

# **ROLE OF SURVEY SAMPLING IN DATA COLLECTION**

By

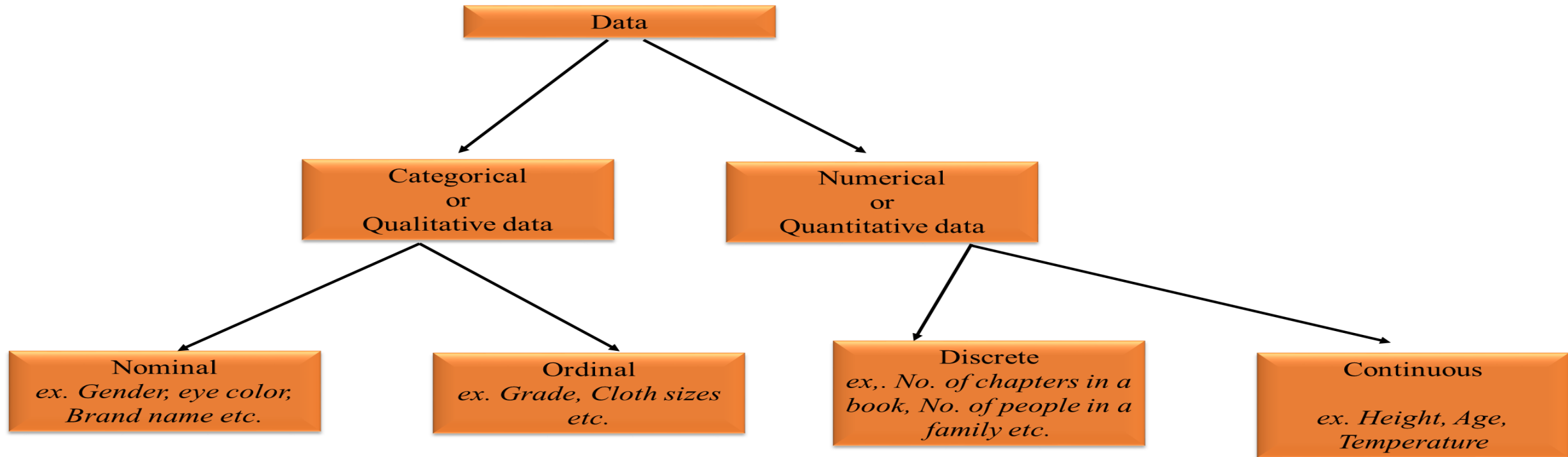
**Prof. G. N. Singh**

Professor & HOD

Department of Mathematics & Computing  
Indian Institute of Technology (Indian School of Mines), Dhanbad.

# ❖ Overview of Statistics

❖ **Data:** Data are individual pieces of factual information recorded and used for the purpose of analysis. It is also a numerical representation of fact.



## ❖ Characteristics of Data

- i. Central tendency
- ii. Dispersion
- iii. Skewness
- iv. Kurtosis

## ❖ Collection of Data

- i. Census or Complete Count
- ii. Survey Sampling

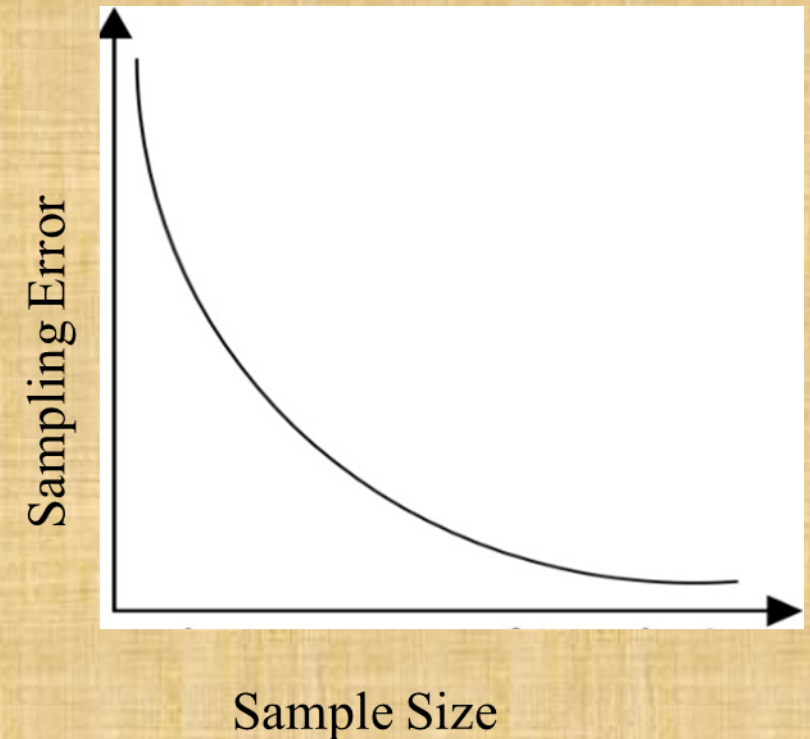


## ❖ Errors in Data Collection

- i. Sampling Error
- ii. Non-Sampling Error

### i. Sampling Error

- If the sample is not the true representative of the whole population.





## **ii. Non-Sampling Error**

The main sources of Non-Sampling errors are:

- Failure to measure some of the units in the selected sample may be due to non-response.
- Observational error due to defective measurement technique.
- Errors introduced in editing, coding and tabulation the results

## ❖ Census or Complete Count

The total count of all units of the population for a certain characteristic or for many characters is known as ***Complete enumeration***, also termed as *Census Survey*.

- Huge amount of cost, man power and time required.
- Some times the complete enumeration is not possible when the units are perishable in nature. For example- Bullet testing.

