STATISTICS FOR DATA SCIENCE

Sampling

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Sampling

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Topics to be Covered

- Statistical Analysis
- Population
- Sample
- Sampling
- Types of Population

Problems to be solved

Suppose, you are interested in finding

- Mean height of all male students of all the universities in India. OR
- Average marks of all female students of PES University. OR
- Relationship between the time a student spends on studying and the grades that he gets. OR
- Impact of rise in number of student assignments on their grades.

STATISTICS FOR DATA SCIENCE

Statistical Analysis

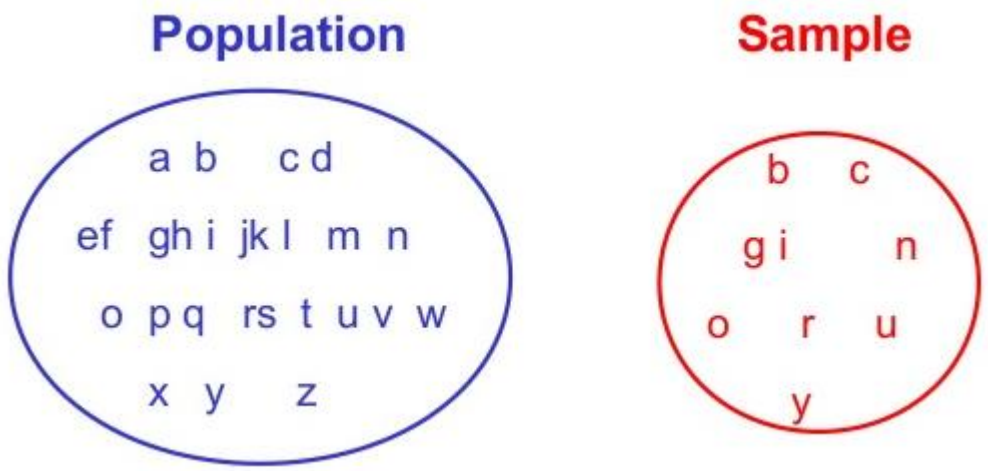
Statistical analysis is the **science of collecting data** and **uncovering patterns and trends**.

Sources: steemkr.com, wifflegif.com

STATISTICS FOR DATA SCIENCE

First step in Statistical Analysis

Identify whether the data set is a Population or a Sample.

