Statistics

Statistics is the science of collecting , organizing , presenting , analyzing and interpreting numerical data to assist in making more effective decisions.

- If we have some data to modify , we use following steps, Generally:

    1.Data is available

    2.Preprocess the data

    3.organize the data

    4.Representation of data

    5.Interpreting the data

    6.Presentation of data

    7.Draw the Conclusion

- **Data** divides into two parts :

    1. Columns

    2. Rows

- **Columns** can be represented as features , variables and predictors.

- **Rows** can be represented as samples , datapoints , tuples , fields.

****Columns** have two types :

    1.Categorical Columns

    2.Numerical Columns

1.Categorical Columns

- Categorical columns (or categorical variables) represent data that can be divided into distinct categories or groups

- categorical data is used to classify items into different categories based on qualitative attributes. **Eg:** degrees , gender ,..etc.

    * The data reprents in English

2.Numerical Columns

- Numerical columns (or numerical variables) in statistics represent data that can be measured and quatified.

- Quantitative

- Numerical data is generally categorized into two main types:

    1.Discrete

    2.Continuous

1.Discrete data :

- Integer type is called as discrete.

- int

- countable

- Eg: Rollnumbers , Backlogs

2.Continuous data :

- Decimal points number is called as continuous.

- Float type data

- Eg: height=152.25cm , weight=65.5kgs

Categorical data	Numerical data
• Qualitative data	• Quantitative data
	• Continoues data :Float
	• Discrete data :Int type

Levels of Measurements

1.Nominal-level Data

2.Ordinal-level Data

3.Interval-level Data

4.Ratio-level Data

1.Nominal-level Data :

- Data that is under the Categorical type.
- Data cannot be arranged in any particular order.
- It is just a name . There is no relation between names . Any names

Properties:

- Observations of a quantitative variables can only be **classified** and **counted.**
- There is no particular order to labels.

EXAMPLES: eye color , gender , religious afflication.

2.Ordinal-level Data :

- Data comes under the Categorical type .
- Data is arranged in some order , but the differences between the datavalues cannot be determined or are meaningless.

Properties:

- Data classifications are represented by sets of labels or names (high , medium , low) that have **relative values.**
- Because of the relative values , the data classified can be **ranked or ordered.**

EXAMPLES: 1.During a taste test of 4 soft drinks.

- Mellow yellow was ranked 1 , Sprite number 2 , Seven-up number 3, and orange crush number 4 .

2. Student score – 1st Rank , 2nd Rand , 3rd Rank

3. Ug , pg , phd

3.Interval-level Data :

- Data comes under the Numerical data.
- Similar to the ordinal level , with additional property.
- It does not have True Zero Point.
- Generally a scale starts from zero ,
- But here if in the scale we see zero also there is **no meaning of zero.**
- It can have negative values .
- Zero means OFF , **no true zero point**.

Properties :

- Data classifications are ordered according to the amount of the characteristics they prossess.

- Equal differences in the characteristics are represented by equal differences in the measurements.

EXAMPLE: 1.Temperature on the Fahrenheit scale.

2. Year data

3. Time data

4.Ratio-level data :

- Data comes under the Numerical data.

- The interval level with an inherent zero starting point. Means have zero scale.

- Differences and ratios are meaningful for this level of measurement. Means ratio is possible

- It does not have any negative values

Properties:

- Data classifications are ordered according to the amount of the characteristics they possess.

- Equal differences in the characteristics are represented by equal differences in the number assigned to the classifications.

- The zero point is the absence of the characteristics and the ratio between two numbers is meaningful.

EXAMPLES: Monthly income , distance travelled ,height , weight.

Q : Suppose the temperature in Hyderabad = 50c , Bangalore = 25c

Can I say : Hyderabad = 2*Bangalore

Here ,

50c = 2*25c

50c ==== 122F

25c ==== 77F

But,

50/25 ! = 122/77

So , by this we understand that ratio is not possible everywhere.

Q : A Father weight = 100kgs , son has a weight = 50kgs

Father = 2*son

100kgs = 2*50kgs

100kgs ==== 220.462 pounds

50kgs ==== 110.231 pounds

100/50 = 220.462/110.231

In this case ratio is possible.

Note: Ratio level means , If you take the values in any measurement

Ratio should be equal

Categorical data	Numerical data
• Qualitative data	• Quantitative data
	• Continoues data :Float
	• Discrete data • :Int type
Nominal	Interval
Ordinal	Ratio

Purpose to know the Level of Measuement of a Data

- The level of measurement of the data dictates the calculations that can be done to summarize and present the data.

- To determine the statistical tests that should to be performed on the data.

Population versus Sample

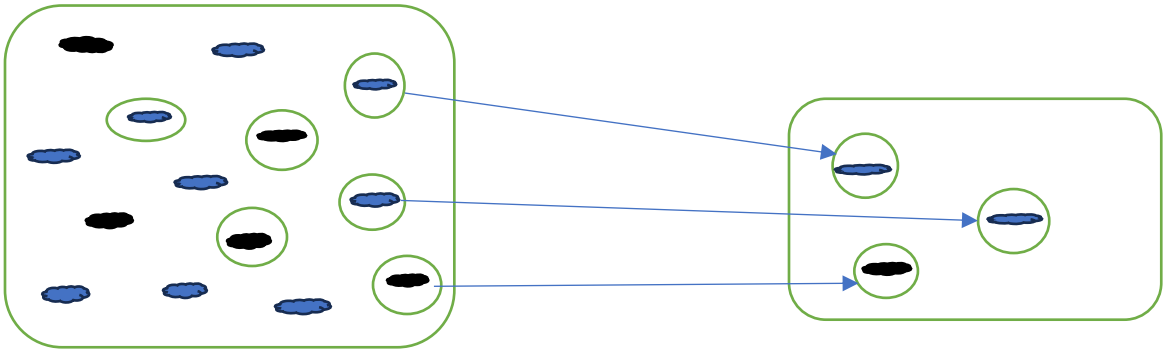
Population is a **collection** of all possible individuals , objects , or measurements of interest.