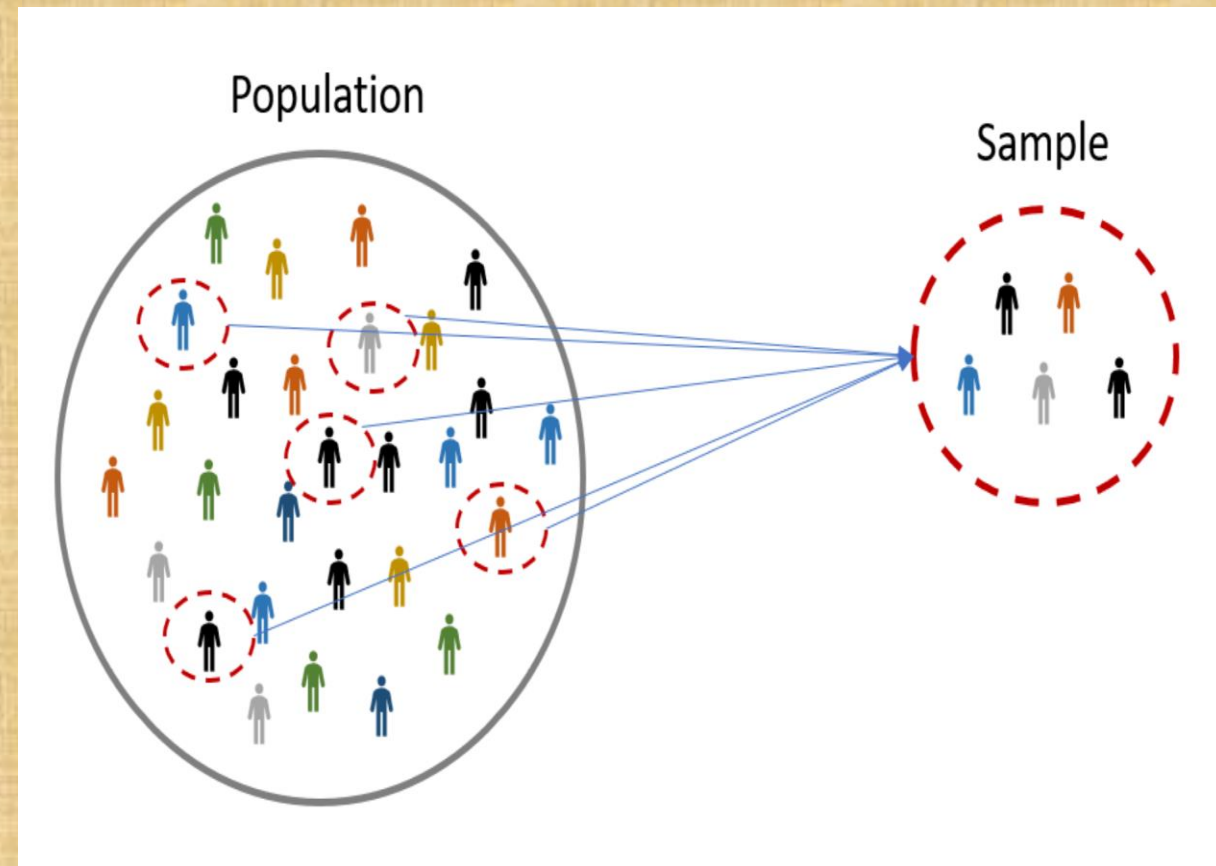


## ❖ Sample

- It is a unit that is selected from population.
- Representative the whole population.
- Purpose to draw the inference.

## ❖ Sampling Frame

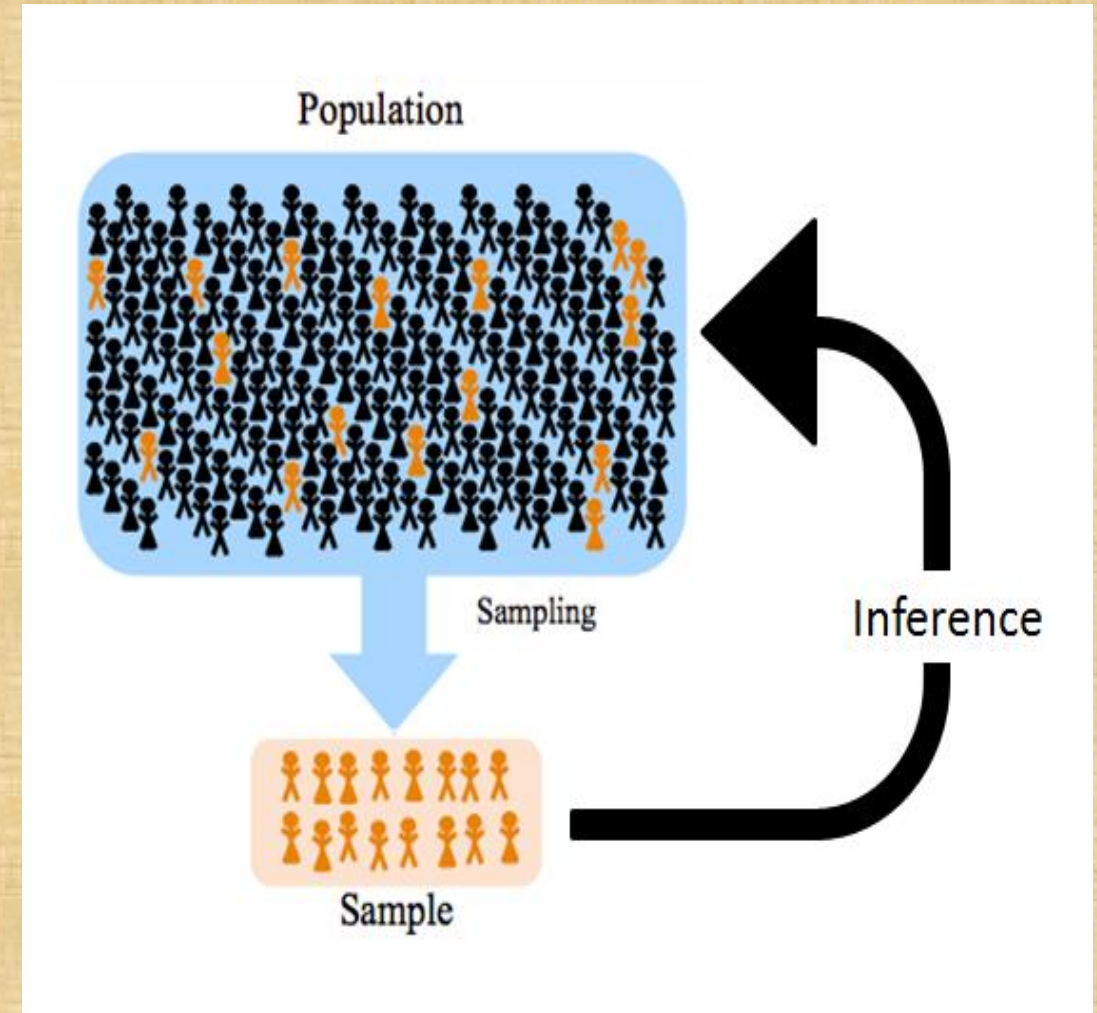
- List of population units for identification.





## ❖ Sampling

- Sampling is the process of selecting desired number of units from a population (a sample) to provide an adequate description and inferences of the population or equivalently about the population parameters.



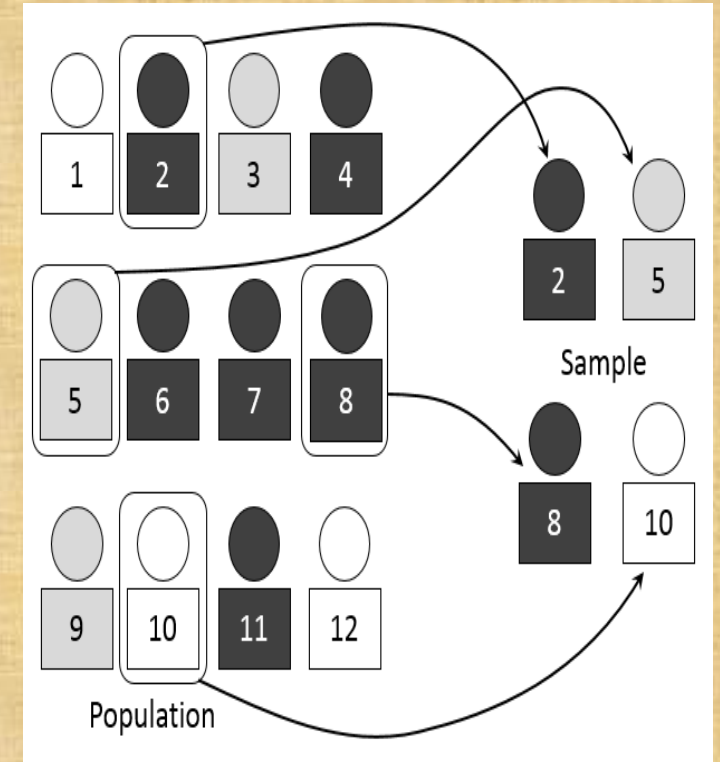


## ❖ Need of Sampling

- Complete enumeration may not be possible.
- Resources: lower cost lesser resources required.
- Speed: Faster results due to lesser coverage.
- Reliable information

To draw conclusions about population from sample, there are two major requirements for a sample

- ✓ Sample size should be large subject to the availability to the resources.
- ✓ Sample has to be selected appropriately so that it is representative of the population. Sample should have all the characteristics of the population.



