



# STATISTICS FOR DATA SCIENCE

## Sampling Methods

---

**Prof. Uma D**

**Prof. Suganthi S**

**Prof. Silviya Nancy J**

Department of Computer Science and Engineering

# STATISTICS FOR DATA SCIENCE

---

## Sampling Methods

**Prof. Uma D**

**Prof. Suganthi S**

**Prof. Silviya Nancy J**

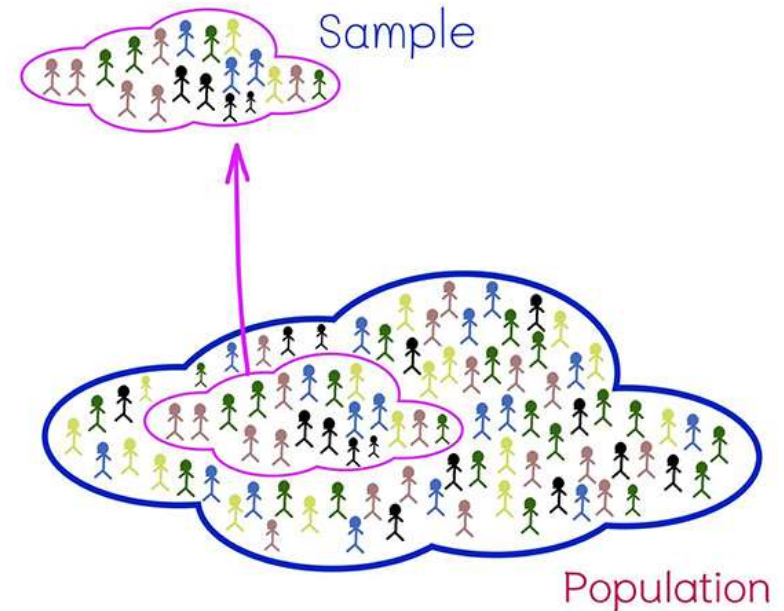
# STATISTICS FOR DATA SCIENCE

## What are sampling methods?

In a statistical study, sampling methods refer to **how we select members** from the population to be in the study.

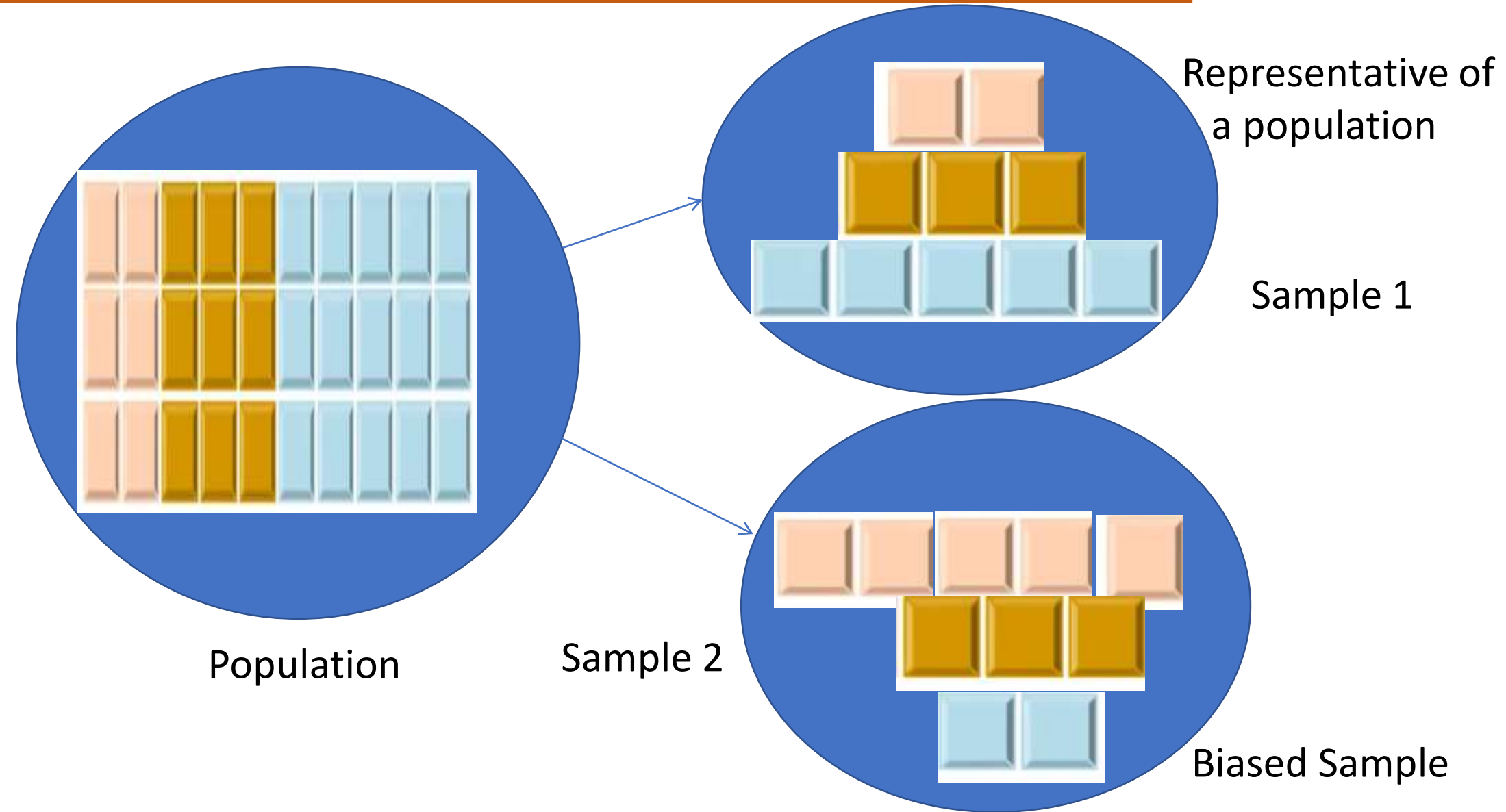
If a sample **isn't randomly selected**, it will probably be **biased in some way** and the **data may not** be **representative of the population**.

There are many ways to select a sample—some good and some bad.

