STATISTICS FOR DATA SCIENCE

Sampling

D. Uma
Department of Computer Science and Engineering
umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

Sampling

D. Uma
Department of Computer Science and Engineering

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

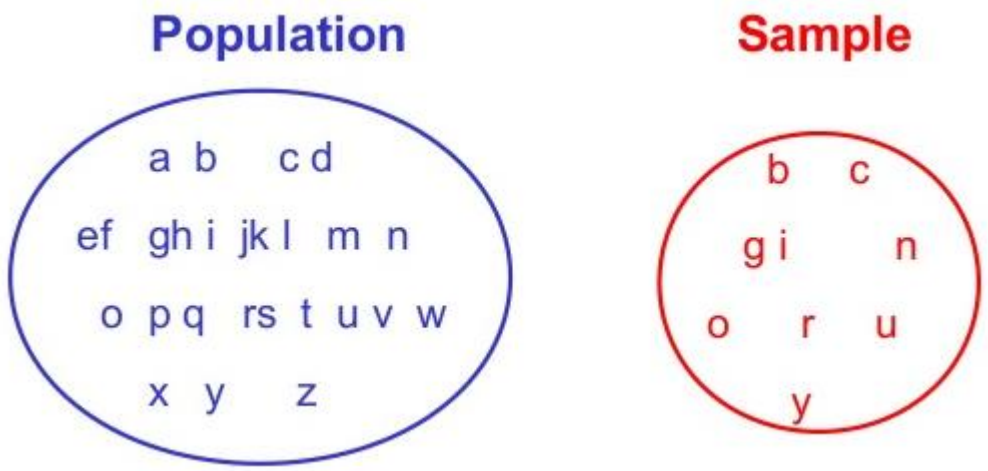
Statistical analysis is the **science of collecting data** and **uncovering patterns and trends**.

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.

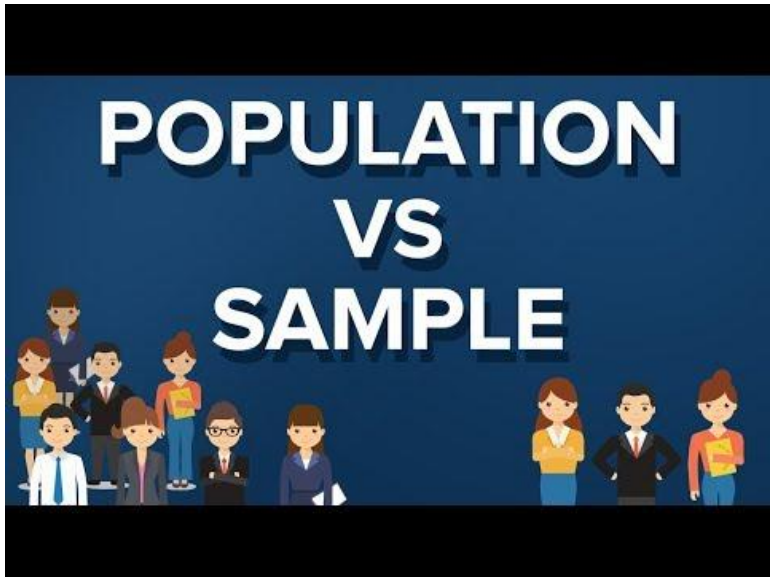


Source: www.slideshare.net,linguistics.byu.edu

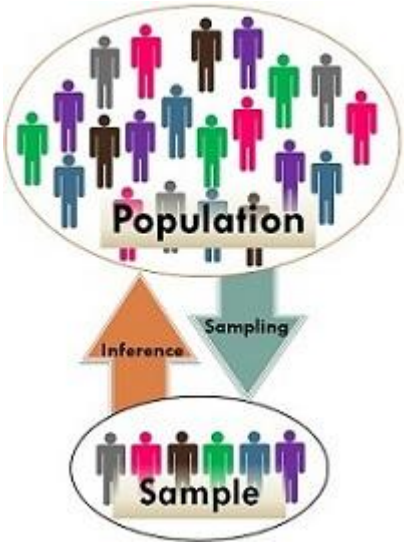
STATISTICS FOR DATA SCIENCE

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

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.

