Sampling

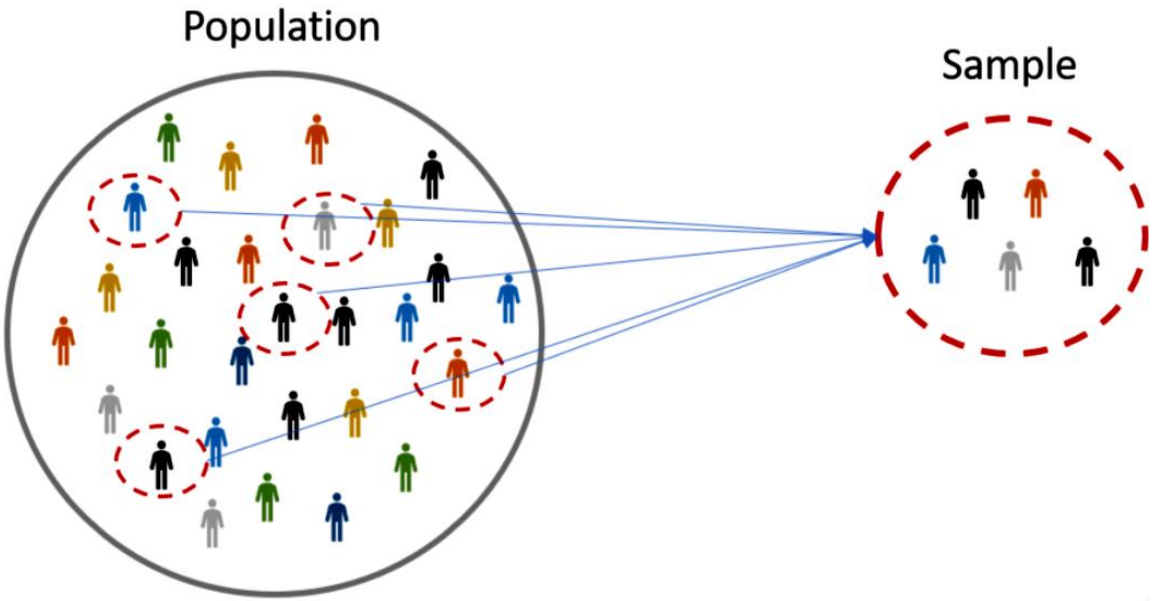
(Data Mining & Data Warehousing, IT-3031)