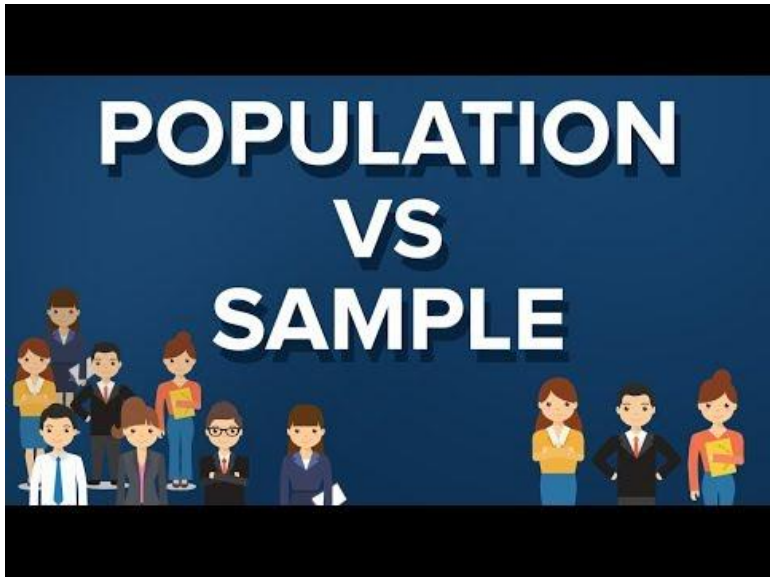


Source: www.slideshare.net,linguistics.byu.edu

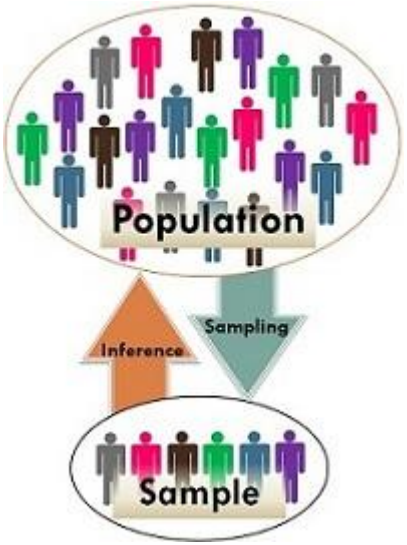
STATISTICS FOR DATA SCIENCE

Population vs. Sample

A **population** is the entire collection of all items(or objects) of interest to our study.



A **sample** is a subset of a population.



Source Image : 365datascience.com

TextBook-Chapter1.1-Pg.No-3

STATISTICS FOR DATA SCIENCE

Is it population or sample?

Study : Survey of the job prospects of the students studying in a university.

Meeting every student in the university to take a survey – Population or Sample?

Source: doc.nv.gov, keydifferences.com, www.2ezwebdesign.com

STATISTICS FOR DATA SCIENCE

Population vs. Sample

So, population is hard to define and hard to observe in real life.

A sample, however, is much easier to contact.

Samples are:

- Easier to contact
- Less time consuming
- Less costly

Source: keydifferences.com

Why Sample ?

Get information about large populations

- Lower cost
- More accuracy of results
- High speed of data collection
- Availability of population elements
- Less field time
- When it is impossible to study the whole population

STATISTICS FOR DATA SCIENCE

What Type of Sample?

Study : Survey of the job prospects of the students studying in a university.

Taking survey from the students who are in Canteen.

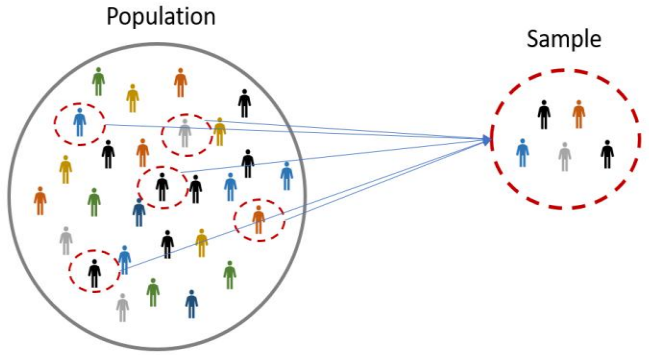
Source: www.slideshare.net, keydifferences.com, , www.2ezwebdesign.com

STATISTICS FOR DATA SCIENCE

What is Good Sample?

The sample must be:

- **representative of the population**
- appropriately sized(larger the better)
- **random(selections occur by chance)**
- unbiased



Source Image: www.sigmamagic.com

STATISTICS FOR DATA SCIENCE

Is it a Good Sample?