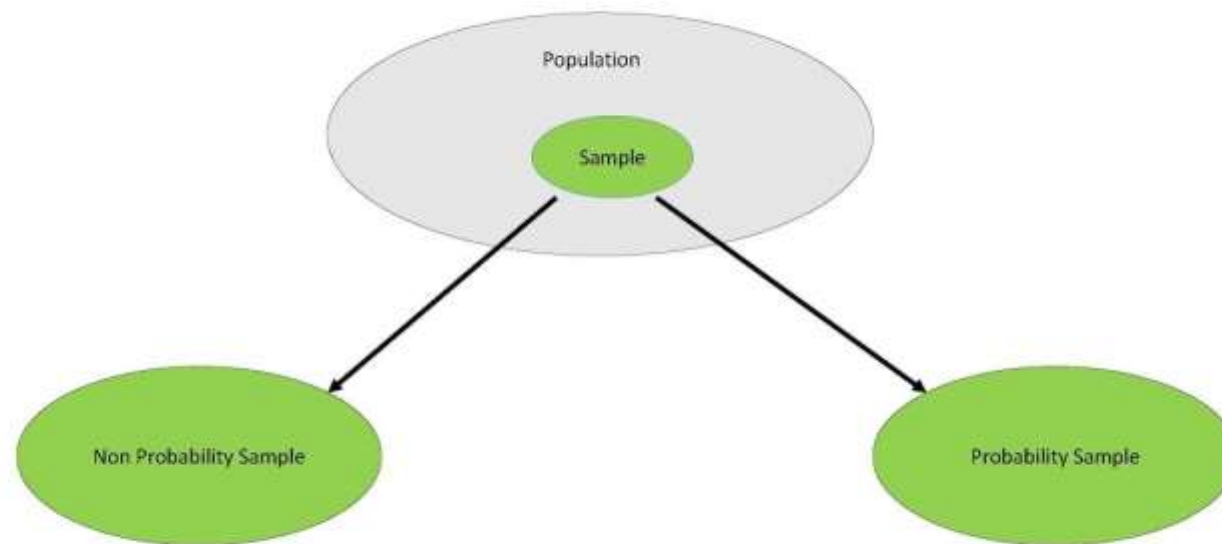


# STATISTICS FOR DATA SCIENCE

## Representative and biased samples



### Sampling Techniques

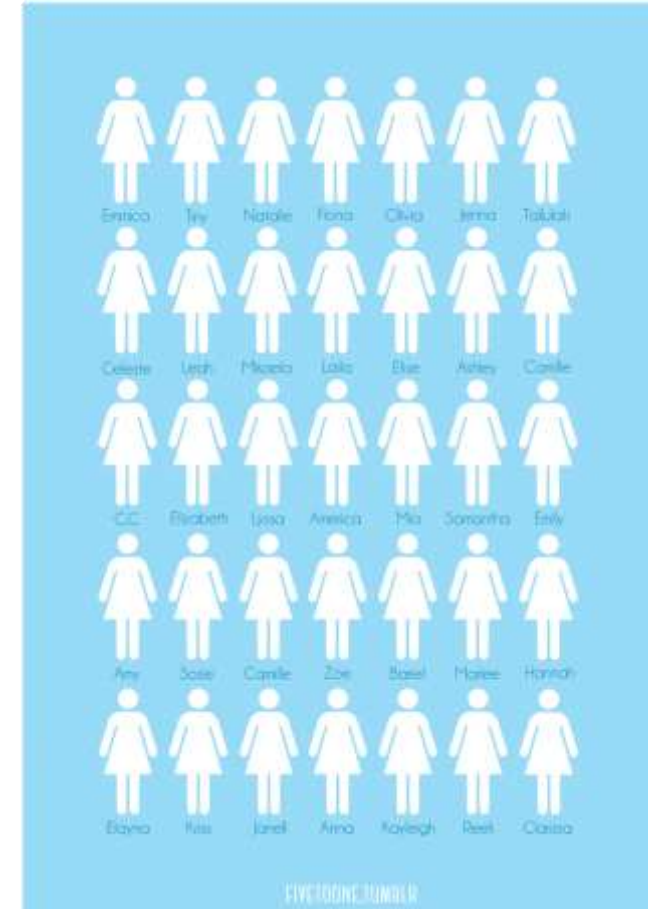


Every unit of the population has the same probability of being included in the sample.

A chance mechanism is used in the selection process.

Eliminates bias in the selection process.

This is also known as Random sampling.