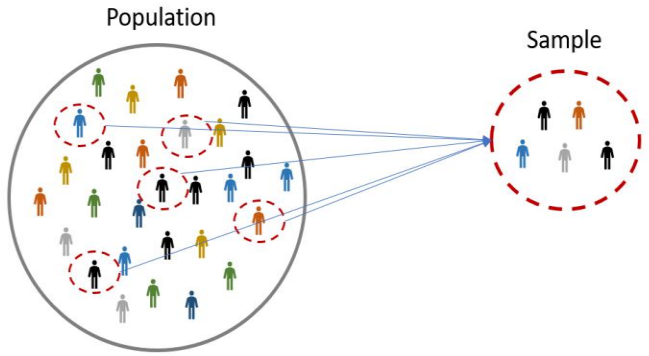
Source: www.slideshare.net, keydifferences.com, , www.2ezwebdesign.com

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?

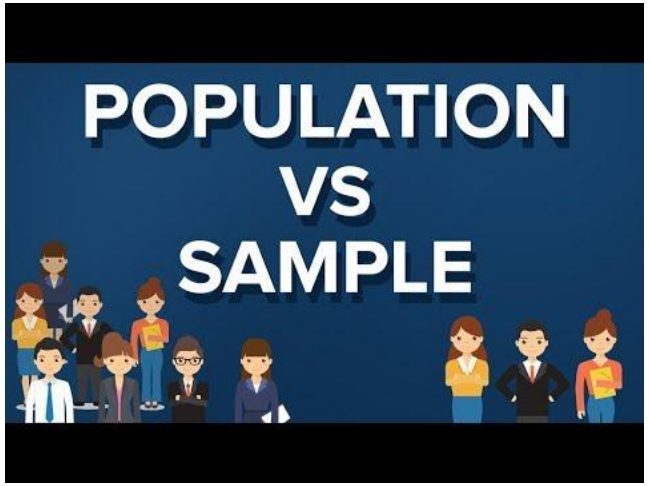
- Is this a representative of population?
- Is this a random sample?

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

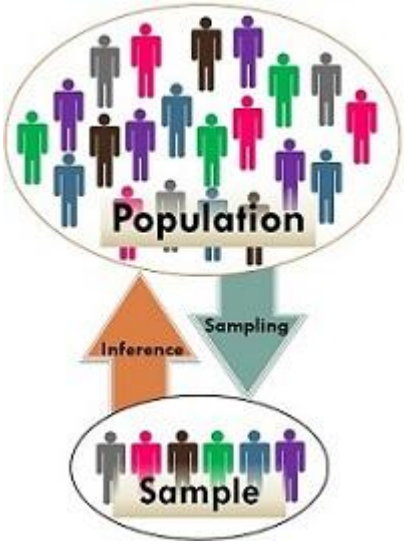
STATISTICS FOR DATA SCIENCE

Population vs. Sample

A **population** is the entire collection of objects or outcomes about which information is sought.



A **sample** is a subset of a population, containing the objects or outcomes that are actually observed.