Dr. Hrudaya Kumar Tripathy
and
Dr. Ajay Kumar Jena

School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar

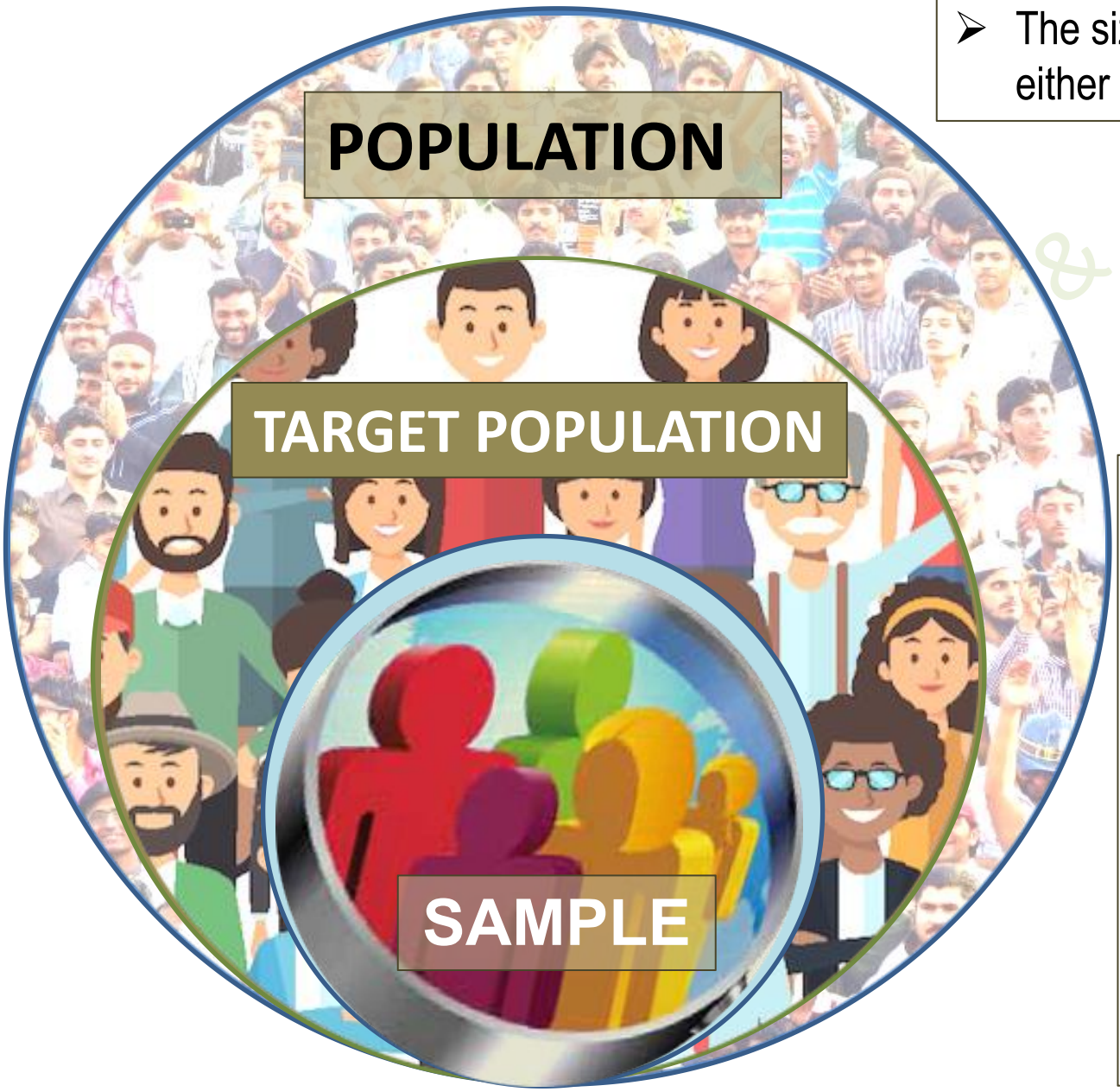
Sampling

- Sampling is a technique of selecting individual members or a subset of the population to make a statistical conclusion in the basis of evidence from them and estimate characteristics of the whole population.
- A sample is a “subgroup of a population”.
- As a way of obtaining a group of people or objects to study that were representative of a large population or universe of interest. (Stacks & Hocking, 1999)



Sampling

- A Population is the set of all possible states of a random variable.
- The size of the population may be either infinite or finite.



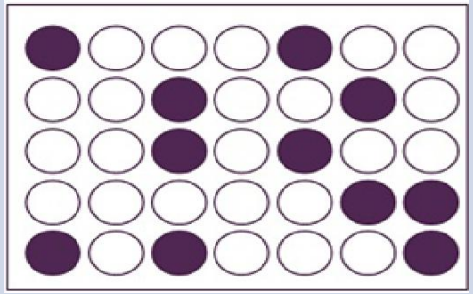
Selection of individual observations

Inference about the population

- A Sample is a descriptive measure & subset of the population.
- Its size is always finite.
- The sample may not be representative for the population.
- A representative sample shows the same properties as the population.

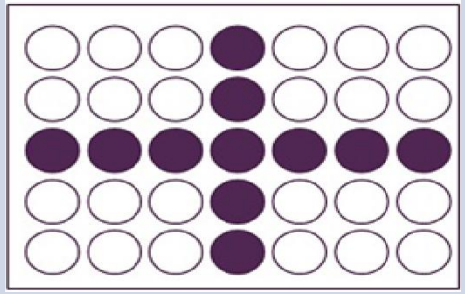
Sampling Methods

Probability Sampling



- Any element can be chosen **randomly** from the population. It deals with choosing the sample randomly.
- The most critical requirement of probability sampling is that everyone in your population has a known and equal chance of getting selected.
- Ex. When an unbiased coin is thrown (randomly), the probability of getting the head is ½.
- Ex. Probability of getting a number i.e 6 when a dice will be thrown.