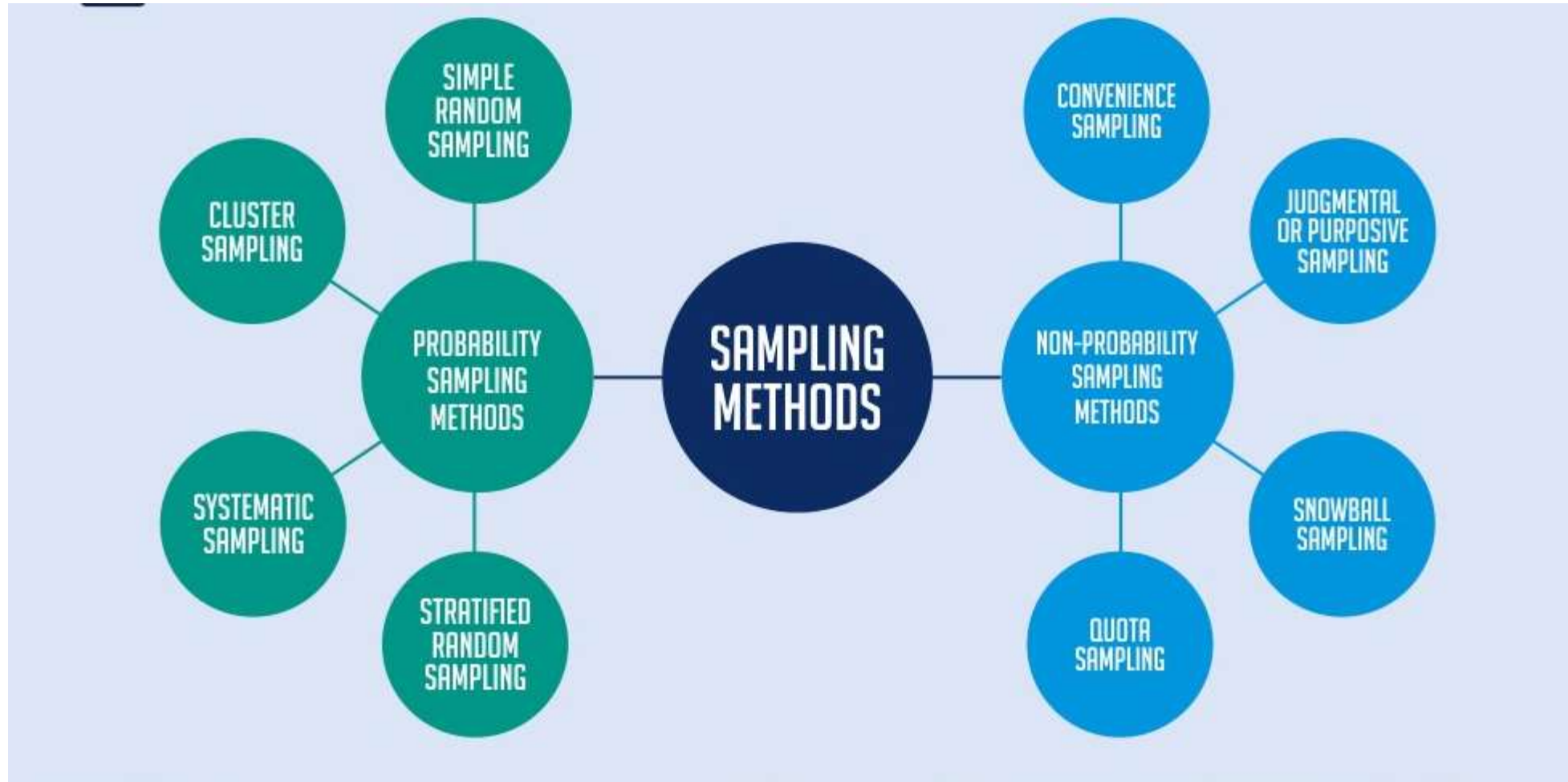


Every unit of the population **does not have the same probability of being included** in the sample.

It opens up the **selection bias**.

**Not an appropriate data collection methods** for most of the statistical analysis.

Also known as **non-random sampling**.





## What type of Sampling?

A teacher puts students' names in a hat and **chooses without looking** to get a sample of students.



**Every member has an equal chance** of being included in the sample.

**Why it's good:** Random samples are usually **fairly representative** since they **don't favor certain members**.

# STATISTICS FOR DATA SCIENCE

## Simple Random Sampling

**Purpose :** Random and Representative Sample from a Larger Population.

**When to Use :** Best to use when population is small and produces better representative.

**Key Thing:** All members of a population has an equal probability.

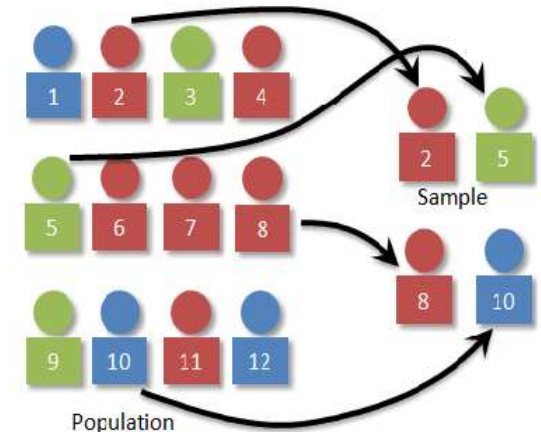
**How :** Assign numbers to members of population & select randomly.

**Small Population:** Manual Lottery Method.

**Larger Population :** System Generated Number.

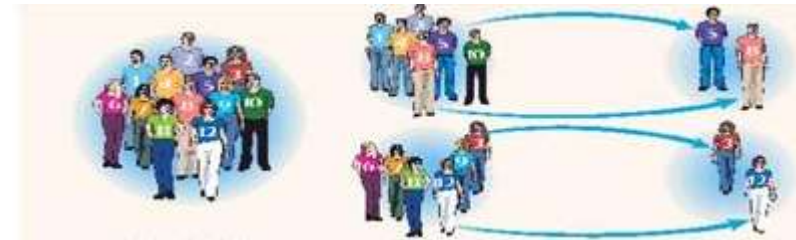
**Advantages:** Easier, Better representativeness, Low sampling error, No prior information is required.

**Disadvantages:** Biased sample, Does not reflect proportionate representation, Difficult when population is larger.



## What type of Sampling?

A student council surveys 50 students by getting random samples of 25 juniors and 25 seniors.



**Why it's good:** This sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

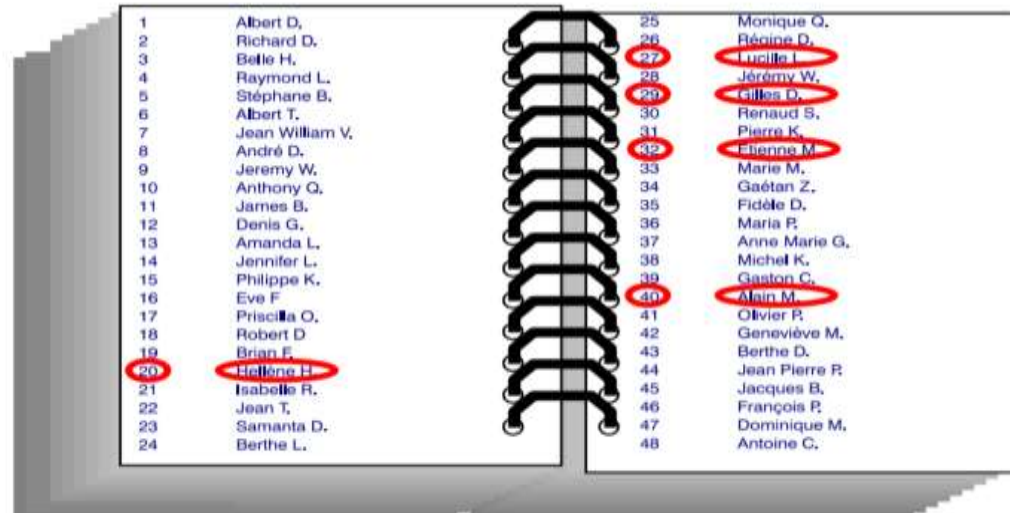
# STATISTICS FOR DATA SCIENCE

## Simple Random Sampling

Definition: Text-Book Reference-Section 1.1,Page-No:3

Example:

- All subsets of the frame are given an equal probability.
- Random number generators



1	Albert D.	25	Monique Q.
2	Richard D.	26	Réjane D.
3	Belle H.	27	Lucille I.
4	Raymond L.	28	Jérémy W.
5	Stéphane B.	29	Gilles D.
6	Albert T.	30	Renaud S.
7	Jean William V.	31	Pierre K.
8	André D.	32	Étienne M.
9	Jeremy W.	33	Marie M.
10	Anthony Q.	34	Gaëtan Z.
11	James B.	35	Fidèle D.
12	Denis G.	36	Maria P.
13	Amanda L.	37	Anne Marie G.
14	Jennifer L.	38	Michel K.
15	Philippe K.	39	Gaston C.
16	Eve F.	40	Alain M.
17	Priscilla O.	41	Olivier P.
18	Robert D.	42	Geneviève M.
19	Brian F.	43	Berthe D.
20	Hélène H.	44	Jean Pierre R.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.