# ❖ Advantage and Disadvantage of Sampling

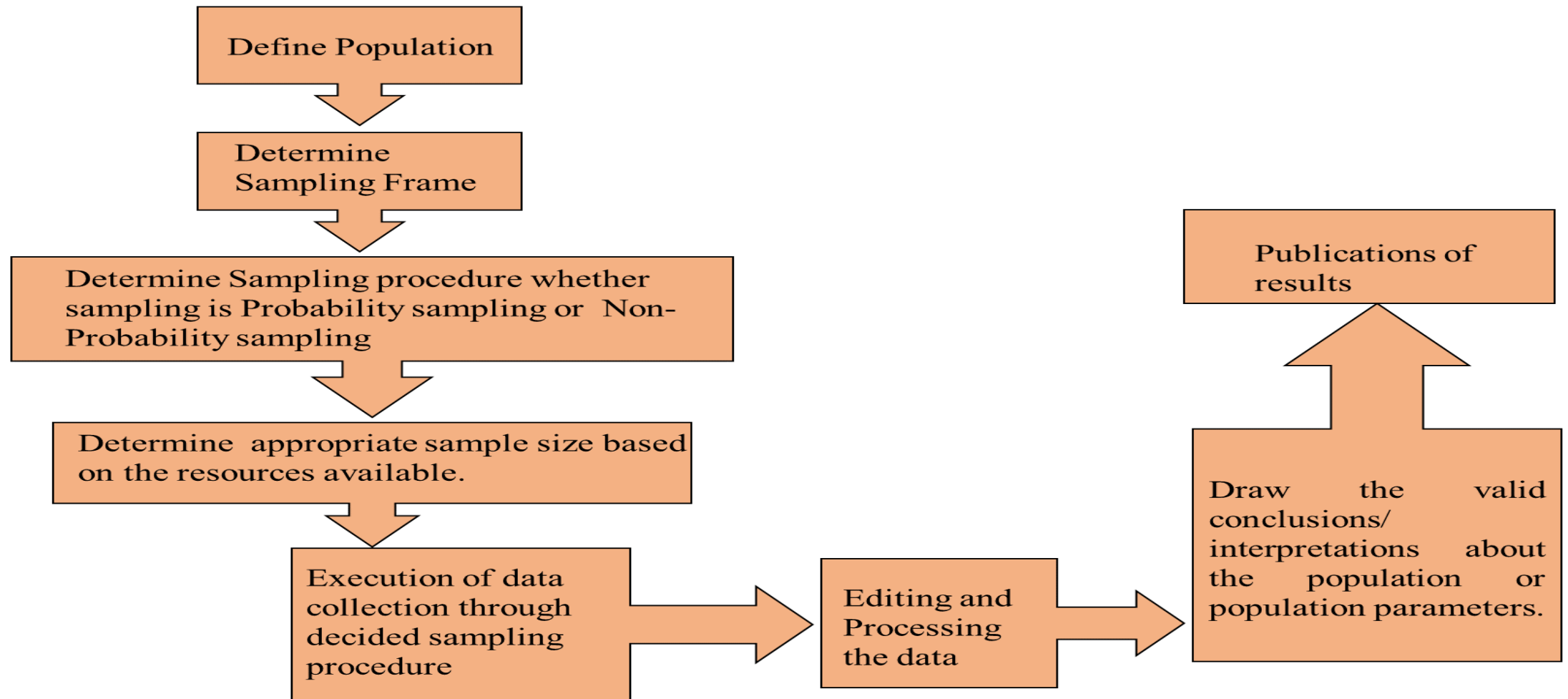
## ❖ Advantages

- Saves time and labor.
- Results in reduction of cost in terms of money and time.
- Ends up with greater accuracy of results
- Has greater scope.
- Has greater adaptability.
- If the population is too large, or hypothetical sampling is the only method to be used.

## ❖ Disadvantages

- There is always a sampling error.
- Sampling may create a feeling of discrimination within the population.
- Sampling may be inadvisable where every unit in the population is legally required to have a record.
- For rare events, small samples may not yield sufficient cases for study.
- Sample may be biased if hidden periodicity in population coincides with that of selection.

# ❖ Steps of Conducting a Sample Survey





# ❖ Principles of Sampling

The theory of sampling is based on three important basic principles:

## ❑ Principle of Statistical Regularity

- A moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group.

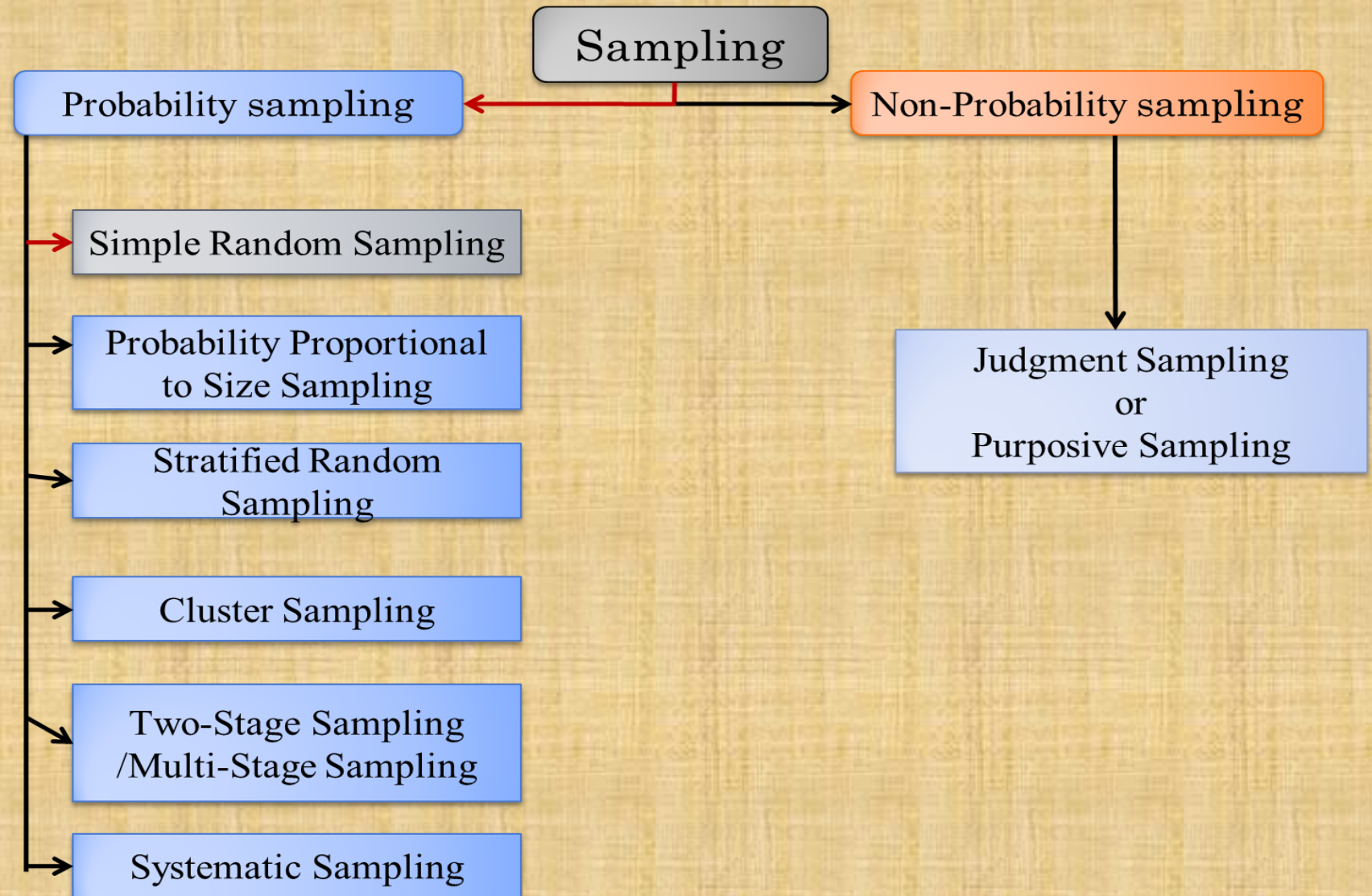
## ❑ Principle of Validity

- The sampling design provides valid estimates about population parameters which means sample should be selected in such a way so that estimates can be explained in terms of probability.