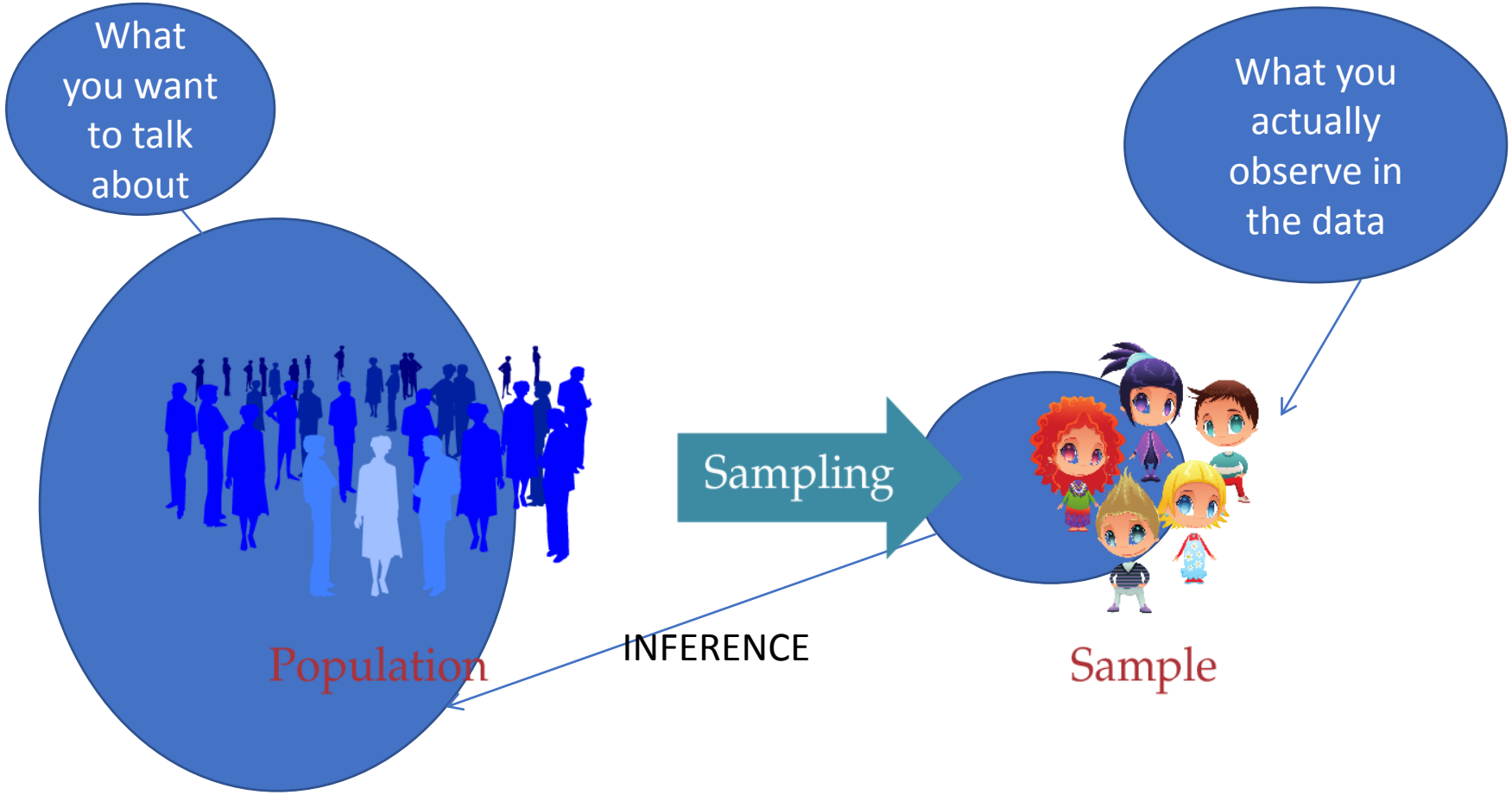
- Is this a representative of population?
- Is this a random sample?

Source: www.slideshare.net, , www.2ezwebdesign.com, giphy.com

STATISTICS FOR DATA SCIENCE

What is sampling?

The **process of selecting observations(a sample**) in order to make an inference that can be generalized to the population.



Source Image : aprendeconalf.es

What is sampling?

- The process of selecting the representative sample units from the population to study the characteristics of the population is called sampling.
- In many empirical studies, data are to be collected from a population under study.
- A population consist number of units usually very large and sometimes infinitely large.
- In many cases, it is not practically possible to include all units of the population for the investigation.
- Therefore a few units of the population have to be selected as a representative of the whole population.
- So sampling is needed in this situation to draw the representative sample of the population.

Why do we need sampling?

Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population's characteristics by directly observing only a portion (or sample) of the population.