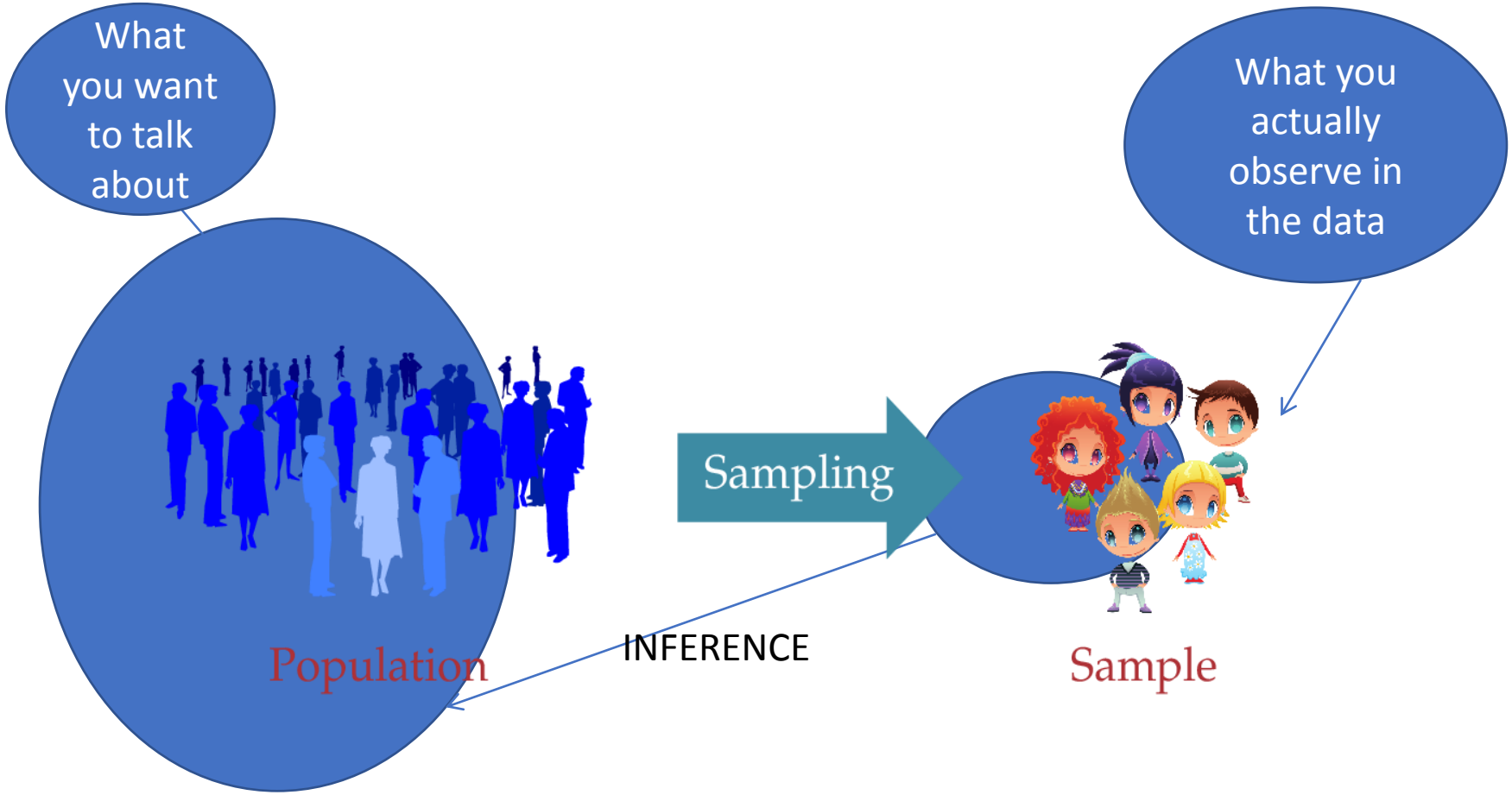


Source Image : 365datascience.com

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.



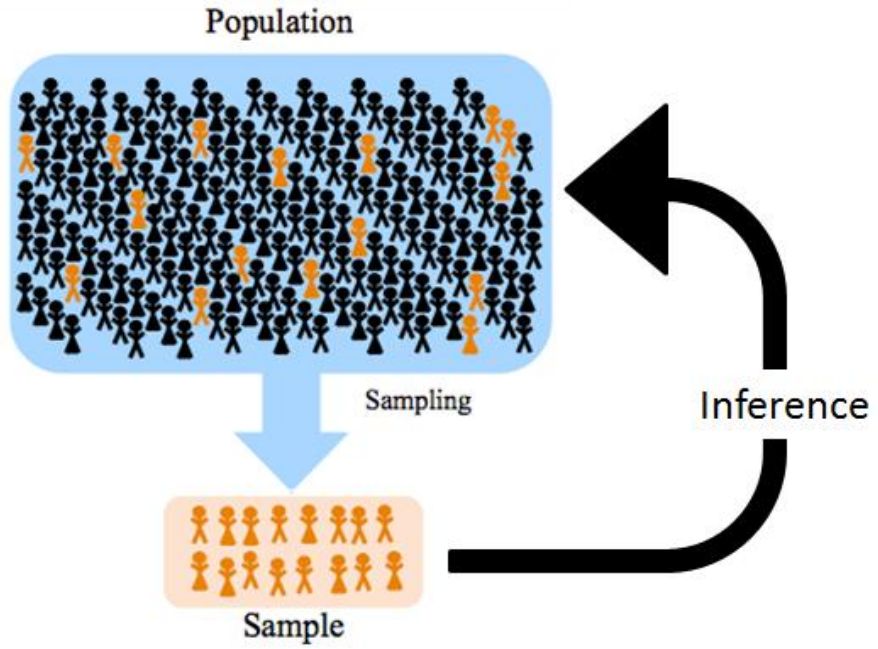
Source Image : aprendeconalf.es

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

STATISTICS FOR DATA SCIENCE

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**.

Examples:

Population of people with brown eye.

Collection of laptops(to check defective or not).

Shipment of calculators(to check defective or not).