Non-Probability Sampling



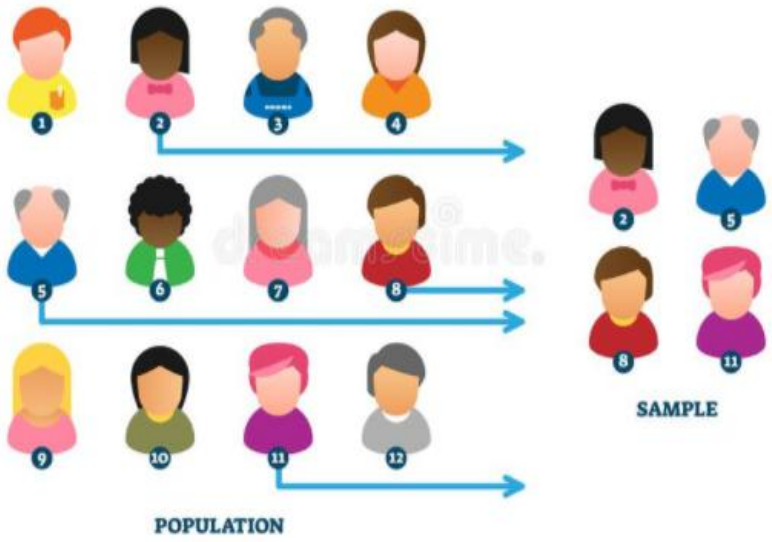
- Every element will be chosen on the subjective judgment (purposefully /intentionally) from the population on the basis of certain past experience & knowledge rather than random selection.
- A sampling process where every single individual elements in the population may not have an opportunity to be choosen as a sample.
- For example, one person could have a 10% chance of being selected and another person could have a 50% chance of being selected.

Probability Sampling

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster sampling
- Multi Stage Sampling

Simple Random Sampling

- Randomly any element can be chosen
- Chance of selection is totally in a randomized fashion.
- No previous knowledge, criteria and procedure is followed at the time of selection of the sample from the population.

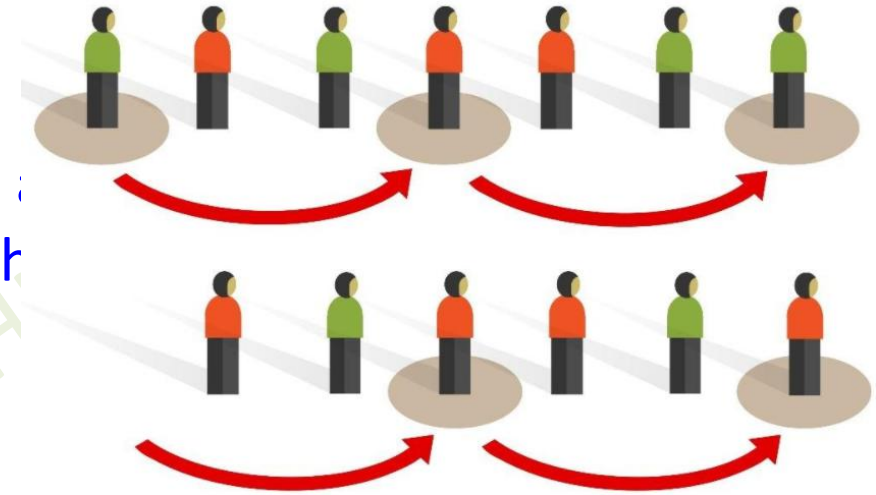


Example: Suppose we would like to select 10 students from any class consists of 75 students. Write the roll numbers of each students in separate cheats and put it in a container and 10 cheats from the container one by one randomly. Here probability of selection is 1/75

ADVT: Every element has an equal chance of getting selected to be the part sample.

Systematic Sampling

- Each member of the sample comes after an equal interval from its previous member.
- All the elements are put together in a sequence first where each element h the equal chance of being selected.
- Select a random starting point and then select the individual at regular intervals



Example: Suppose we would like to select 10 students from any class consists of 75 students. Choosing a random stating roll choose every 5th student.

ADVT: As each student has a chance of getting selected there is no biasness in selection.

Systematic Clustering (cont..)

For a sample of size **n**, we divide our population of size **N** into subgroups of **k** elements.

We select our first element randomly from the first subgroup of k elements.

To select other elements of sample, perform following:

We know number of elements in each group is **k** i.e **N/n**

So if our first element is $\mathbf{n_1}$ then Second element is $\mathbf{n_1+k}$ i.e $\mathbf{n_2}$

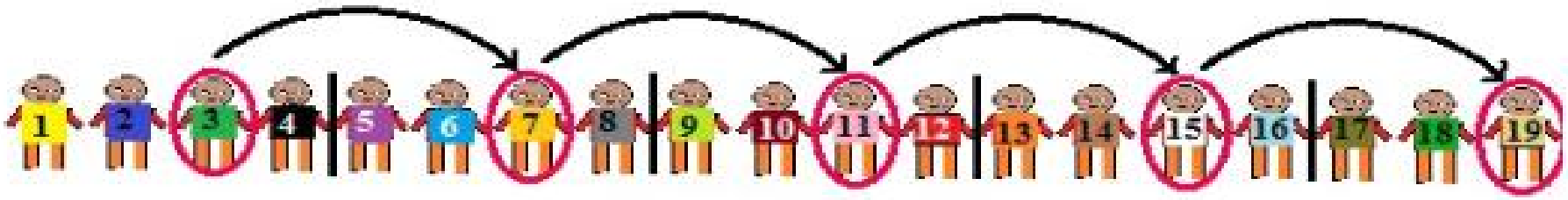
Third element $\mathbf{n_2+k}$ i.e $\mathbf{n_3}$ and so on..