# STATISTICS FOR DATA SCIENCE

## Simple Random Sampling

---



Example Problems: Text-Book Reference-Section 1.1

- Example 1.1,Page-No:4
- Example 1.7,Page-No:8

# STATISTICS FOR DATA SCIENCE

## Stratified Random Sampling

**Purpose :** Unbiased Random Sample from a Larger Population.

**When to Use :** Population proportion should be reflected in sample.

**Key Thing:** Sample proportion is same as Population proportion,  
Strata is homogeneous.

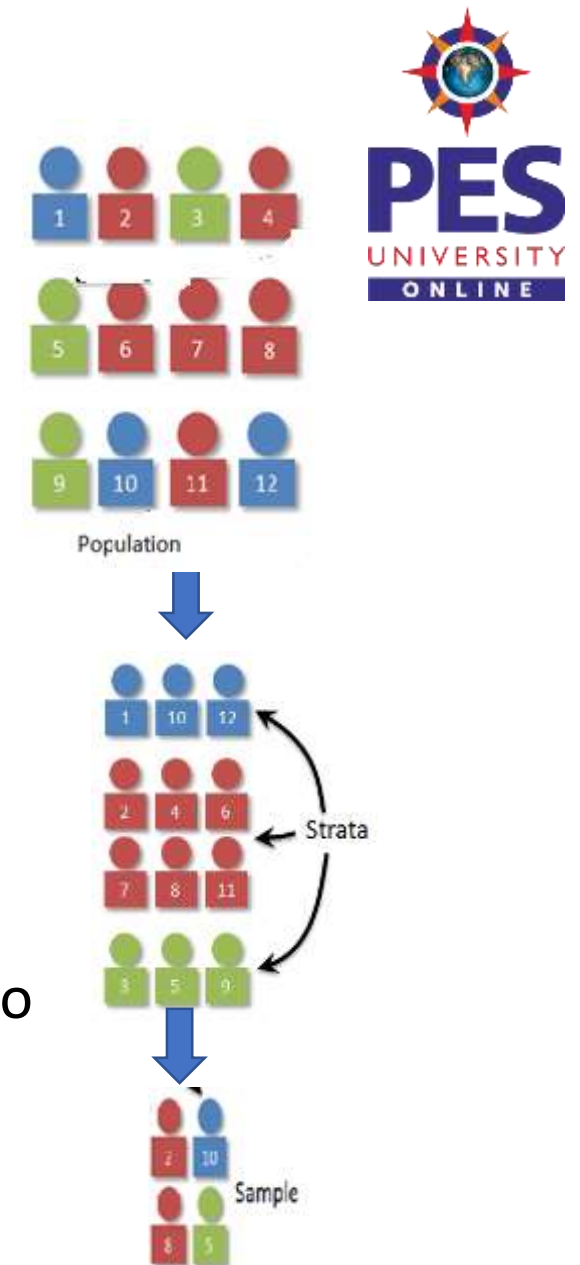
**How :** Divide the population into Strata or Groups.

**Criteria :** Gender, Hair Color, Eye Color, Salary, Designation, Age etc.

**Selection :** Simple Random Sampling approach to Strata.

**Advantages:** Enhancement of representativeness of a sample, Easy to carry out, Higher statistical efficiency.

**Disadvantages:** Classification error, time consuming, expensive.



## What type of Sampling?

Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 4<sup>th</sup> student is selected to take a survey.



**Why it's good:** This sample includes every 4<sup>th</sup> person. But, first person is chosen randomly.

# STATISTICS FOR DATA SCIENCE

## Examples



	Example 1	Example 2	Example 3
Population	All people in US	All PSU intercollegiate athletes	All elementary students in the local school district
Groups (Strata)	4 Time Zones in the U.S. (Eastern, Central, Mountain, Pacific)	26 PSU intercollegiate teams	11 different elementary schools in the local school district
Obtain a Simple Random Sample	500 people from each of the 4 time zones	5 athletes from each of the 26 PSU teams	20 students from each of the 11 elementary schools
Sample	$4 \times 500 = 2000$ selected people	$26 \times 5 = 130$ selected athletes	$11 \times 20 = 220$ selected students



