X) – Min (X)

  = 54 – 23

  = 31

Variance:

- It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.
- $\mathbf{S^2 =} \sum^{n}_{i=1} [(x_i - \overline{x})^2 \div n]$

Where, **n** represent total data points

$\overline{\mathbf{x}}$ represent mean of data points

$\mathbf{x_i}$ represent individual data points.

Standard Deviation

- It is defined as the square root of the variance. It is calculated by finding the Mean, then subtracting each number from the Mean which is also known as the average, and squaring the result. Adding all the values and then dividing by the no of terms followed by the square root.

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Quartiles in Statistics:

- Similar to the median which divides the data into half so that 50% of the estimation lies below the median and 50% lies above it, the quartile splits the data into quarters so that 25% of the estimation are less than the lower quartile, 50% of estimation are less than the mean, and 75% of estimation are less than the upper quartile. Usually, the data is ordered from smallest to largest:

- First quartile: 25% from smallest to largest of numbers
- Second quartile: between 25.1% and 50% (till median)
- Third quartile: 51% to 75% (above the median)
- Fourth quartile: 25% of largest numbers
- Interquartile range (*IQR*) and is defined as *IQR* = *Q*3 - Q1.

Quartiles Example

- **Find the quartiles of the following data: 4, 6, 7, 8, 10, 23, 34**

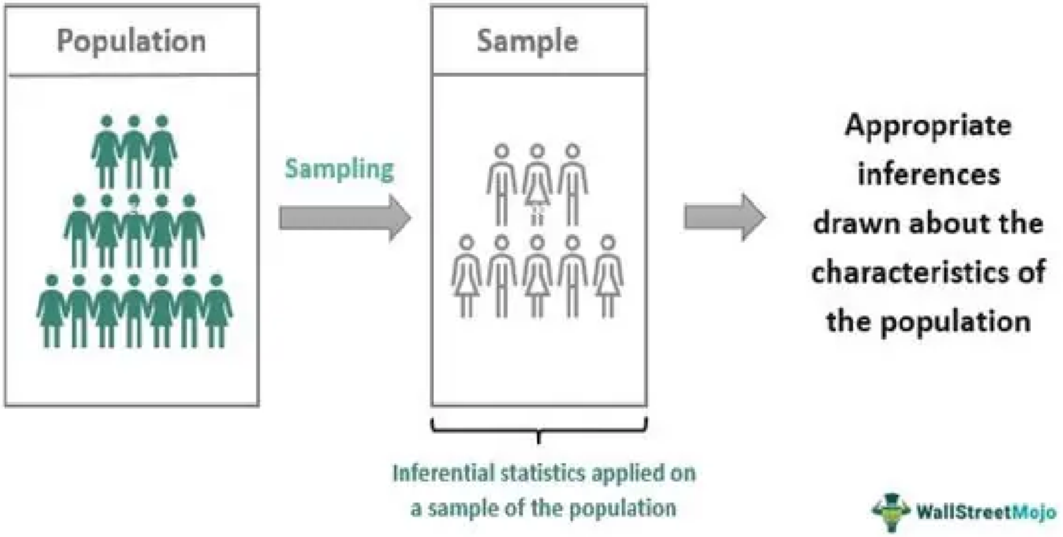
- Q_1 = [(n+1)/4]th item
- Q_2 = [(n+1)/2]th item
- Q_3 = [3(n+1)/4]th item

- Here the numbers are arranged in the ascending order and number of items, n = 7
- Lower quartile, Q_1 = [(n+1)/4] th item
- Q_1= 7+1/4 = 2nd item = 6
- Median, Q_2 = [(n+1)/2]th item
- Q_2= 7+1/2 item = 4th item = 8
- Upper Quartile, Q_3 = [3(n+1)/4]th item
- Q_3 = 3(7+1)/4 item = 6th item = 22

Inferential Statistics:

- Definition
- ***Inferential Statistics*** consists of methods that use sample results to help make decisions or predictions about a population.
- Inferential Statistics makes inference and prediction about population based on a sample of data taken from population
- It generalizes a large dataset and applies probabilities to draw a conclusion
- Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis.
- Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population

Inferential Statistics



Appropriate inferences drawn about the characteristics of the population

Inferential statistics applied on a sample of the population

Types of Inferential Statistics

- Inferential statistics can be broadly categorized into two types:

- **Hypothesis Testing**
- **Regression Analysis**

Hypothesis Testing

- Hypothesis testing involves using a sample of data to test a hypothesis about a population parameter.
- A hypothesis is a statement about a population parameter, such as the population mean or proportion, that the researcher wants to test