- Each and every point in the world Population.

Sample is a **portion** or **part** , of the population of interest.

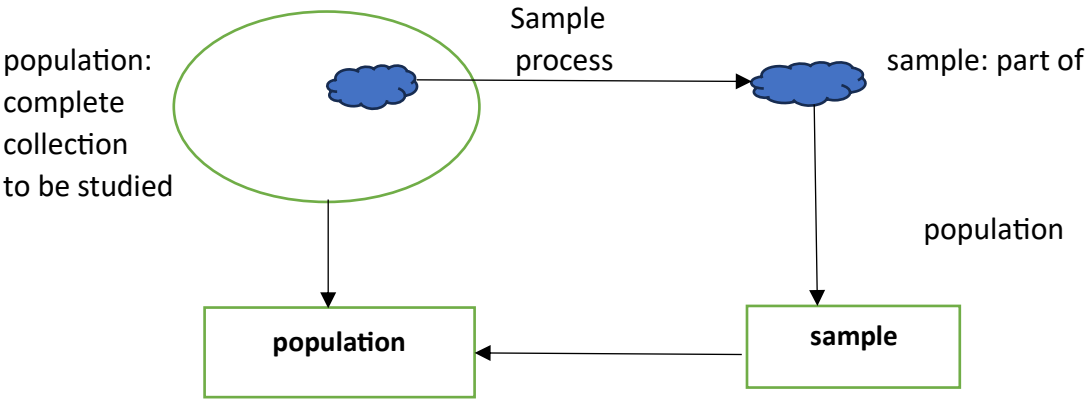
Population : All items

Sample : Items selected from the population



Population:All items

Sample: selected from population



Why we taking sample instead of studying every member of population?

- Prohibitive cost of census.

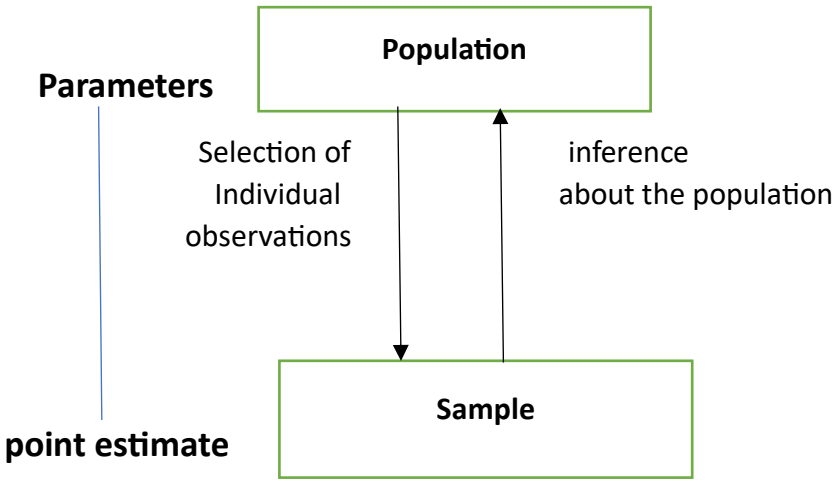
- Destruction of item being studied may be required.

- Not possible to test or inspect all members of a population being studied.

Use of Sample in Learning about Population

- Using a sample to learn something about a population is done extensively in business , agriculture , politics , and government.

EXAMPLE : Television networks constantly monitor the popularity of their programs by hiring Neilsen and other organizations to sample the preferences of TV veiwers.



Why Statistics ?

- Numerical information is everywhere.

- Statistics techniques are used to make decisions that affect our daily lives.

- The knowledge of Statistical methods will help you understand how decisions are made and give you a better understanding of how they affect you.

No matter what line of work you select, you will find yourself faced with decisions where an understanding of data analysis is helpful.

What is mean by Statistics ?

- In the more common usage, statistics refers to numerical information

Examples: the average starting salary of college graduates, the number of deaths due to alcoholism last year, the change in the Dow Jones Industrial Average from yesterday to today, and the number of home runs hit by the Chicago Cubs during the 2007 season.

- We often present statistical information in a graphical form for capturing reader attention and to portray a large amount of information**.**

Some examples of the need for data collection

- Research analysts for Merrill Lynch evaluate many facts of a particular stock before making a "buy" or "sell" recommendation.

- The marketing department at Colgate-Palmolive Co., a manufacturer of soap products, has the responsibility of making recommendations regarding the potential profitability of a newly developed group of face soaps having fruit smells.

- The United States government is concerned with the present condition of our economy and with predicting future economic trends.

- Managers must make decisions about the quality of their product or service.

Who Uses Statistics ?

Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, etc...

Types of Statistics:

1.Descriptive Statistics

2.Inferential Statistics

1.Descrptive Statistics :

- Methods of organizing, summarizing, and presenting data in an informative way.

- Analyse on the population, draw the conclusion on population

EXAMPLE 1: The United States government reports the population of the United States was 179,323,000 in 1960; 203,302,000 in 1970; 226,542,000 in 1980; 248,709,000 in 1990

EXAMPLE 2: According to the Bureau of Labor Statistics, the average hourly earnings of production workers was $17.90 for April 2008.

2.Inferential Statistics :

- A decision, estimate, prediction, or generalization about a population, based on a sample.

Note: In statistics the word population and sample have a broader meaning.

A population or sample may consist of individuals or objects.

- Analyse on the sample, draw the conclusion on population.