- Selecting a sample requires less time than selecting every item in a population
- Sample selection is a cost-efficient method
- Analysis of the sample is less cumbersome and more practical than an analysis of the entire population

Features of good sampling

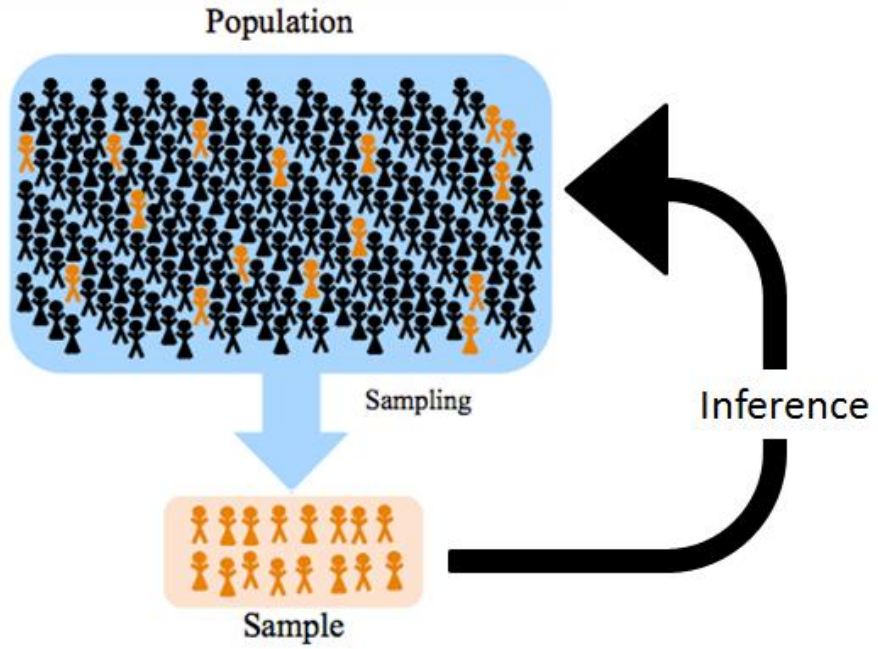
- Result in a truly representative sample.
- Sample design can have small sampling error.
- Systematic bias can be controlled.
- Results of the sample study can be applied, in general, for the universe with a reasonable level of confidence.

STATISTICS FOR DATA SCIENCE

Population

Whom do you want to generalize your results?

- Students aged 20 to 25 years
- Men aged 40 to 50 years
- All Five Star Hotels
- All students of a university
- All customers of a Restaurant



Source Image: onlinecourses.science.psu.edu

Types of population

- Tangible or Concrete Population
- Conceptual Population

STATISTICS FOR DATA SCIENCE

Tangible population

Populations where the **members are physical objects,** such as persons, calculators, cars, apples, bolts etc. are called **Tangible** or **Concrete** populations.

Such populations are assumed to be **always finite** and therefore involves **counting**.

Source: https://www.hindivarta.com/jansankhya-ki-samasya-aur-samadhan-par-nibandh/

TextBook-Chapter1.1-Pg.No-6

STATISTICS FOR DATA SCIENCE

Tangible population

Examples:

Population of people with brown eye.

Collection of laptops(to check defective or not).

Shipment of calculators(to check defective or not).

Conceptual population

Populations that do **not consists of physical or actual** are **objects** called **Conceptual populations**.

A conceptual population is a population that consists of a **not well-defined group** of which **all elements are not available** at the time the sample is collected(because the population increases every day).

The size of a conceptual population is **usually large.**

Conceptual populations are mostly the **result of a measurement.**

Conceptual population

Examples:

The population that consists of all the readings that a scale can produce – collection of lengths of nails, collection of weights of items.

Geologist weighs a rock several times on a sensitive scale.

The population of patients who take aspirin to reduce blood clotting.

Often the result of an experiment.