# STATISTICS FOR DATA SCIENCE

## Systematic Sampling

**When to Use :** When project budget is tight and less time to complete.

**Key Thing:** Find the kth value to select every kth member.

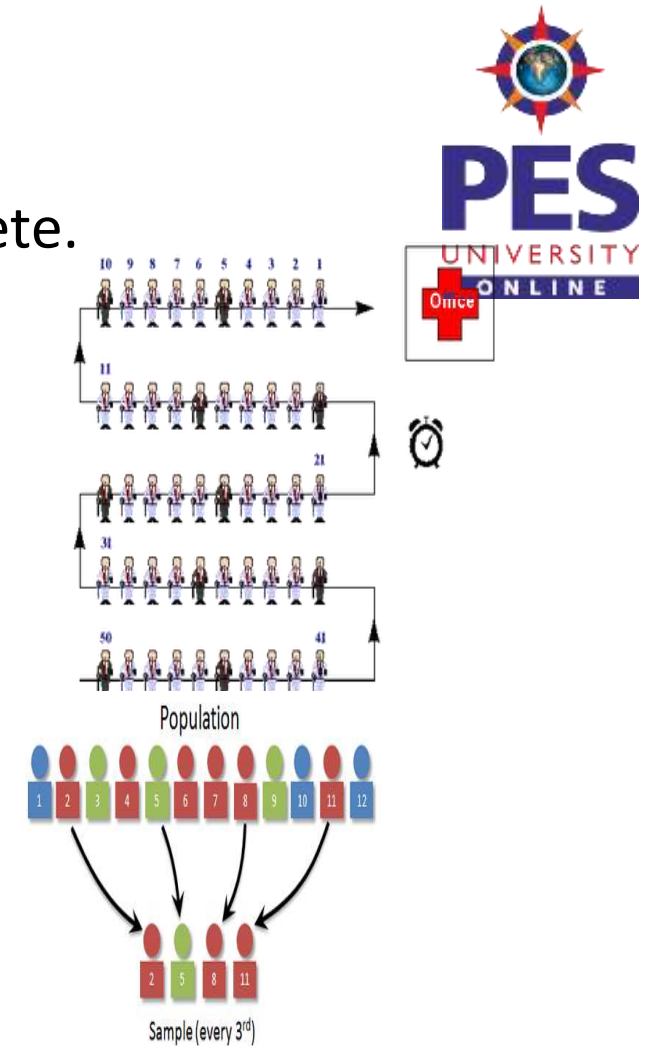
$$k = N / n$$

**How:** Assign numbers to each population member.

**Selection :** Randomly select first person and then select every kth person.

**Advantages:** Easy to select, Sample evenly spread over entire reference population, cost effective.

**Disadvantages:** Sample may be biased, Each element does not have equal chance, Ignorance of all elements between two kth element.



## What type of Sampling?

An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.



**Why it's good:** This sample gets every member from some of the groups, so it's good when each group reflects the population as a whole.

- Assume that in a population of 10,000 people, a statistician selects **every 100th person for sampling**.
- If a local NGO is seeking to form a systematic sample of 500 volunteers from a population of 5000, they can select **every 10th person in the population** to build a sample systematically.

# STATISTICS FOR DATA SCIENCE

## Cluster Sampling

**When to Use :** When population is already broken up into groups(clusters).

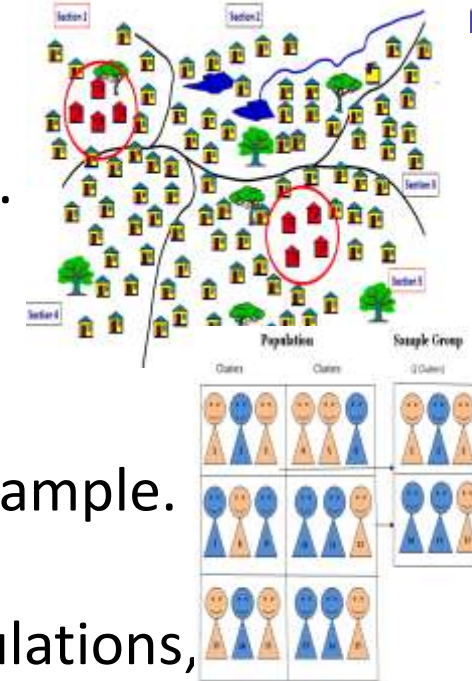
**Key Thing:** Heterogeneous members in each group.

**How:** Population is divided into non-overlapping areas(clusters).  
Each cluster is a miniature or microcosm of a population.

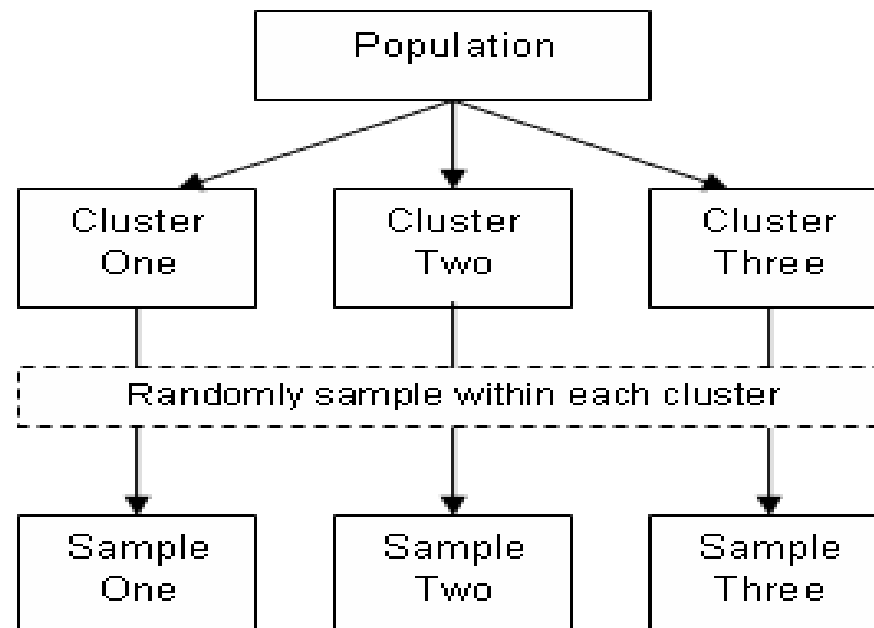
**Selection :** Clusters are selected randomly and all elements are included or elements are chosen using simple random sample.

**Advantages:** More convenient for geographically dispersed populations,  
Less travel cost, Simplified administration of the survey.

**Disadvantages:** Statistically less efficient, Sampling error is higher,  
problems are higher than simple random sampling.



The population is divided into subgroups (clusters) like families. A simple random sample is taken of the subgroups and then all members of the cluster selected are surveyed.