Example of Hypothesis Testing

- Suppose a car manufacturer claims that their cars have an average fuel efficiency of 30 miles per gallon (mpg). A researcher wants to test whether this claim is true.
- **The null hypothesis (H0) is that the average fuel efficiency of the cars is 30 mpg**
- **And the alternative hypothesis (Ha) is that the average fuel efficiency is different from 30 mpg.**

- The researcher takes a sample of 50 cars and measures their fuel efficiency. The sample has a mean fuel efficiency of 28 mpg and a standard deviation of 4 mpg.
- The researcher may then report the results and recommend that the car manufacturer re-evaluate their claim or investigate potential factors that may be causing the lower-than-expected fuel efficiency.

Regression Analysis

- Regression analysis is used to model the relationship between one or more independent variables and a dependent variable
- The goal of regression analysis is to create a mathematical equation that can be used to predict the value of the dependent variable based on the values of the independent variables

Example of Regression Analysis

- Suppose you’re a real estate agent and you want to predict the selling price of a house based on its size (in square feet) and the number of bedrooms it has. To do this, you collect data on recently sold houses in a particular neighborhood, including the selling price, size, and number of bedrooms for each house. You can then use regression analysis to build a model that predicts the selling price of a house based on its size and number of bedrooms.

1.2 Population Versus Sample

- Definition
- A ***population*** consists of all elements – individuals, items, or objects – whose characteristics are being studied. The population that is being studied is also called the ***target population***.

- A portion of the population selected for study is referred to as a ***sample***.

POPULATION VERSUS SAMPLE

- Definition
- A survey that includes every member of the population is called a ***census***. The technique of collecting information from a portion of the population is called a ***sample survey***.

POPULATION VERSUS SAMPLE

- Definition
- A sample that represents the characteristics of the population as closely as possible is called a ***representative sample***.

POPULATION VERSUS SAMPLE

- Definition
- A sample drawn in such a way that each element of the population has a chance of being selected is called a ***random sample***. If all samples of the same size selected from a population have the same chance of being selected, we call it ***simple random sampling***. Such a sample is called a ***simple random sample***.

POPULATION VERSUS SAMPLE

- A sample may be selected with or without replacement.

- In sampling ***with replacement***, each time we select an element from the population, we put it back in the population before we select the next element.

- Sampling ***without replacement*** occurs when the selected element is not replaced in the population.

1.3 Basic Terms

- Definition
- An ***element*** or ***member*** of a sample or population is a specific subject or object (for example, a person, firm, item, state, or country) about which the information is collected.

BASIC TERMS

- Definition
- A ***variable*** is a characteristic under study that assumes different values for different elements. In contrast to a variable, the value of a ***constant*** is fixed.

BASIC TERMS

- Definition
- The value of a variable for an element is called an ***observation*** or ***measurement***.

- A ***data set*** is a collection of observations on one or more variables.