Corn yield after applying fertilizer.

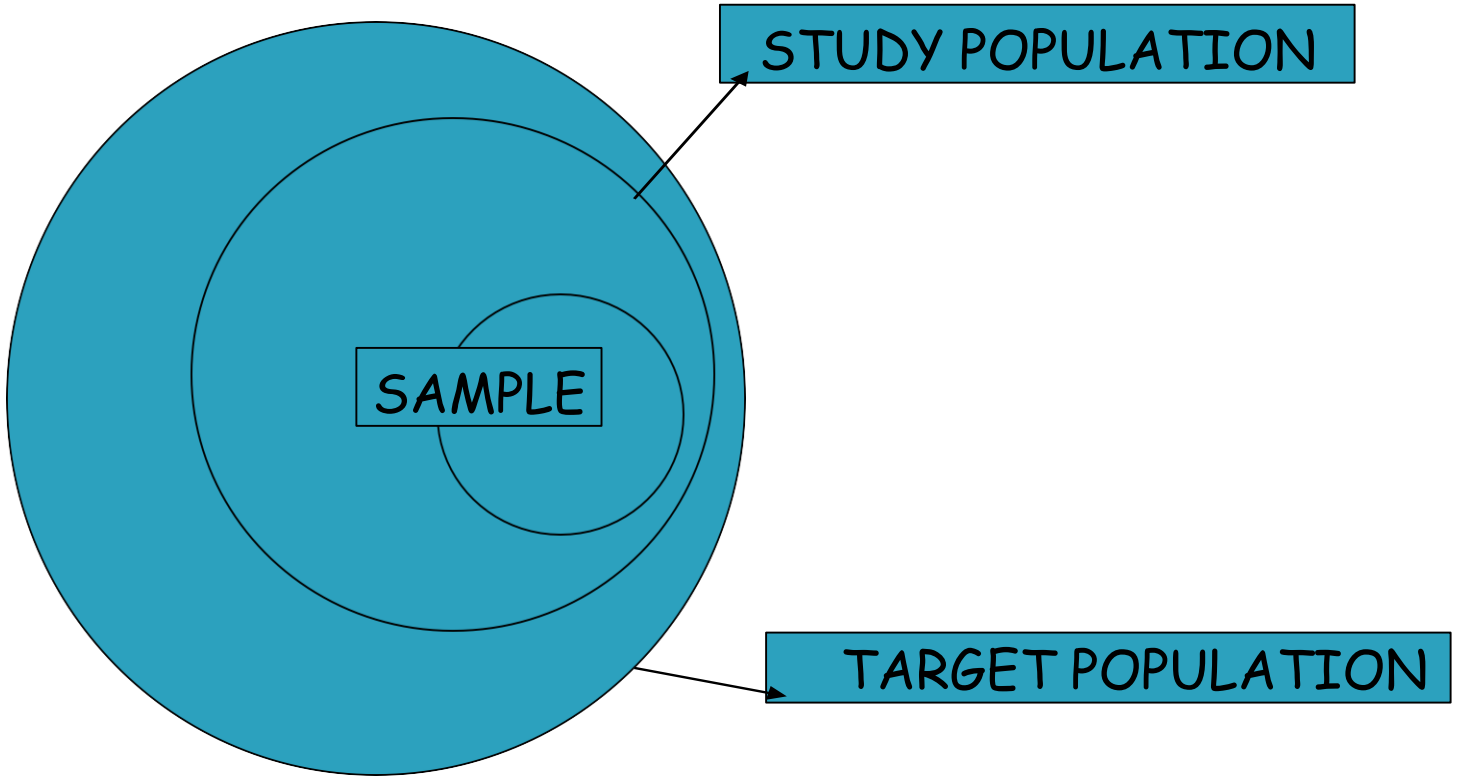
Corrosion level after applying a protective coating.

