Essence of Sampling

- Sampling is a process of selecting subset of observations/records from a population to make inference about various population parameters such as mean, proportion, standard deviation, etc

- It is an important step in inferential statistics since an incorrect sample may lead to wrong inference about the population

Sampling is necessary when it is difficult or expensive to collect data on the entire population. The inference about the population is made based on the sample that was collected; incorrect sample may lead to incorrect inference about the population.

POPULATION PARAMETERS

- Measures such as mean and standard deviation calculated using the entire population are called *population parameters*

- The population parameters mean and standard deviation are usually denoted using symbols μ and σ, respectively

SAMPLE STATISTIC

- When population parameters are estimated from sample they are called *sample statistic* or *statistic*

- The sample statistic is denoted using symbols (for mean) and S (or s for standard deviation)

SAMPLING

The process of identifying a subset from a population of elements is called sampling process or simply sampling

Steps used in any Sampling process:

- Identification of target population that is important for a given problem under study
- Decide the sampling frame.
- Determine the sample size
- Sampling method

Random Sampling

- Shewhart (1931) defines random sample as a ‘sample drawn under conditions such that the law of large number applies’

- Random sampling is usually carried out **without replacement**, that is, an observation which is selected in the sample is removed from the population for further consideration

- Random samples can also be created **with replacement**, that is, an observation which is selected for inclusion in the sample can again be considered since it is replaced (not removed) in the population.

Stratified Sampling

- The population can be divided into mutually exclusive groups using some factor (for example, age, gender, marital status, income, geographical regions, etc.). The groups, thus, formed are called **stratum**
- It is important that the groups are mutually exclusive and exhaustive of the population.

Stratified Sampling -Examples

a) Amount of time spent by male and female users in sending messages in a day. Here the strata are male and female users.

b) Efficacy of a drug among different age groups. Age group can be classified into categories such as less than 40, between 41 and 60, and over 60 years of age.

c) Performance of children in school and the parents' marital status. Here, marital status can be (a) Married, (b) Divorced. In this case we assume that the parent's marital status may influence children's academic performance.

d) Television rating points for a program across different geographical regions of a country. For India, geographical regions could be different states of the country.

Steps in creating stratified sample

a) Identify the factor that can be used for creating strata (for example: factor = Age; Strata 1: age less than 40; Strata 2: age between 41 and 60; and Strata 3: Age more than 60).

b) Calculate the proportion of each stratum in the population (say p_1, p_2, and p_3 for three strata identified in step 1).