Table 1.1 Total Revenues for 2010 of Six Companies

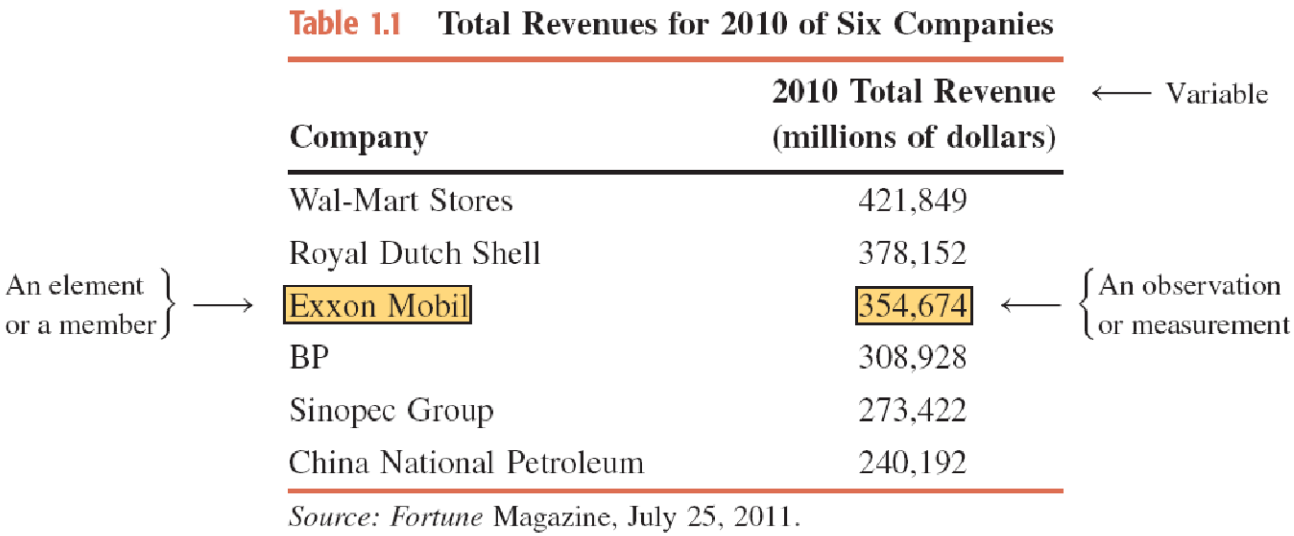


Table 1.1 **Total Revenues for 2010 of Six Companies**

Company	2010 Total Revenue (millions of dollars)
Wal-Mart Stores	421,849
Royal Dutch Shell	378,152
Exxon Mobil	354,674
BP	308,928
Sinopec Group	273,422
China National Petroleum	240,192

Source: Fortune Magazine, July 25, 2011.

← Variable

An element or a member →

← An observation or measurement

1.4 Types of Variables

- Quantitative Variables
  - Discrete Variables
  - Continuous Variables

- Qualitative or Categorical Variables

Quantitative Variables

- Definition
- A variable that can be measured numerically is called a ***quantitative variable***. The data collected on a quantitative variable are called ***quantitative data***.

Quantitative Variables: Discrete

- Definition
- A variable whose values are countable is called a ***discrete variable***. In other words, a discrete variable can assume only certain values with no intermediate values.

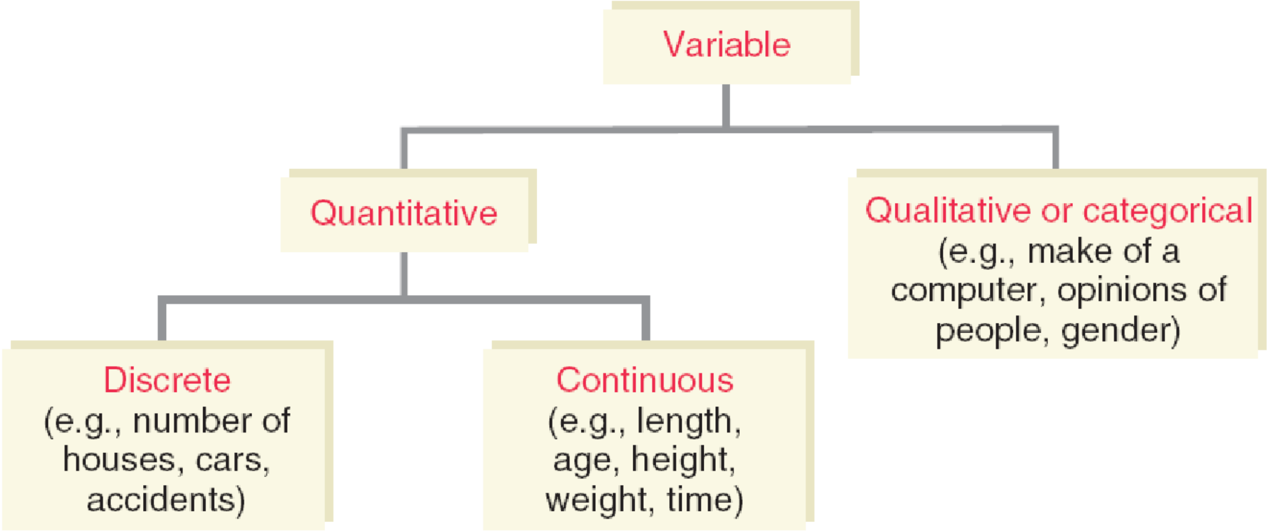
Quantitative Variables: Continuous

- Definition
- A variable that can assume any numerical value over a certain interval or intervals is called a ***continuous variable***.

Qualitative or Categorical Variables

- Definition
- A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a ***qualitative*** or ***categorical variable***. The data collected on such a variable are called ***qualitative data***.

Figure 1.2 Types of Variables


Variable
Quantitative
Qualitative or categorical (e.g., make of a computer, opinions of people, gender)
Discrete (e.g., number of houses, cars, accidents)
Continuous (e.g., length, age, height, weight, time)

1.6 Sources of Data

- Data may be obtained from
  - Internal Sources
  - External Sources
  - Surveys and Experiments