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

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.

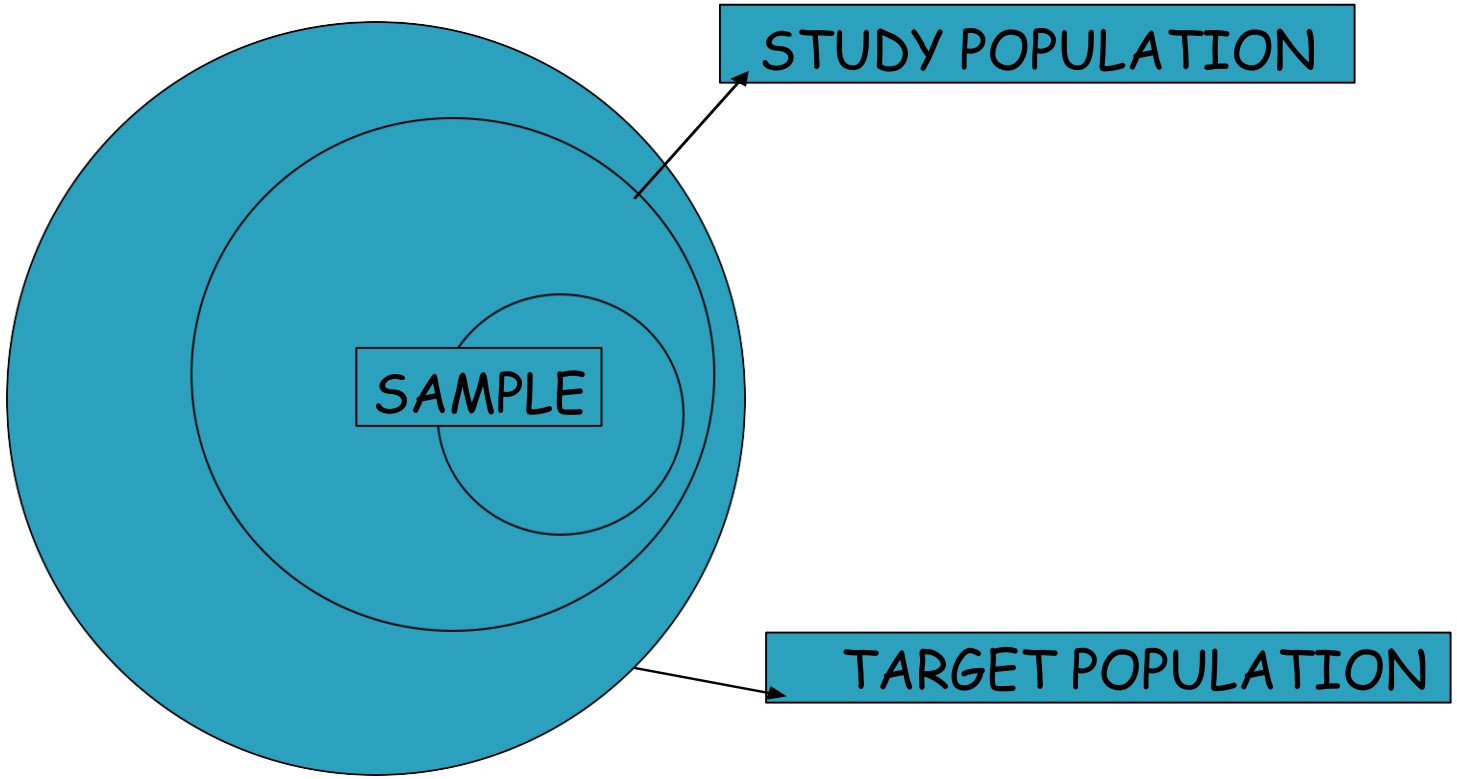
Corn yield after applying fertilizer.

Corrosion level after applying a protective coating.

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.



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

STATISTICS FOR DATA SCIENCE

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.

Identify the population and sample in this setting.

a. The population is all high school students in Karnataka; the sample is all of the seniors at PES school.

b. The population is all students at PES school; the sample is all of the seniors at PES school.

c. The population is all seniors at PES school; the sample is the 50 seniors surveyed.

Do it yourself !!!!!!!!

Identify the population and sample in this setting.

a. The population is every car at the factory; the sample is the 1 car he is curious about.

b. The population is every car at the factory, the sample is the 30 selected points.

c. The population is every possible point on the car; the sample is the selected points.

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335