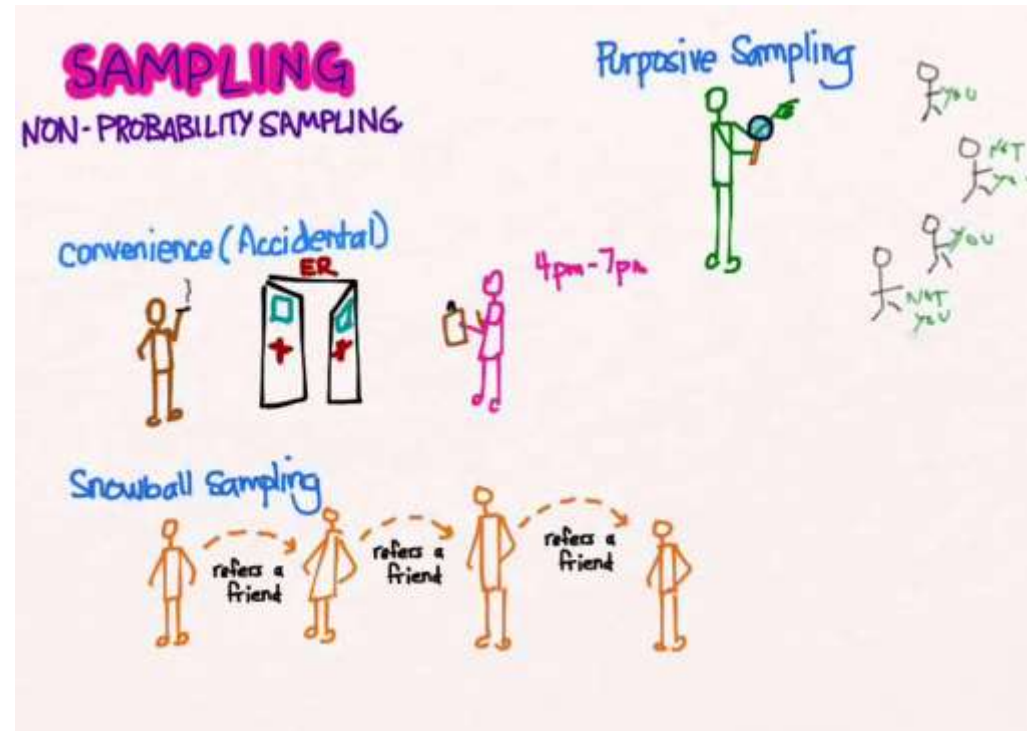
- Consider a scenario where an organization is looking to survey the **performance of smartphones across India**.
- **Step 1:** They can divide the entire country's population into cities (clusters) and select further towns with the highest population and also filter those using mobile devices.
- **Step 2: Single-stage cluster sampling** – An NGO wants to create a sample of girls across five neighboring towns to provide education. Using single-stage sampling, the NGO randomly selects towns (clusters) to form a sample and extend help to the girls deprived of education in those towns.

- **Step 3: Two-stage cluster sampling** – A business owner wants to explore the performance of his/her plants that are spread across various parts of the U.S. The owner creates clusters of the plants. He/she then selects random samples from these clusters to conduct research.
  - **Step 4: Multiple-stage cluster sampling** - For conducting effective research across multiple geographies, one needs to form complicated clusters that can be achieved only using the multiple-stage sampling technique.
- An organization intends to survey to analyze the performance of smartphones across India. They can divide the entire country's population into cities (clusters) and select cities with the highest population and also filter those using mobile devices.

# STATISTICS FOR DATA SCIENCE

## Non-random Sampling Methods

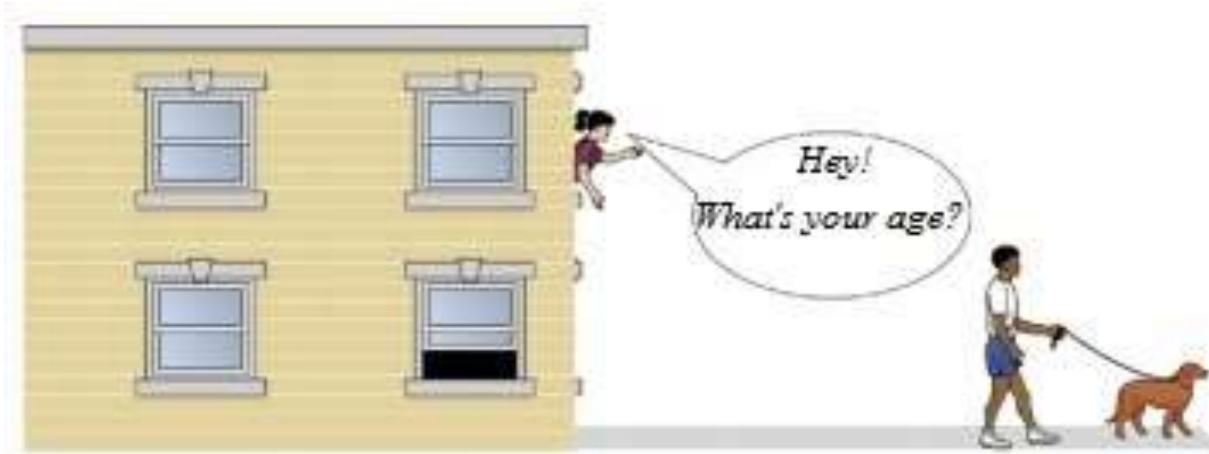
- ☛ Convenience
- ☛ Judgmental/Purposive
- ☛ Quota
- ☛ Snowball





## What type of Sampling?

A researcher polls people as they walk by on the street.



**Bad ways to sample:** The researcher chooses a sample that is readily available in some non-random way.

**Why it's probably biased:** The location and time of day and other factors may produce a biased sample of people.

**When to Use :** When population is not clearly defined or sampling unit is not clear or complete source list is not available.

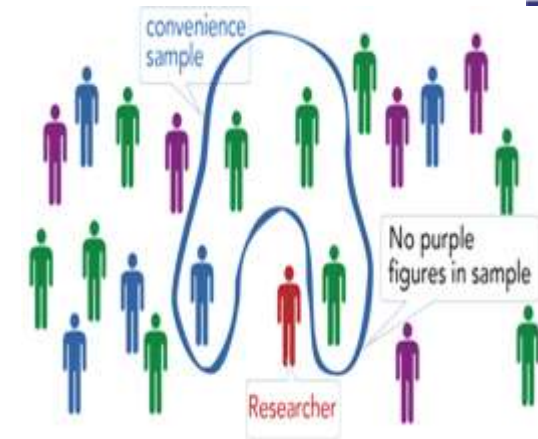
**Key Thing:** Subjects for a study are easily available within the proximity of the researcher.

**How:** It is done at the “convenience” of the researcher.

**Selection :** Whichever individuals are easiest to reach and they are convenient.

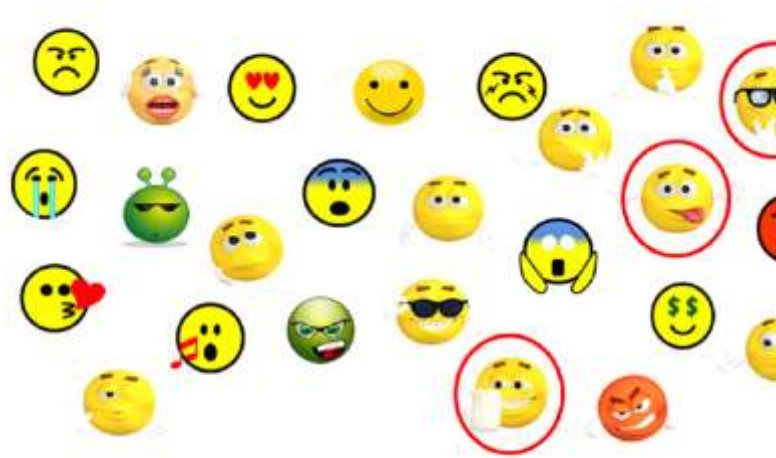
**Advantages:** Ease of availability, saves time, money, Useful in pilot study.