## ❑ Principle of Optimization

- Given level of efficiency at minimum cost.
- Maximum possible efficiency with given cost.

# ❖ Types of Survey Sampling



# ❖ Judgmental Sampling

- It is also known as purposive sampling.
- Respondents are selected according to an experienced researcher's belief that they will meet the requirements of the study.





# ❖ Advantage and Disadvantage of Judgmental Sampling

## ❖ Advantage

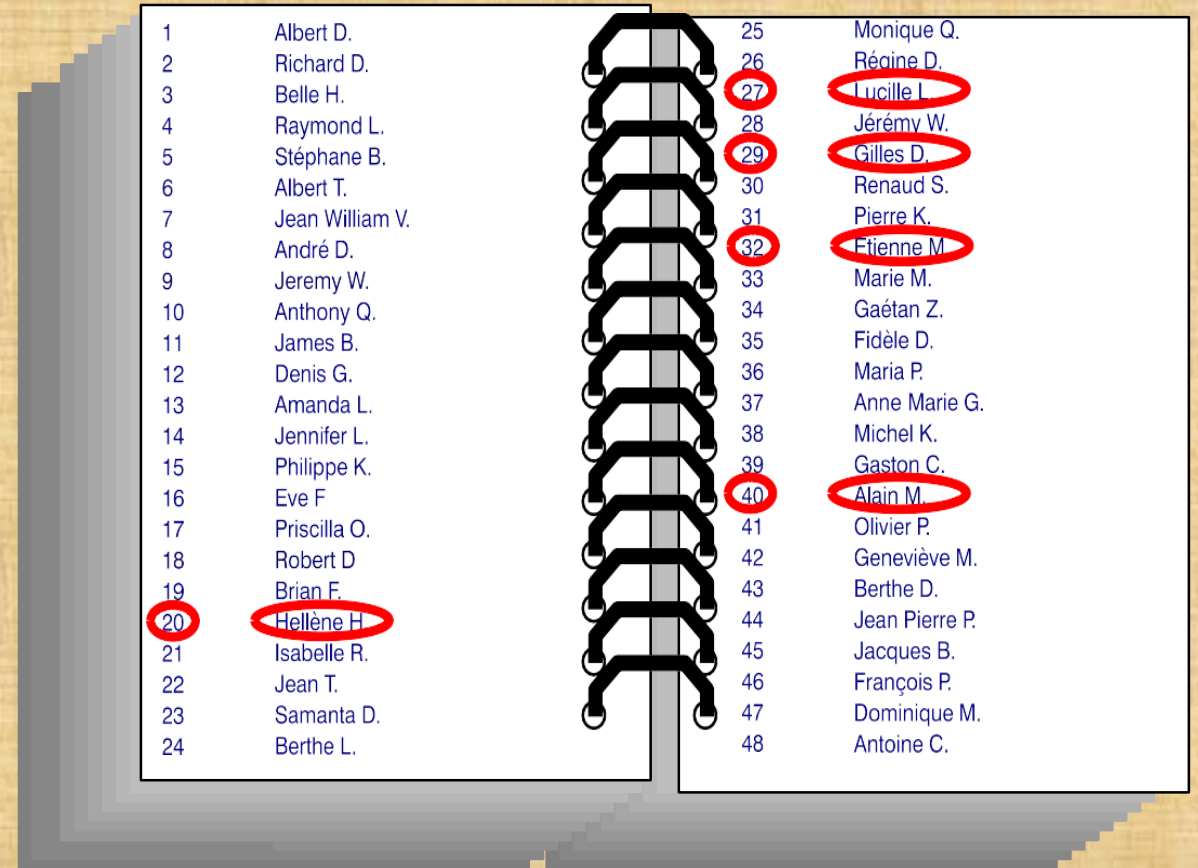
- Low cost.
- Less time involved.
- A select number of people who are known to be related to the topic are part of the study which means that there are lesser chances of having people who will distort the data.
- Meet the specific objective.
- Mostly used in demographic surveys.

## ❖ Disadvantage

- It can be subject to researcher's bias.
- The group selected may not be represent all the population.

## ❖ Simple Random Sampling

- Simplest and most common method.
- Also known as *Unrestricted random sampling*.
- Each unit of the population has equal chance of selection.



1	Albert D.	25	Monique Q.
2	Richard D.	26	Réine D.
3	Belle H.	27	Lucille L.
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	Etienne M.
9	Jeremy W.	33	Marie M.
10	Anthony Q.	34	Gaétan 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	Hellène H.	44	Jean Pierre P.
21	Isabelle R.	45	Jacques B.
22	Jean T.	46	François P.
23	Samanta D.	47	Dominique M.
24	Berthe L.	48	Antoine C.

# ❖ Advantage and Disadvantage of Simple Random Sampling

## ❖ Advantage

- Minimal knowledge of population needed.
- Easy to analyze data.
- High probability of representative sample.
- Meets assumptions of many statistical procedures.

## ❖ Disadvantage

- Low frequency of use.
- Does not use researchers' expertise.
- Larger risk of sampling error.



## ❖ Probability Proportional to Size Sampling

- This type of sampling procedure where the probability of sample is proportional to the size of the unit.
- Probability of drawing any specified unit differs from draw to draw.



# ❖ Advantage and Disadvantage of PPS Sampling

## ❖ Advantage

- Results in smaller sample sizes.
- Effective for overstatement errors.
- Generally simpler to use than classical variable sampling.
- SRS is a particular case of this sampling.

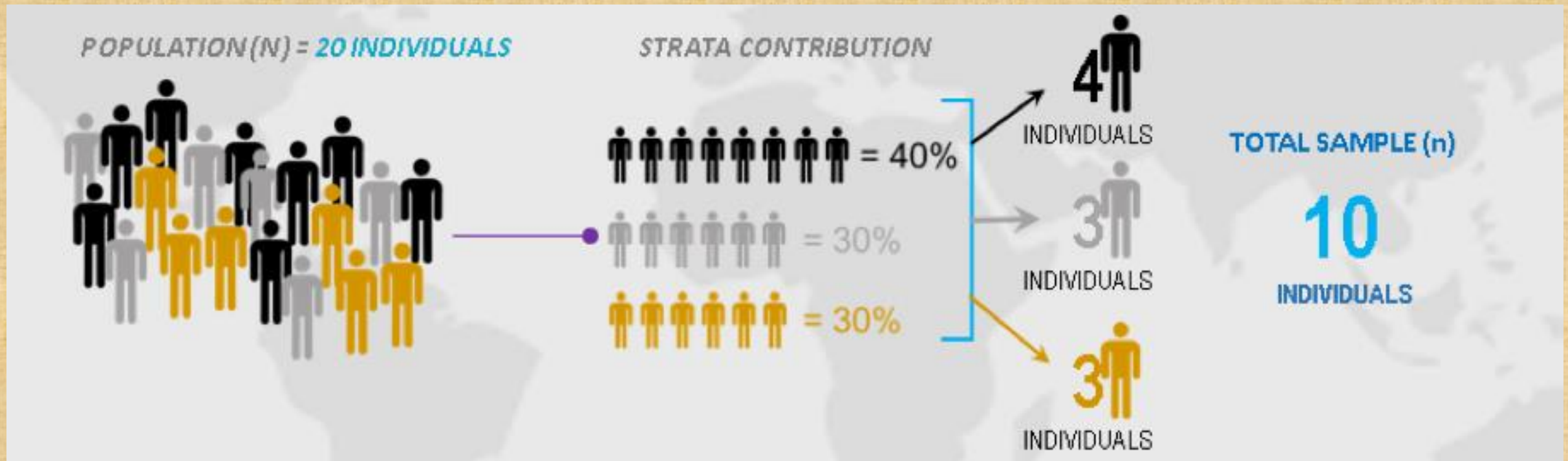
## ❖ Disadvantage

- Expanding a PPS sample is difficult.
- Provide a conservative (higher) estimate of misstatement.