STATISTICS FOR DATA SCIENCE

Target and study population

Target or **Theoretical population** refers to the entire group of individuals or objects to which researchers are interested in generalizing the conclusions.

The **accessible population** is the population in research to which the researchers can apply their conclusions.

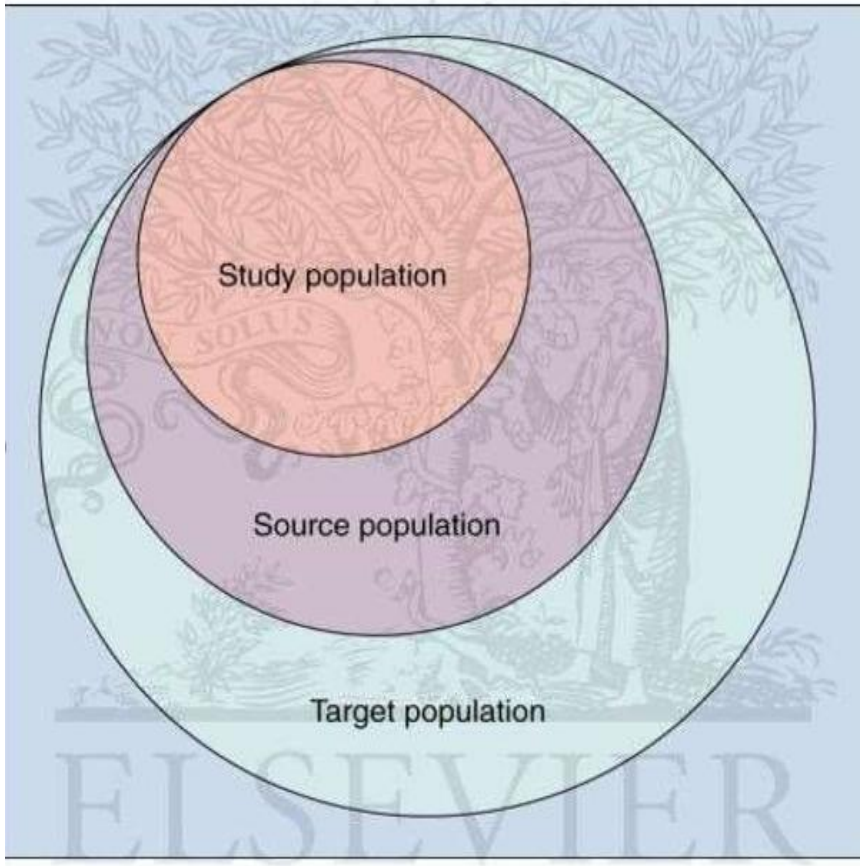


Source Image : www.slideshare.net

STATISTICS FOR DATA SCIENCE

Target population

A set of elements larger than or different from the population sampled and to which the researcher would like to generalize study findings.



Source Image : www.Elsevier.org

STATISTICS FOR DATA SCIENCE

Terminologies in sampling

Study : Find the mean weight of all students of all universities in India.

Whom to you want to generalize results?
All universities in India

Theoretical Population

What population can you get access to?
All universities in Karnataka

Study Population

How can you get access to them?
List of Universities in Karnataka

Sampling Frame

Who is in your study?
Two Universities from Karnataka

Sample

Source Image : Sunil Kumar

Terminologies in sampling

Target or Theoretical Population: The population to which the investigator wants to generalize his results.

Sampling Frame : The sampling frame is the list from which the potential respondents are drawn.