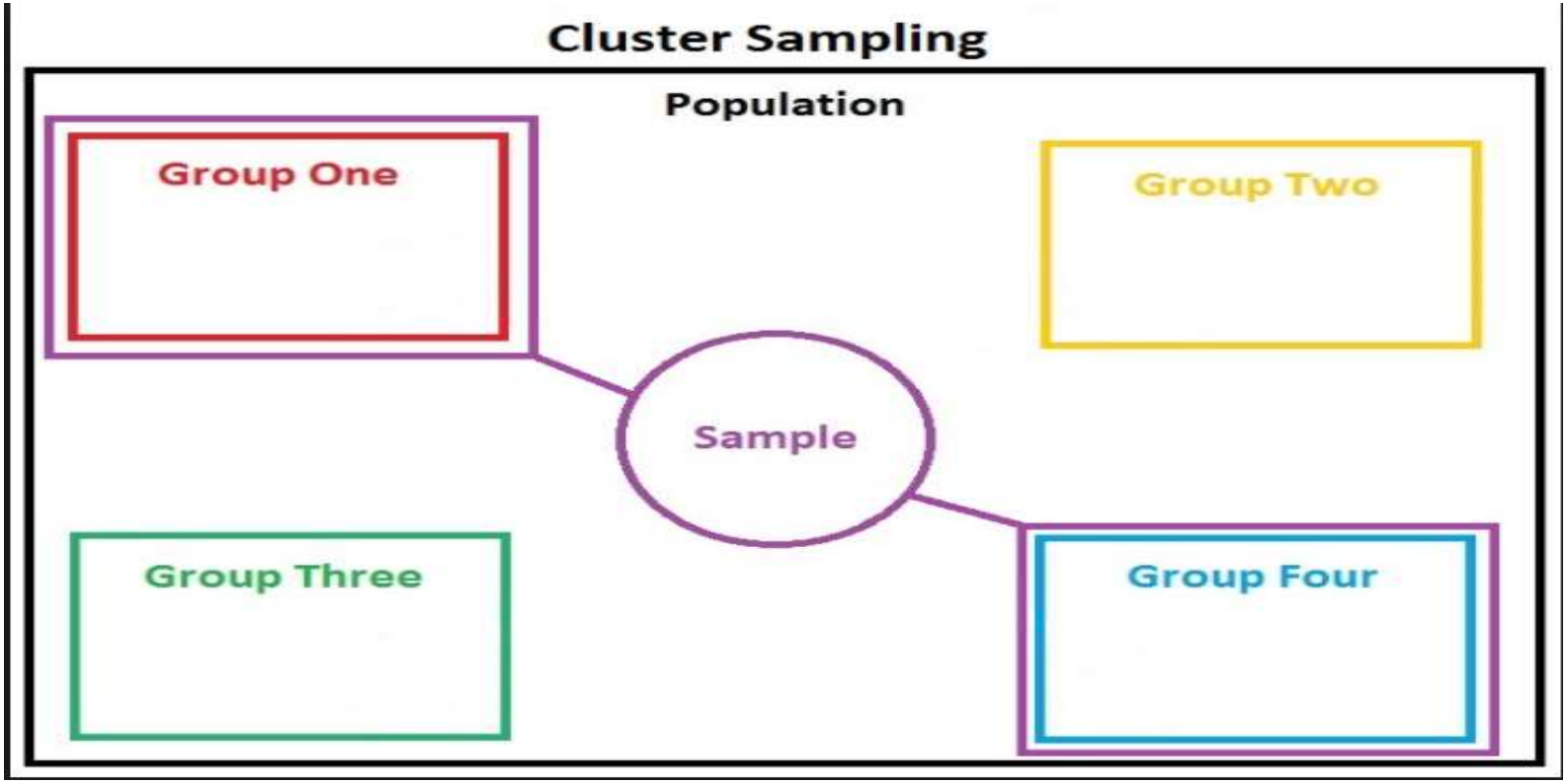
c) Calculate the sample size (say N). The sample size for strata 1, 2, and 3 identified in step 2 are $p_1 \times N$, $p_2 \times N$, and $p_3 \times N$, respectively.

d) Use random sampling procedure explained to generate random samples in each strata.

e) Combine samples from each stratum to create the final sample.

Cluster Sampling

- In cluster sampling, the population is divided into mutually exclusive clusters