## ❖ Stratified Random Sampling

- Population is divided into the homogeneous groups (with respect to study character) called *Strata*.
- Samples are randomly selected independently from each strata.





# ❖ Advantage and Disadvantage of Stratified Random Sampling

## ❖ Advantage

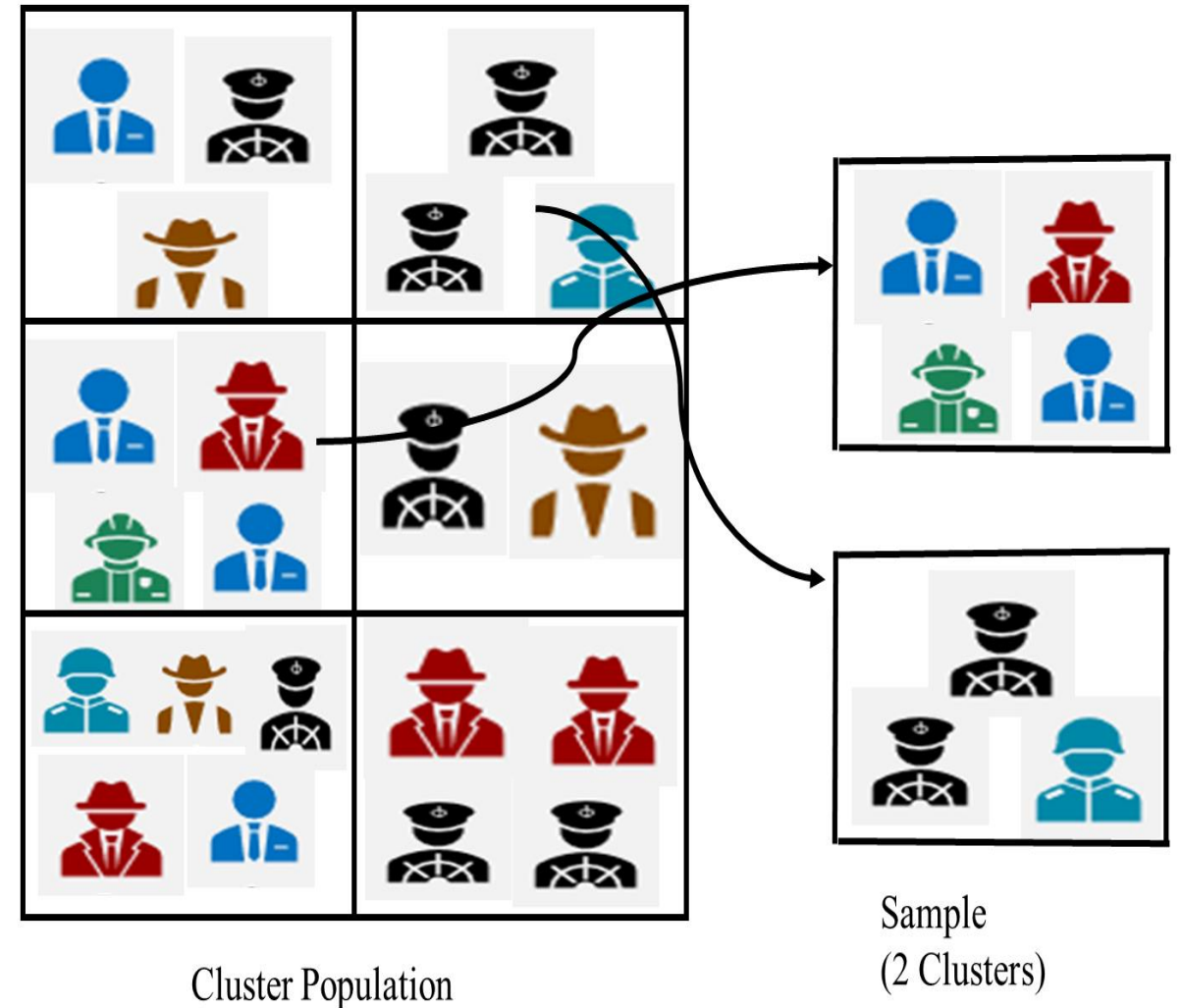
- Minimizing sample selection bias.
- Less variability.
- Cost reduced if strata already exists.
- It can be used for both proportional and non-proportional samples.
- Stratification brings a gain in the precision in estimation of a characteristic of a population.

## ❖ Disadvantage

- It is unusable when researchers cannot confidently classify every member of the population into a subgroup
- The sorting process becomes more difficult, rendering stratified random sampling an ineffective and less than ideal method
- High cost

## ❖ Cluster Random Sampling

- The population is divided into subgroups (clusters) according to some well defined rule. (for ex. Families).
- Choose a sample of clusters according to some procedure.
- Carry out a complete enumeration of the selected clusters



# ❖ Advantage and Disadvantage of Cluster Random Sampling

## ❖ Advantage

- Very useful when populations are large and spread over a large geographical region.
- Simple as complete list of sampling units within population not required.
- Less travel / resources required.
- Economically efficient.