Taking an example of N=20, n=5

No of elements in each of the subgroups is **N/n** i.e **20/5 =4= k**

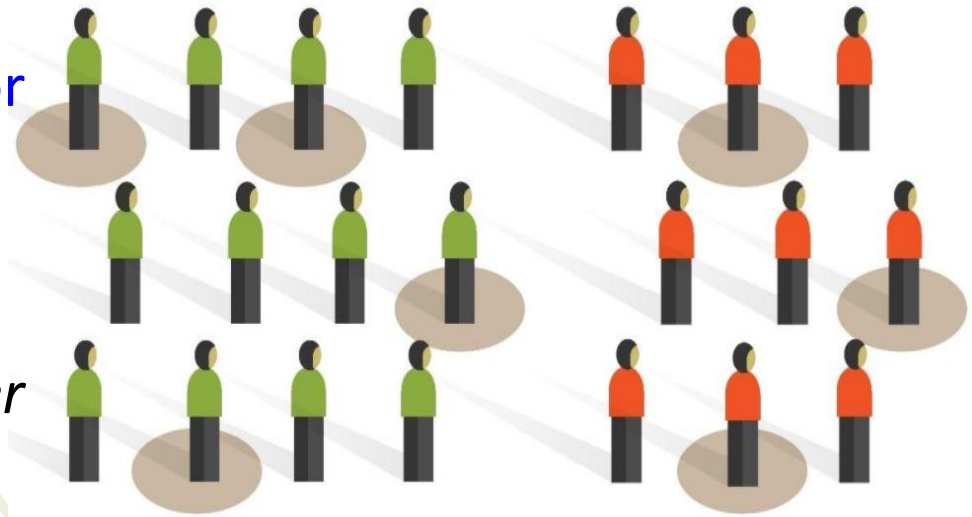
Now, randomly select first element from the first subgroup.

If we select $\mathbf{n_1= 3}$, $\mathbf{n_2 = n_1+k = 3+4 = 7}$, $\mathbf{n_3 = n_2+k = 7+4 = 11}$



Stratified Sampling

- The population is divided into smaller homogeneous groups or strata by some characteristics.
  i.e *the elements within the group are homogeneous and heterogeneous among the other subgroups formed.*



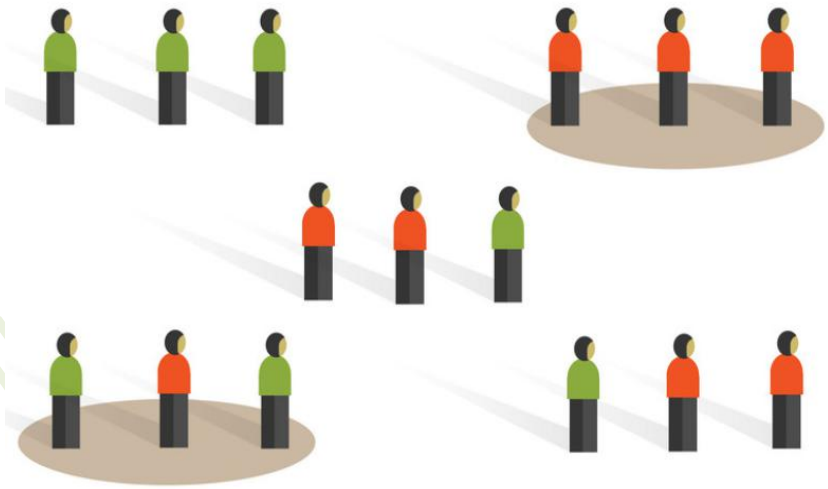
- The samples are selected randomly from these strata.
- We need to have prior information about the population to create subgroups

Example: Suppose we would like to select some students from any class consists of 75 students. The students will be divided into groups of boys and girls. Then some students will be chosen from boys and some from the girls.

ADVT: Members of each category or group will be chosen without any bias.

Cluster Sampling

- From the big population, choose a small group by diving it in to clusters/sections i.e area wise
- The clusters are randomly selected.
- All the elements of the cluster are used for sampling.