List of Universities, List of Students, List of Airline Companies, Telephone Directory

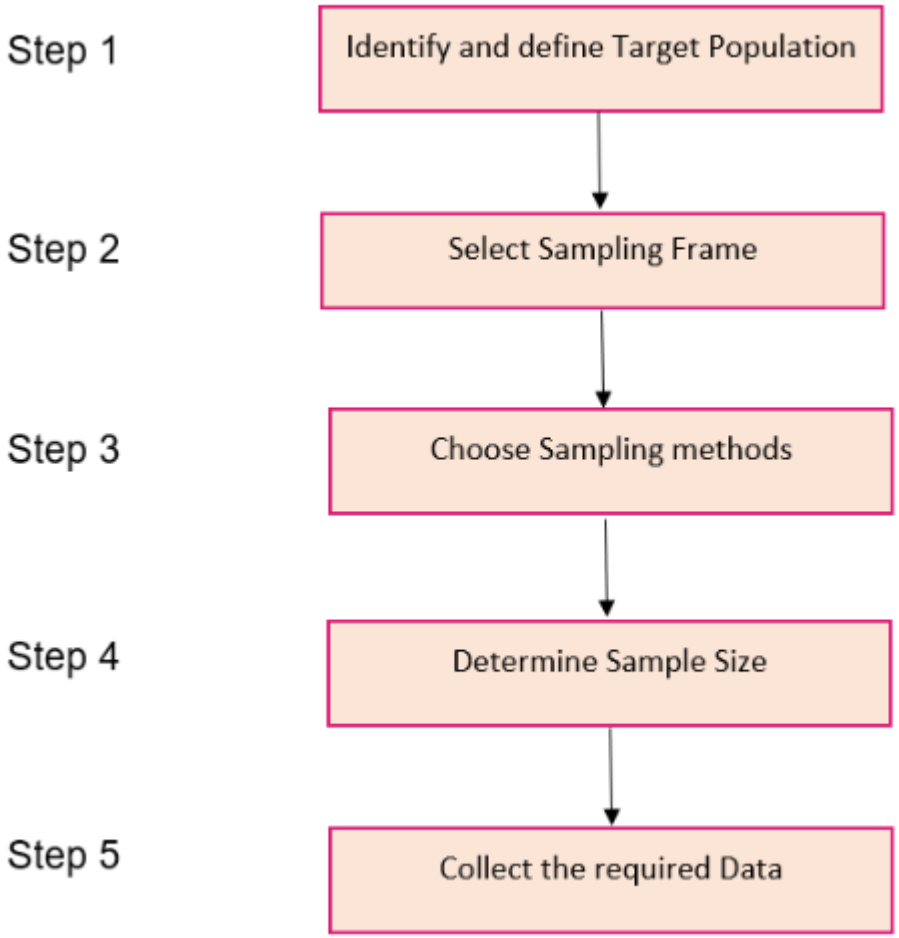
Sampling Unit : Smallest Unit from which sample can be selected.

Sampling Scheme: Method of selecting sampling units from sampling frame.

Sample: All selected respondent are sample.

STATISTICS FOR DATA SCIENCE

Steps in sampling



Source Image: https://www.analyticsvidhya.com

Steps in sampling

Step 1: Identify and Define the population:

- Population must be defined in terms of elements, sampling units, extent and time.
- Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

Step 2: Select sampling frame:

- A decision has to be taken concerning a sample unit before selecting sample.
- Sampling unit can be geographic unit such as state, district, or Construction unit such as flat, house, or social unit.
- The list of sampling unit is called as Sampling Frame

Steps in Sampling

Step 3: Choose Sampling methods:

There are several methods of selecting population units to be included in the sample. Broadly they are classified as.

- **Probability sampling/Random sampling:** Under this method, each unit of the population has the certain probability of being included in the sample.
- **Non -probability sampling/ non-random sampling:** Under this method, there is no pre-assigned probability of selection of sample units to be included in the sample.

Steps in Sampling

Step 4: Determine Sample Size:

- This refers to the number of units/items to be selected from the universe to constitute a sample.
- The Size of sample should be neither too large / too small
- An optimum sample is one which satisfies the requirements of efficiency, flexibility and reliability.

Step 5: Data Collection:

- No irrelevant information should be collected and no essential information should be discarded.

STATISTICS FOR DATA SCIENCE

Example- steps in sampling

Sampling Example :Click on this link