**Disadvantages:** Biased sample, Sampling errors, Results can't be generalized.



## What type of Sampling?

People who own the qualities expected by the researcher.



The researcher uses their own knowledge and judgement to include or exclude people in the sample.

- A marketing student needs to get feedback on the “scope of content marketing in 2020.”
- The student may quickly **create an online survey**, send a link to all the contacts on your phone, share a link on social media, and talk to people you meet daily, face-to-face.
- If an interviewer needs to **conduct a survey at a shopping center** early in the morning on a given day, the people that he could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area.

# STATISTICS FOR DATA SCIENCE

## Judgmental Sampling

**When to Use :** The sample is selected based upon judgment.  
Also, the researcher must be confident that the chosen sample is truly representative of the entire population.

**Key Thing:** The researcher selects a sample based on experience or knowledge of the group to be sampled.

**How:** Basis of the researcher's knowledge and judgment.

**Selection :** People who own the qualities expected by the researcher.

**Advantages:** Consumes minimum time, real-time results.

**Disadvantages:** Selection bias, Selection of proper sample size.



# STATISTICS FOR DATA SCIENCE

## What type of Sampling?

An interviewer may be told to **sample 200 females and 300 males** between the **age of 45 and 50**.



Sample the people **till the quota is met**.

- Consider a scenario where a panel decides to understand what are the factors which lead a person to **select ethical hacking as a profession**.
- Ethical hacking is a skill which has been recently attracting youth. More and more people are selecting it as a profession.
- The researchers who understand what ethical hacking is will be able to decide who should form the sample to learn about it as a profession.
- That is when **judgmental sampling** is implemented. Researchers can easily filter out those participants who can be eligible to be a part of the research sample.



# STATISTICS FOR DATA SCIENCE

## Quota Sampling

**When to Use :** If a study aims to investigate a trait or a characteristic of a certain subgroup, this type of sampling is the ideal technique.

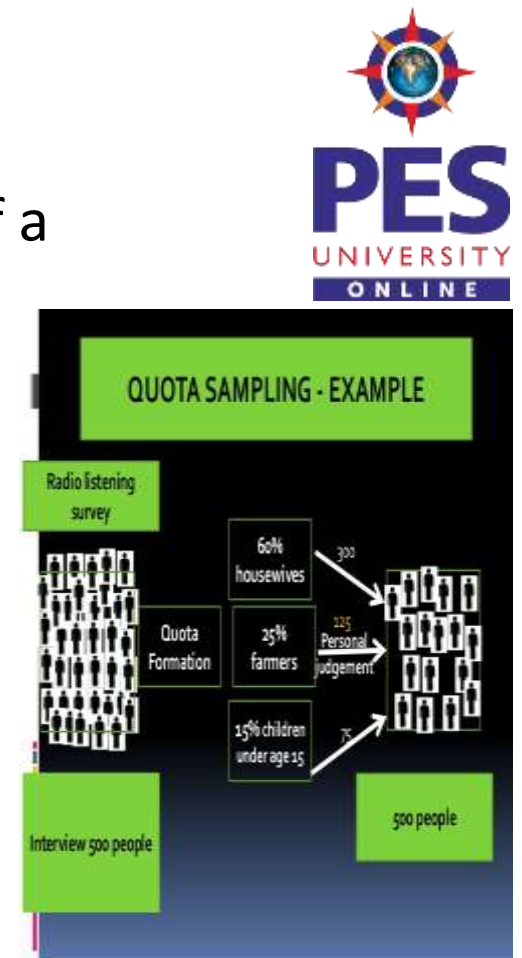
**Key Thing:** Sample elements are selected until the quota controls are satisfied.

**How:** First identify the strata and their proportions as they are represented in the population.

**Advantages:** When the respondent refuses to cooperate, he may be replaced by another person who is ready to furnish information, Less expensive, speedy, Convenience in execution.

**Disadvantages:** Bias, lack of valuable data.

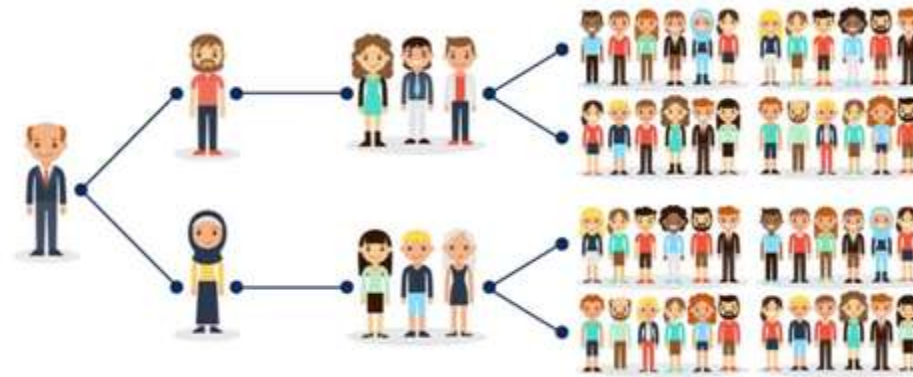
**Note:** It is the non-probability equivalent of stratified sampling.



**PES**  
UNIVERSITY  
ONLINE



Survey the people who are involved in illegal activities.



Identify an initial object.

A researcher wants to survey individuals about what smartphone brand they prefer to use. He/she considers a sample size of 500 respondents. Also, he/she is only interested in surveying ten states in the US. Here's how the researcher can divide the population by quotas:

- **Gender:** 250 males and 250 females
- **Age:** 100 respondents each between the ages of 16-20, 21-30, 31-40, 41-50 & 51+
- **Employment status:** 350 employed and 150 unemployed people.
  - (Researchers apply further nested quotas . For eg, out of the 150 unemployed people, 100 must be students.)
- **Location:** 50 responses per state

Depending on the type of research, the researcher can apply **quotas based on the sampling frame**. It is not necessary for the researcher to divide the quotas equally.