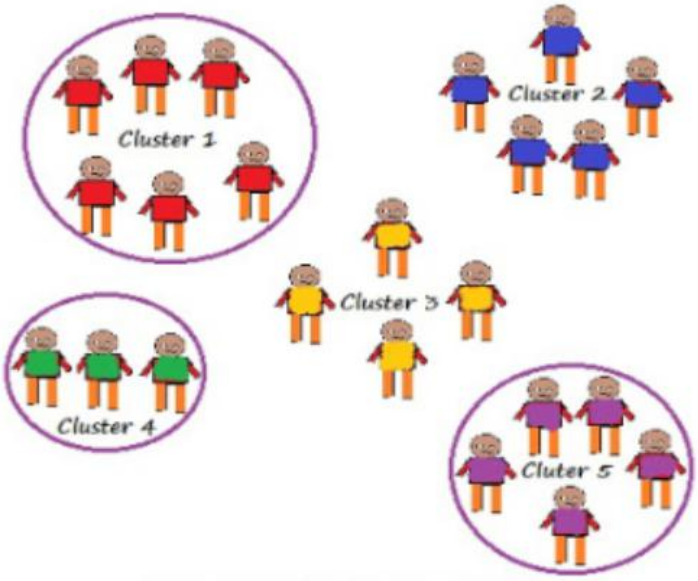
Example: Suppose we would like to know the awareness about COVID in a city. Instead of going the details survey of the entire city one can divide the city into clusters and randomly choose a cluster from that. All the members of the cluster will be considered.

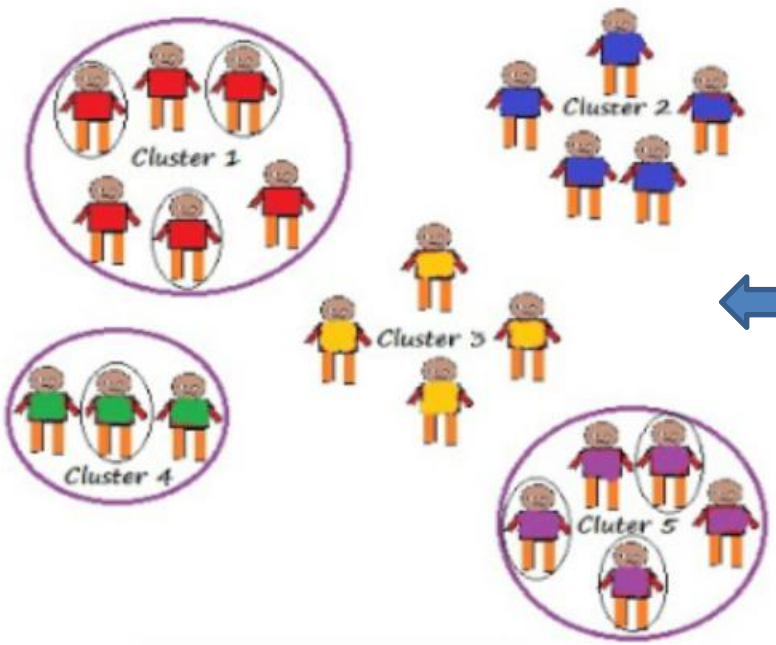
Cluster sampling can be done in following ways:

- **Single Stage Cluster Sampling**
- **Two Stage Cluster Sampling**

Single and Two stage Cluster Sampling

Dividing the entire population into clusters. Out of many clusters one cluster is selected randomly for sampling.