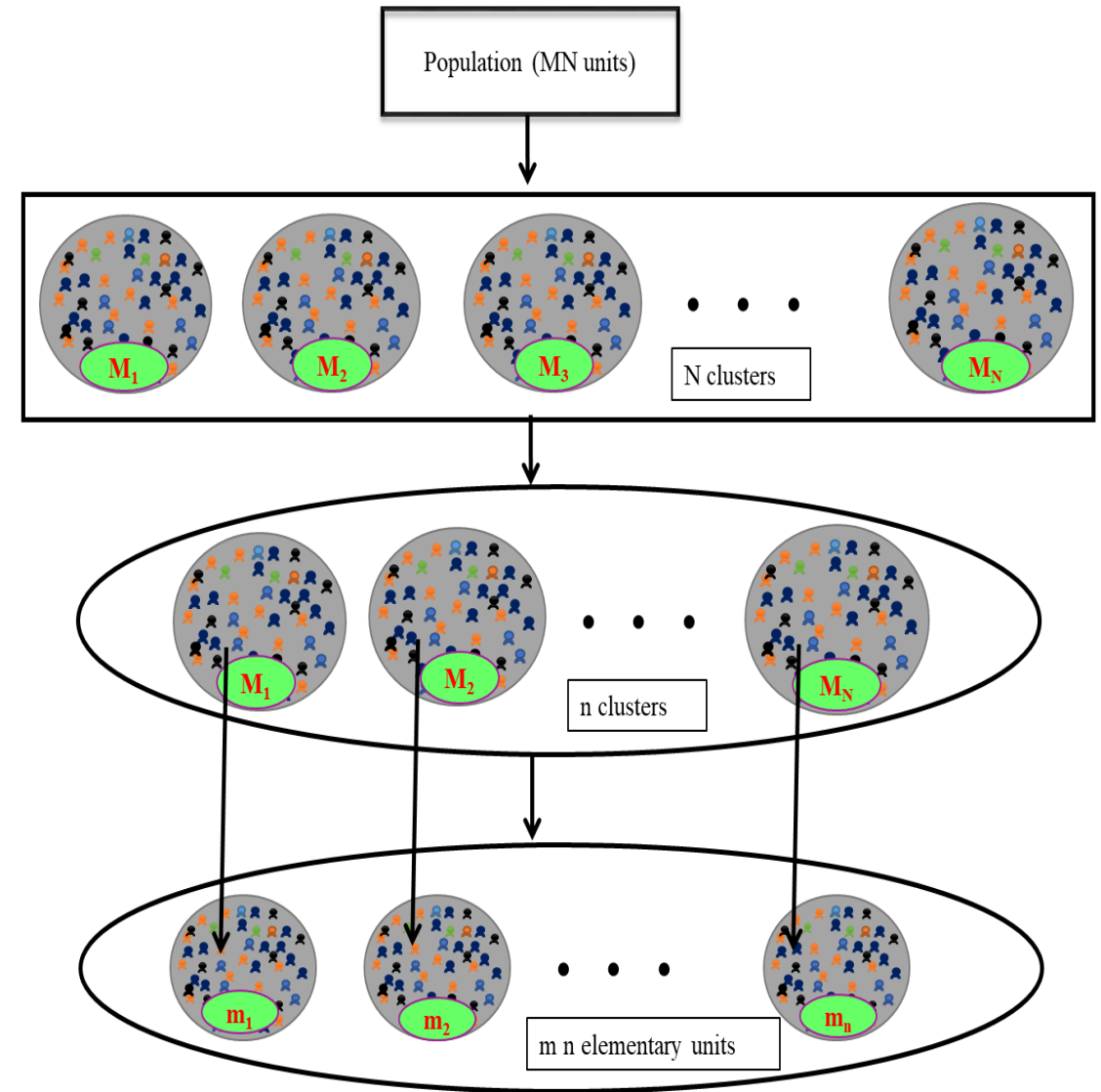
## ❖ Disadvantage

- Cluster may not be representative of whole population but may be too alike.
- Statistically less efficient i.e. standard error of the estimate is likely to be large.



## ❖ Two-Stage Sampling

- In this sampling first select the clusters.
- Selecting a specified number of elements from each selected cluster.



# ❖ Advantage and Disadvantage of Two-Stage Sampling

## ❖ Advantage

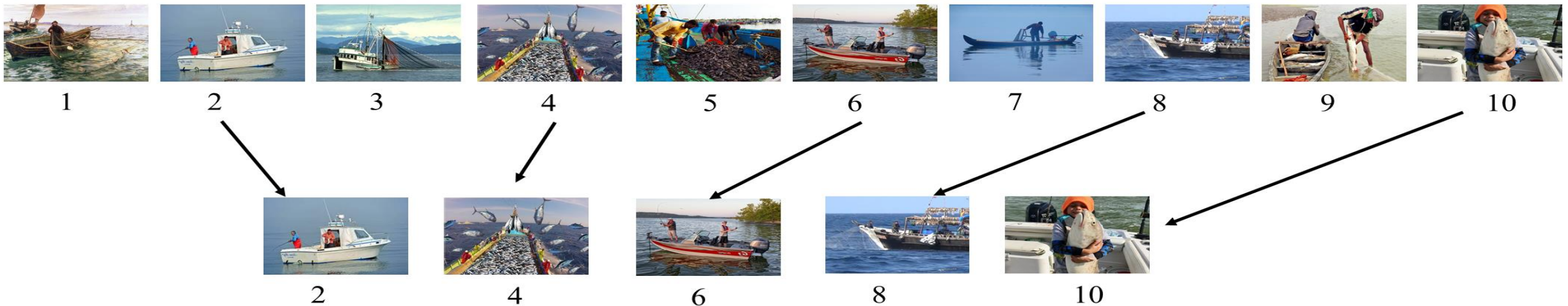
- More Accurate.
- More Effective.
- Most feasible approach for large population.

## ❖ Disadvantage

- Each stage in sampling introduces sampling error-the more stages there are, the more error there tends to be.
- Needs more man power.

# ❖ Systematic Random Sampling

- First unit is selected with the help of random numbers.
- Rest get selected automatically according to some pre-assigned pattern .
- Found very useful in forest surveys for estimating the volume of timber, in fisheries for estimating the total catch of fish etc.
- $N=n*k$ , where  $k$  is an integer.



In this case, every second person is systematically selected.

In this case,  $N=10$ ;  $n=5$ ;  $k=2$



# ❖ Advantage and Disadvantage of Systematic Random Sampling

## ❖ Advantage

- Moderate cost; moderate usage.
- Simple to draw sample.
- Easy to verify.

## ❖ Disadvantage

- Periodic ordering required.
- Sample sizes may vary.
- Not possible to estimate the population variance.

# ❖ Methods of Estimation

## ❖ About Auxiliary Information

- Auxiliary information means additional information about the character under study.
  - It is easily available for each unit of the population
- For ex: Number of doctors, beds, supporting staff & patients which is known can be used as auxiliary information on hospital survey.



# ❖ Methods of Estimation with and without Auxiliary Information

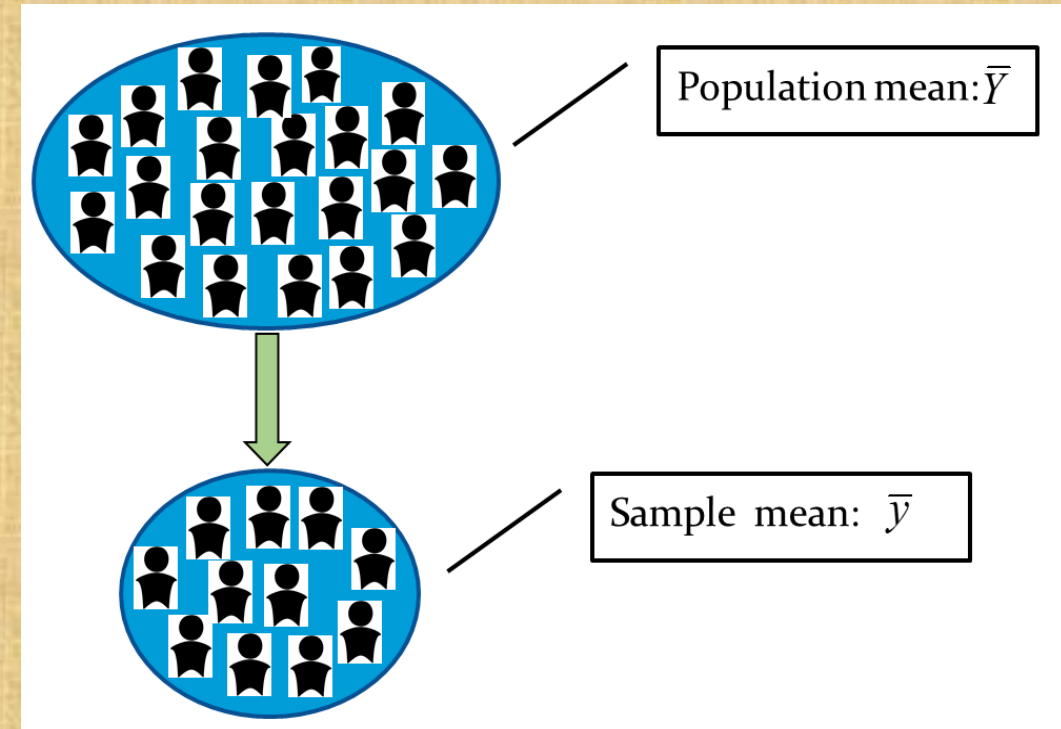
- Without using Auxiliary Information.
  - For ex. Sample mean
- By using Auxiliary Information.
  - For ex. (i). Ratio estimator
    - (ii). Product estimator
    - (iii). Regression estimator