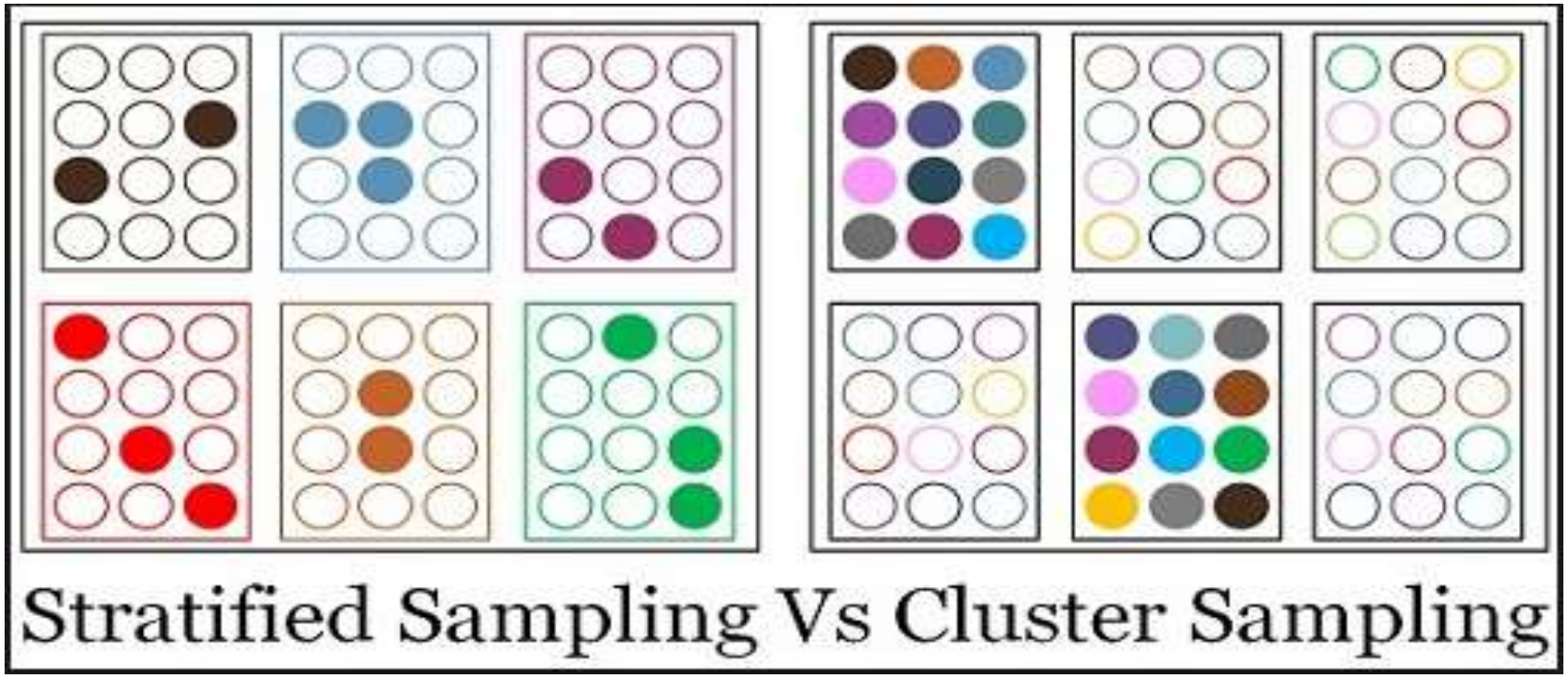
Cluster Sampling - Steps

- Identify the clusters (example: different models of smart phones sold by a manufacturer, customers from different geographical locations).
- Using random sampling select the clusters.
- Select all units in the clusters selected in step 2 and form the sample. If the size is too large, a random sampling within the clusters identified in step 2 may be used for final sample.

- Stratified sampling and cluster sampling are similar; the major difference is that in a stratified sample, all strata will be represented in the sample, whereas in a cluster sampling, not all clusters will be represented

Stratified Sampling Vs Cluster Sampling

- Stratified sampling and cluster sampling are similar; the major difference is that in a stratified sample, all strata will be represented in the sample, whereas in a cluster sampling, not all clusters will be represented



Bootstrap Aggregating (Bagging)

- Bootstrap Aggregating (known as Bagging) is sampling with replacement used in machine learning algorithms, especially the random forest algorithm (Breiman, 1996)

- The size of each sample and the number of samples are determined based on factors such as population size, target accuracy of the model developed using bagging and convergence, etc

- Bagging is frequently used in ensemble methods (in which several models are developed and the final prediction is usually based on the majority voting)

Non-Probability Sampling

- In a non-probability sampling, the selection of sample units from the population does not follow any probability distribution
- Sample units are selected based on convenience and/or on voluntary basis.