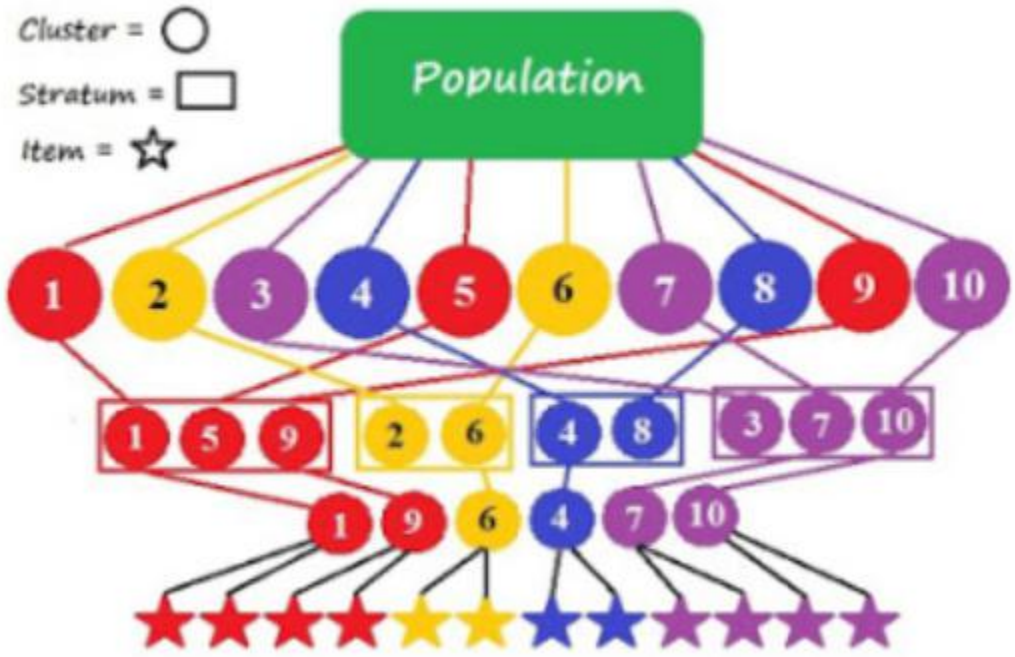


Dividing the entire population into clusters. Randomly select two or more clusters and then from those selected clusters again randomly select elements for sampling.

Example: An airline company wants to survey its customers one day, so they randomly select 55 flights that day and survey every passenger on those flights.

Multi Stage Sampling

- Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity.
- One or more clusters can be randomly selected from each stratum.
- This process continues until the cluster can't be divided anymore.
- Example : A country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.



Non-Probability Sampling

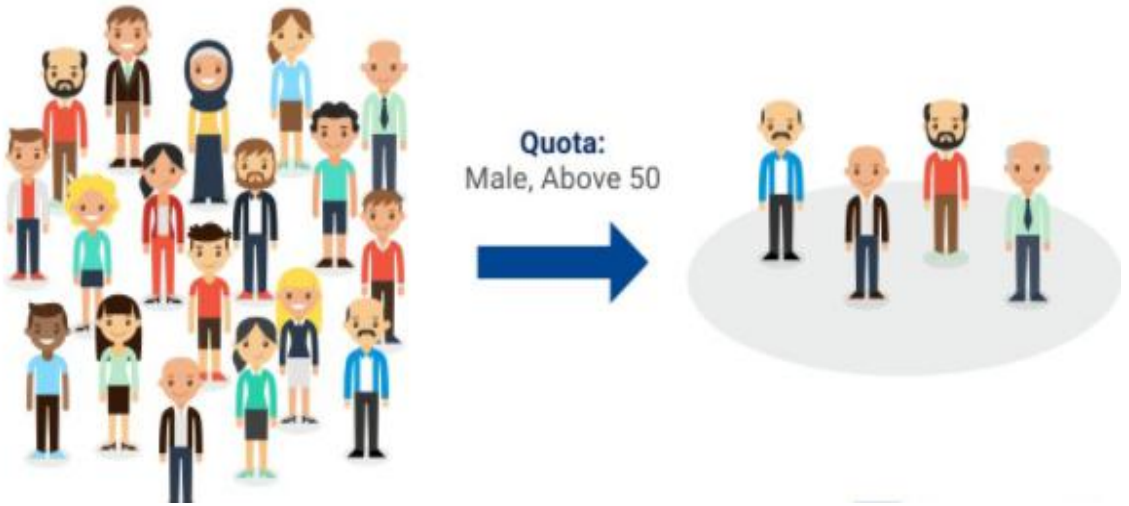
- Every element will be chosen purposefully/intentionally from the population on the basis of certain past experience and knowledge.
- It is a less stringent method.
- This sampling method depends heavily on the expertise of the researchers.
- It is carried out by observation, and researchers use it widely for qualitative research.
- Mainly classified into
  - **Quota Sampling**
  - **Purpose Sampling/Judgemental Sampling**
  - **Convenience Sampling**
  - **Referral / Snowball Sampling**

Quota Sampling

- Quota sampling works by first dividing the selected population into exclusive subgroups.
- The proportions of each subgroup are measured, and the ratio of selected subgroups are then used in the final sampling process.
- The proportions of the selected subgroups are used as boundaries for selecting a sample population of proportionally represented subgroups.

There are two types of quota sampling:

- **proportional**
- **non proportional.**



Proportional Quota Sampling

In proportional quota sampling you want to represent the major characteristics of the population by sampling a proportional amount of each.

The problem here is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?

For example, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota."

Non-Proportional Quota Sampling

Use when it is important to ensure that a number of sub-groups in the field of study are well-covered.

Use when you want to compare results across sub-groups.

Use when there is likely to a wide variation in the studied characteristic within minority groups.