## ❖ Sample mean

- Let  $y_i, i = 1, 2, \dots, n$  denote the sample value of the  $i^{\text{th}}$  unit selected in the sample, then the sample mean is defined as:

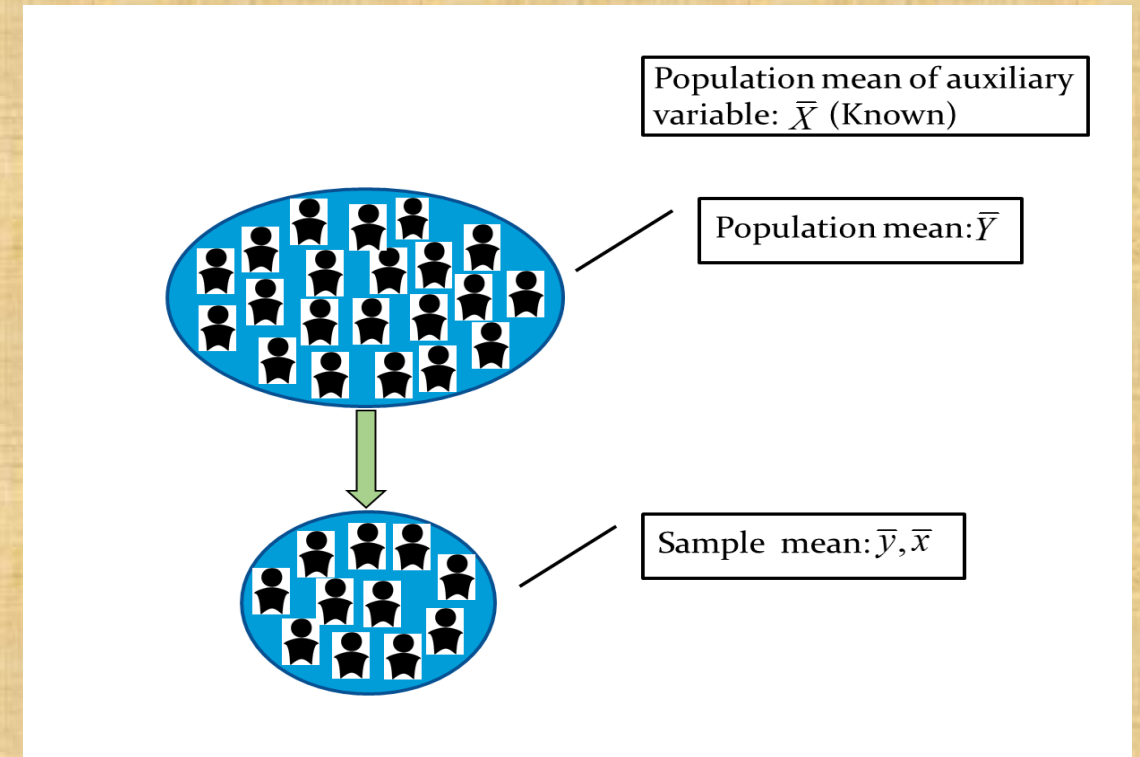
$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$



## ❖ Ratio estimator

- Cochran (1940) was the first to show the contribution of known auxiliary information in improving the efficiency of the estimator population mean in survey sampling.

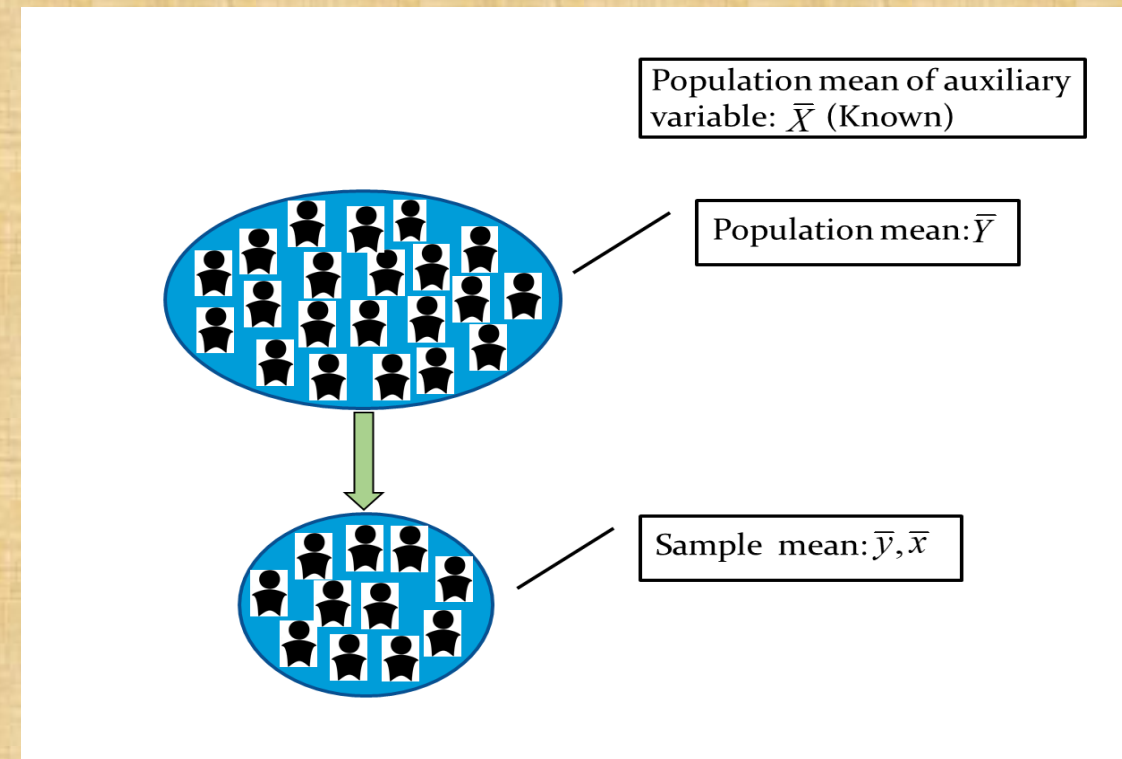
$$\bar{y}_R = \bar{y} \left( \frac{\bar{X}}{\bar{x}} \right)$$