# STATISTICS FOR DATA SCIENCE

## Snowball Sampling

**When to Use :** When the desired sample characteristic is rare.

**Key Thing:** Research starts with a key person and introduce the next one to become a chain. It may be extremely difficult or cost prohibitive to locate respondents in these situations.

**How:** Identify an initial subject and ask these people to identify others.

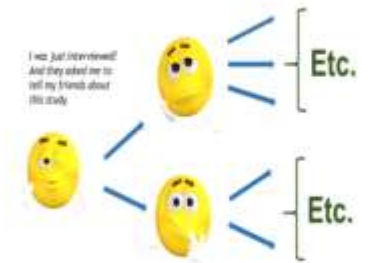
**Selection :** This technique relies on referrals from initial subjects to generate additional subjects.

**Advantage:** Lowers search cost.

**Disadvantage:** Introduces bias.



Snowball sampling



Number of referred individuals not restricted or specified. No formal documentation of link between referred and referee.

- **No official list of names of the members:** This sampling technique can be used for a population, where there is no easily available data like their [demographic](#) information. For example, homeless or list of members of an elite club, whose personal details cannot be obtained easily.
- **Difficulty to locate people:** People with rare diseases are quite difficult to locate. However, if a researcher is carrying out a [research study](#) similar in nature, finding the primary data source can be a challenge. Once he/she is identified, they usually have information about more such similar individuals.

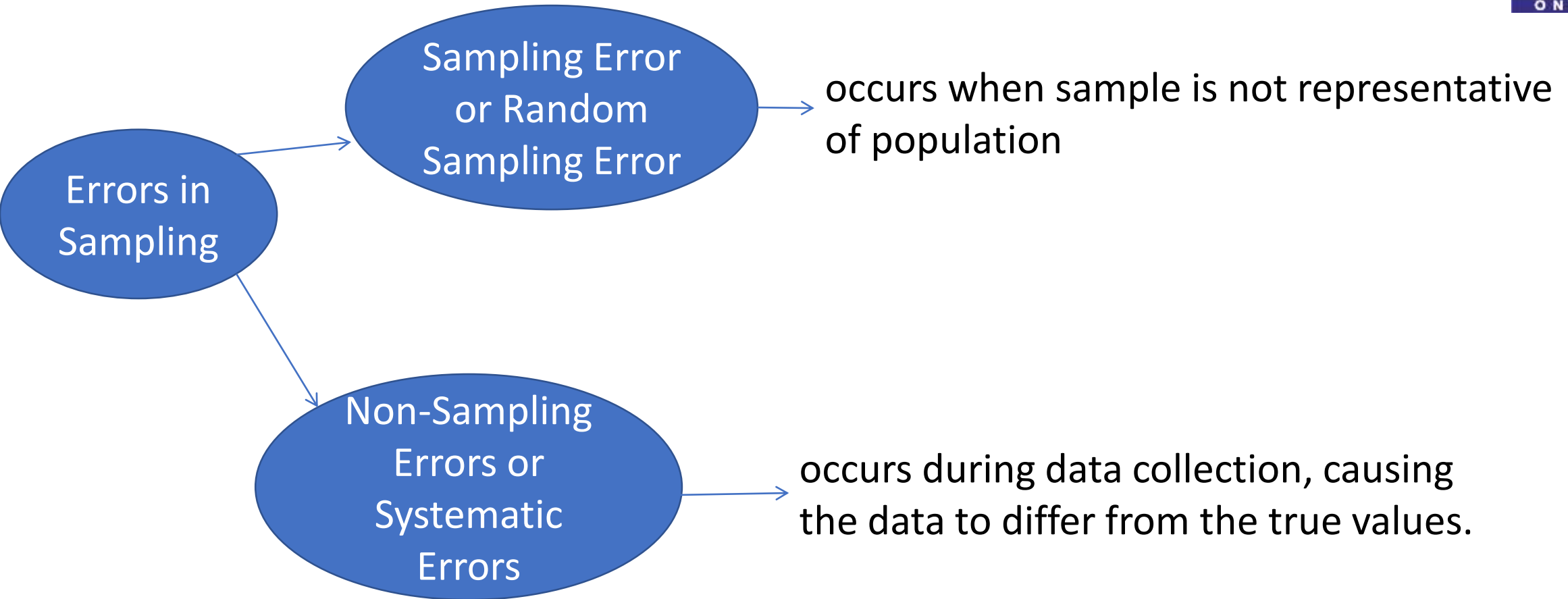
- **Sampling variance** refers to variation of a particular statistic (e.g. the mean, proportion etc) calculated in sample, if to repeat the study (sample-creation/data-collection/statistic-calculation) many times.
- **Sampling bias** occurs when a chosen sample is not representative of the larger population. It occurs due to the sampling technique/method used to perform data collection.
- It can be either **selection bias** and **nonresponsive bias**.

- A sampling method has **sampling bias** if all subjects in the population are not equally likely to be included in a sample.
- **Selection bias** is a type of sampling bias that occurs when objects are selected from the population in a non-random fashion. With selection bias, the exclusion of certain objects from possible samples affects statistical results based on those samples.
- **Nonresponse bias** is a type of sampling bias that occurs because of the absence of certain objects or subjects from a sample. For example, some subjects don't respond to surveys because they refuse, cannot be contacted, or have a lack of interest in the survey content.

### Bias in selection

#### Sources of bias

- Omitting people in hard to reach groups
- Replacing selected individuals with others, because they are in accessible
- Use of outdated list as sample frame





**Sampling error** arises due to the variability that occurs by chance because a sample is not the complete picture of the population.

- Typically, larger sample size leads to an increase in the precision of an statistic (can reduce sampling error)
- Exact measurement of sampling error is generally not feasible since the true population values are unknown
- However, sampling error can often be estimated by probabilistic modelling of the sample.

**Non-sampling errors** are the results of mistakes made in implementing data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors.

- ◆ Data from **nonrandom samples** are **not appropriate for analysis** by inferential statistical methods.
- ◆ **Sampling Error** occurs when the **sample is not representative** of the population.
- ◆ **Non-sampling Errors**
  - Missing Data, Recording, Data Entry, and Analysis Errors
  - Poorly conceived concepts , unclear definitions, and defective questionnaires
  - Response errors occur when people so not know, will not say, or overstate in their answers



# THANK YOU

---

**Prof. Uma D**

**Prof. Suganthi S**

**Prof. Silviya Nancy J**

Department of Computer Science and Engineering