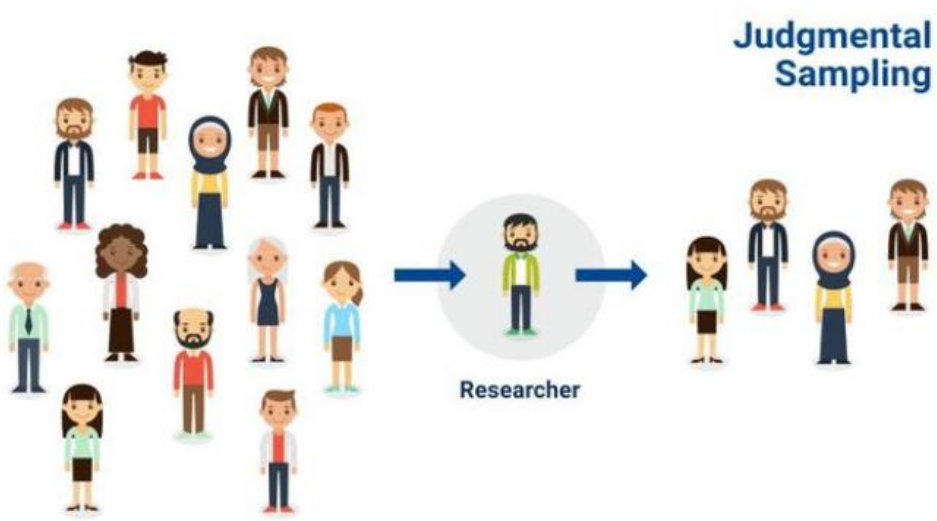
- Identify sub-groups from which you want to ensure sufficient coverage. Specify a minimum sample size from each sub-group.
- Here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population.

Example

A study of the prosperity of ethnic groups across a city, specifies that a minimum of 50 people in ten named groups must be included in the study. The distribution of incomes across each ethnic group is then compared against one another.

Purpose Sampling/Judgemental Sampling

- Samples are chosen only on the basis of the researcher's knowledge and judgement.
- It enables the researcher to select cases that will best enable him to answer his research questions that meet the objective.
- Choosing a sample because of represent the certain purpose.



Example-1: In online live voting for selecting a GOOD Singer from a competition, the people who have interest in singing can be selected in the sample .

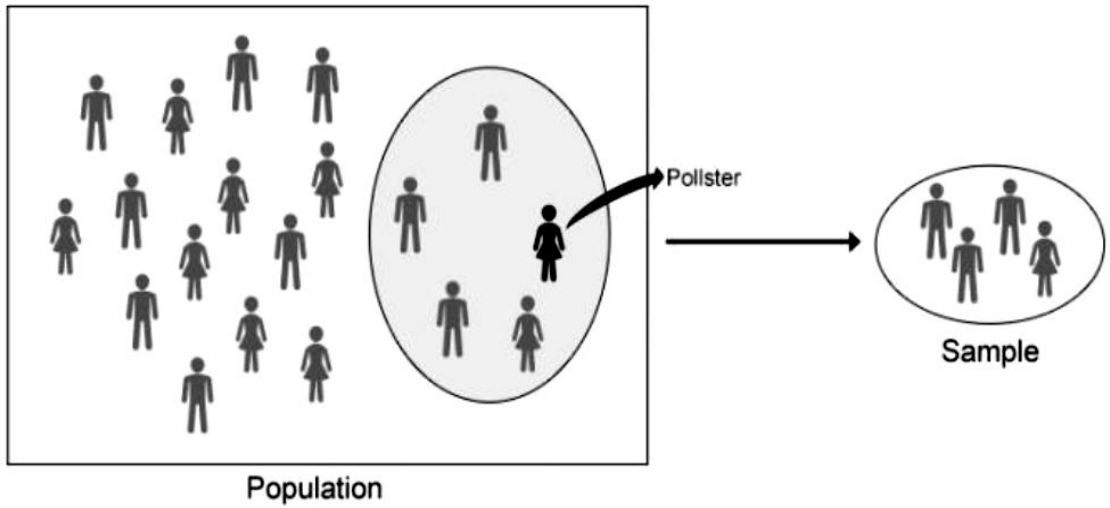
Example-2: If we want to understand the thought process of the people who are interested in pursuing master's degree then the selection criteria would be "Are you interested for Masters in..?"
All the people who respond with a "No" will be excluded from our sample.

Convenience Sampling

- Convenience sampling ***(also called accidental sampling or grab sampling)*** is where you include people who are easy to reach.
- Sample are taken mainly on basis of the readily available.
- Sample which is convenient to the researcher or the data analyst can be chosen. The task is done without any principles or theories.