## ❖ Product estimator

- Murthy (1964) considered another estimator of population mean using known population mean of auxiliary variable as a Product estimator

$$\bar{y}_P = \bar{y} \left( \frac{\bar{x}}{\bar{X}} \right)$$





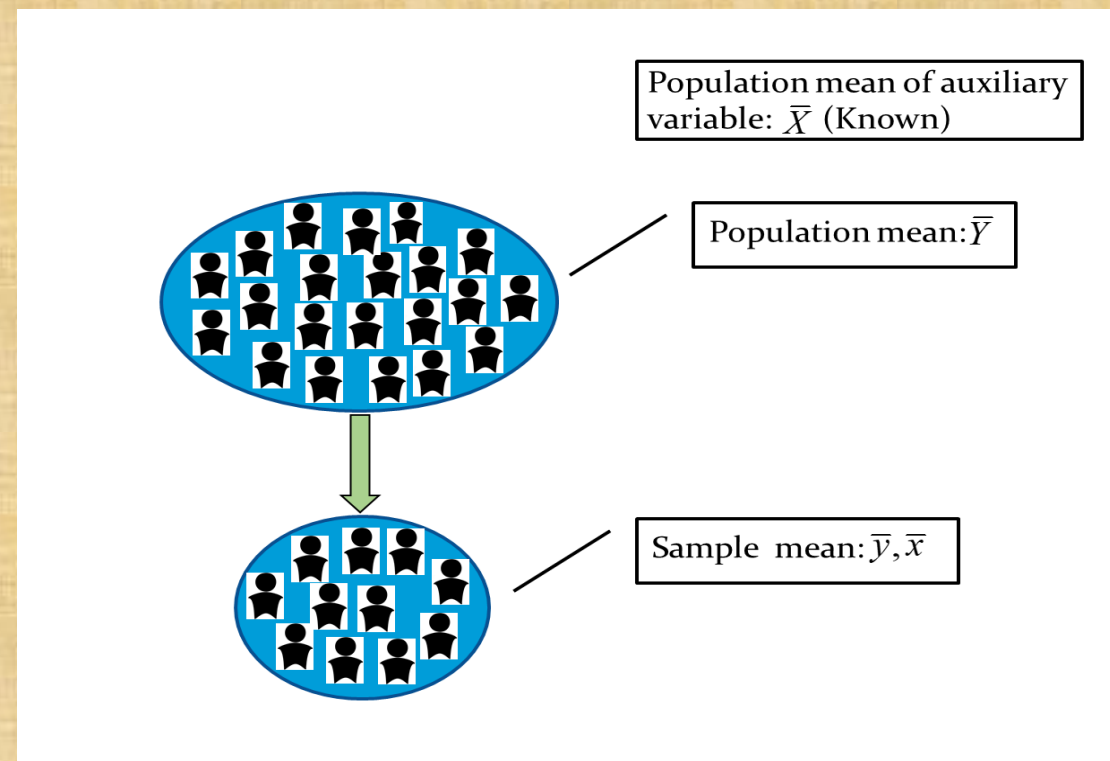
## ❖ Regression estimator

- Hansen, Hurwitz and Madow (1953) consider the linear difference/ regression estimator of the population mean as:

$$\bar{y}_{LR} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$$

$$\bullet \text{Regression Coefficient: } \beta = \frac{S_{xy}}{S_x^2}$$

$$\bullet \hat{\beta} = \frac{s_{xy}}{s_x^2}$$

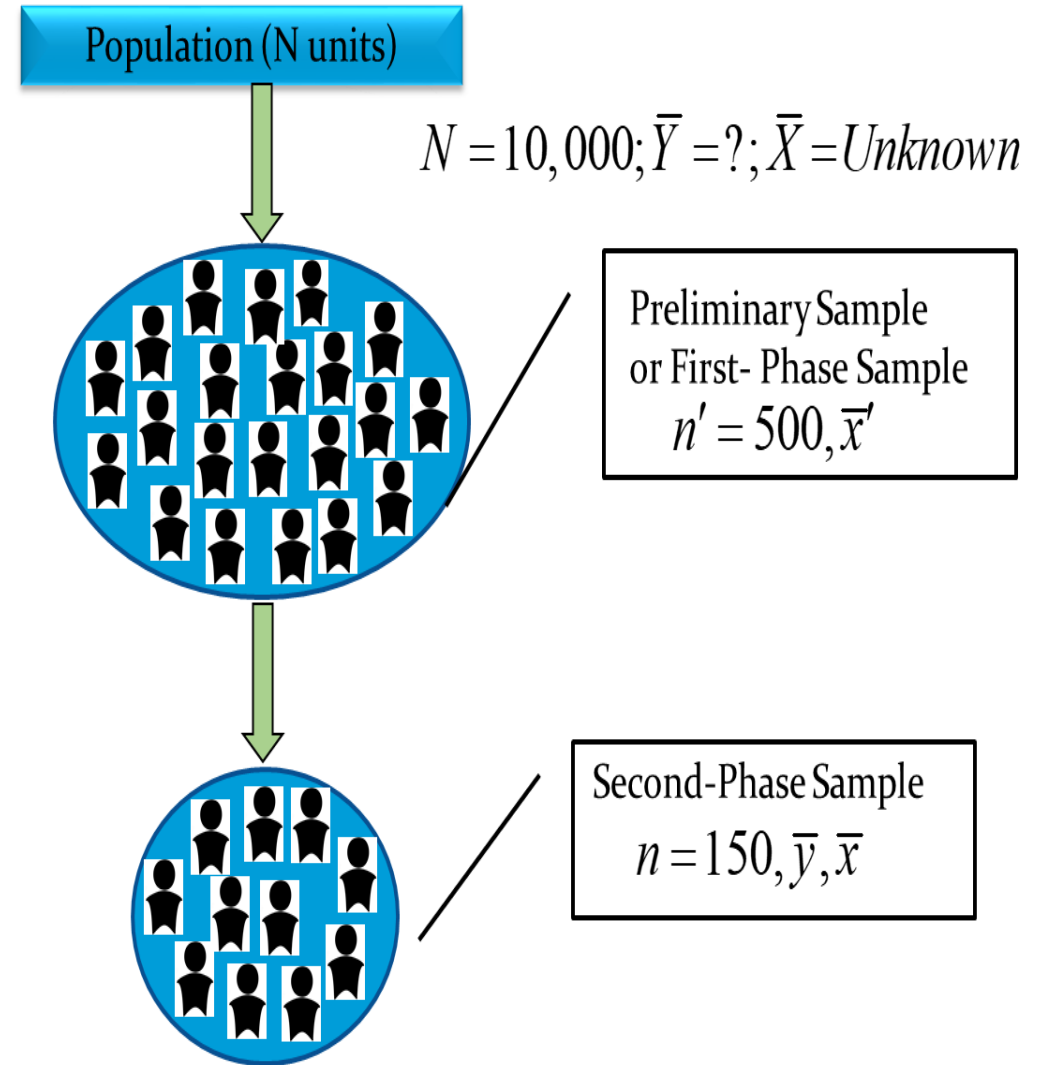


## ❖ Know more about Ratio, Product and Regression estimators

Ratio Estimator	Product Estimator	Regression Estimator
<ul style="list-style-type: none"><li>The correlation between <math>y</math> and <math>x</math> must be positive and high (within <math>+0.5</math> and <math>+1.0</math>)</li></ul>	<ul style="list-style-type: none"><li>The correlation between <math>y</math> and <math>x</math> must be negative and high (within <math>-1.0</math> and <math>-0.5</math>)</li></ul>	<ul style="list-style-type: none"><li>The correlation between <math>y</math> and <math>x</math> must be non-zero within the range <math>[-1.0</math> and <math>+1.0]</math>.</li></ul>

## ❖ Two-Phase Sampling

- Also known as *Double Sampling*.
- The randomization is done twice.
- First a random sample is drawn from the population.
- Then again a random second phase sample is drawn from the first sample.





# ❖ Advantage and Disadvantage of Two-Phase Sampling

## ❖ Advantage

- Offers more detailed information on the topic of the study.
- Cost effective.

## ❖ Disadvantage

- Necessary to have a complete sampling frame of units.

## ❖ The Two-Phase Ratio, Product and Regression Estimators

- **Ratio estimator**

$$\bar{y}_R = \bar{y} \left( \frac{\bar{x}'}{\bar{x}} \right)$$

- **Product estimator**

$$\bar{y}_P = \bar{y} \left( \frac{\bar{x}}{\bar{x}'} \right)$$

- **Regression estimator**

$$\bar{y}_{LR} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$$



THANK  
YOU