For example, you could survey people from:

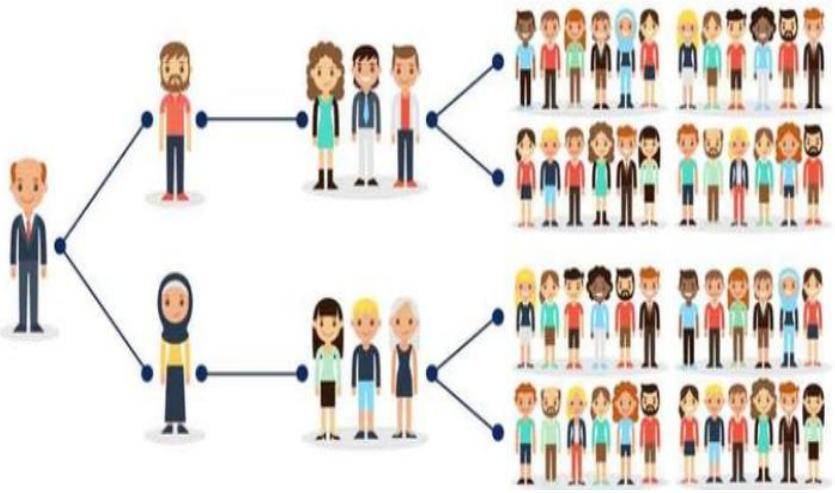
- Your workplace,
- Your school,
- A club you belong to,
- The local mall.



Example: Suppose I would like to select 5 students from any class consists of 75 students. Choosing the 5 students who sits near by me without any principle of selection.

Referral / Snowball Sampling

- Snowball sampling method is purely based on referrals and that is how a researcher is able to generate a sample.
- So the researcher will take the help from the first element which he select for the population and ask him to recommend others who will fit for the description of the sample needed.
- So this referral technique goes on, increasing the size of population like a snowball.



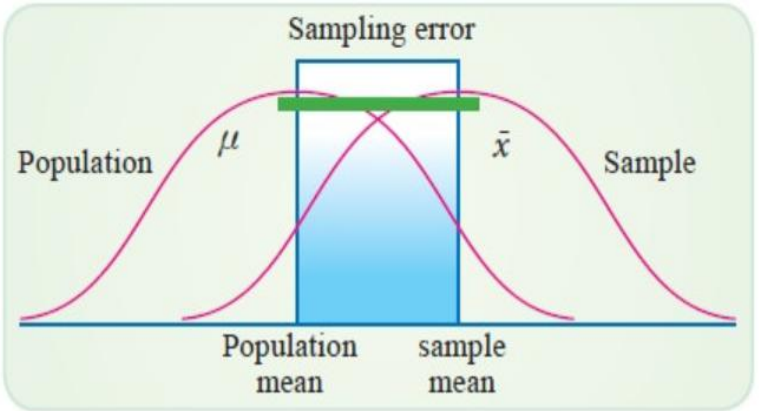
Example: ***If you are studying the level of customer satisfaction among the members of an elite country club, you will find it extremely difficult to collect primary data sources unless a member of the club agrees to have a direct conversation with you and provides the contact details of the other members of the club.***

Sampling Errors

- Sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data.
- The results found in the sample thus do not represent the results that would be obtained from the entire population.
- Sampling error can be reduced by randomizing sample selection and/or increasing the number of observations.
- It mainly happens when the sample size is very small (10 to 100).

For example, if you wanted to figure out how many people out of a thousand were under 18, and you came up with the figure 19.357%. If the actual percentage equals 19.300%, the difference (19.357 – 19.300) of 0.57 or 3% = the margin of error. If you continued to take samples of 1,000 people, you'd probably get slightly different statistics, 19.1%, 18.9%, 19.5% etc, but they would all be around the same figure. This is one of the reasons that you'll often see sample sizes of 1,000 or 1,500 in surveys: they produce a very acceptable margin of error of about 3%.

Formula: the formula for the margin of error is $1/\sqrt{n}$, where n is the size of the sample. For example, a random sample of 1,000 has about a $1/\sqrt{n}$; = 3.2% error.



Five Common Types of Sampling Errors

- **Population Specification Error**—This error occurs when the researcher does not understand who they should survey.
- **Sample Frame Error**—A frame error occurs when the wrong sub-population is used to select a sample.
- **Selection Error**—This occurs when respondents self-select their participation in the study – only those that are interested respond. Selection error can be controlled by going extra lengths to get participation.
- **Non-Response**—Non-response errors occur when respondents are different than those who do not respond. This may occur because either the potential respondent was not contacted or they refused to respond.
- **Sampling Errors**—These errors occur because of variation in the number or representativeness of the sample that responds. Sampling errors can be controlled by (1) careful sample designs, (2) large samples, and (3) multiple contacts to assure representative response.

Thank you