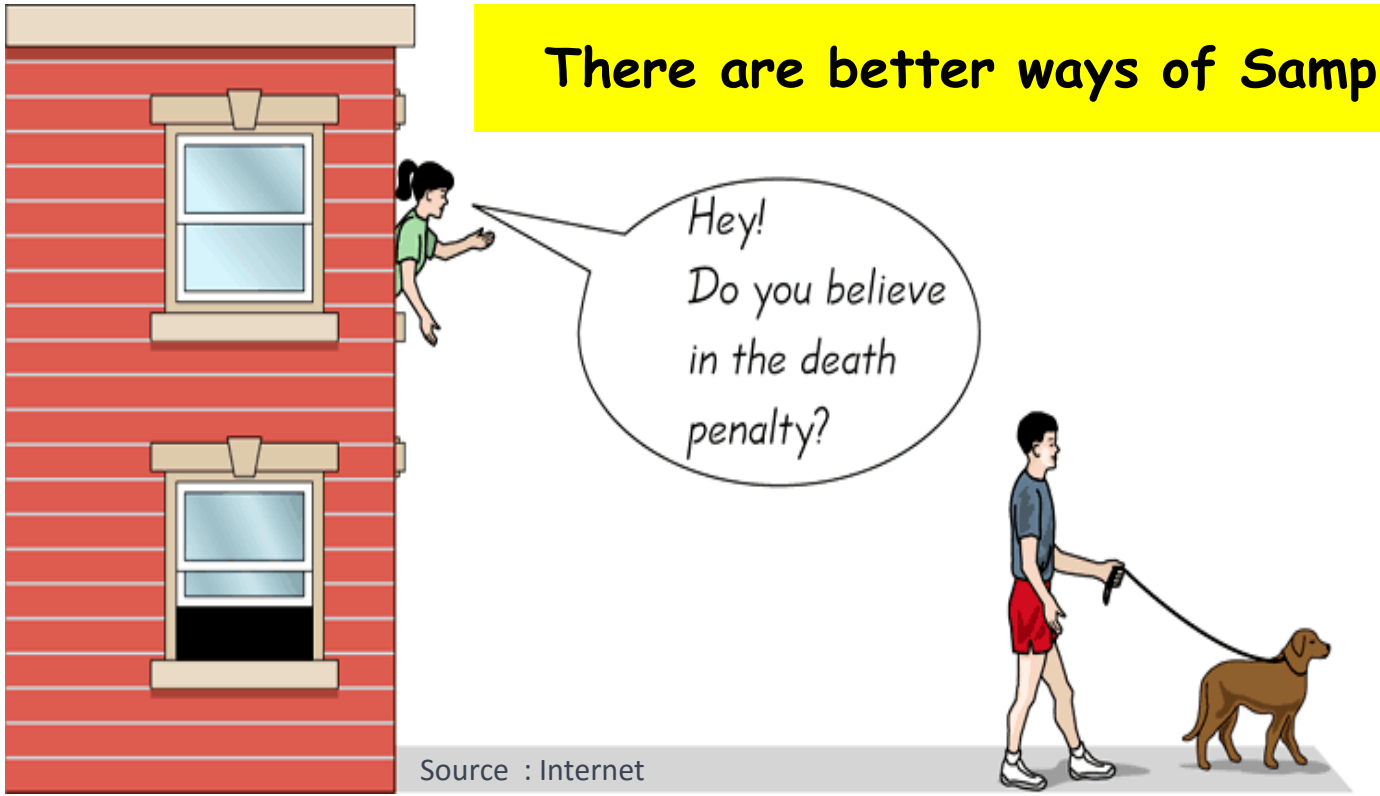
Statistical Sampling

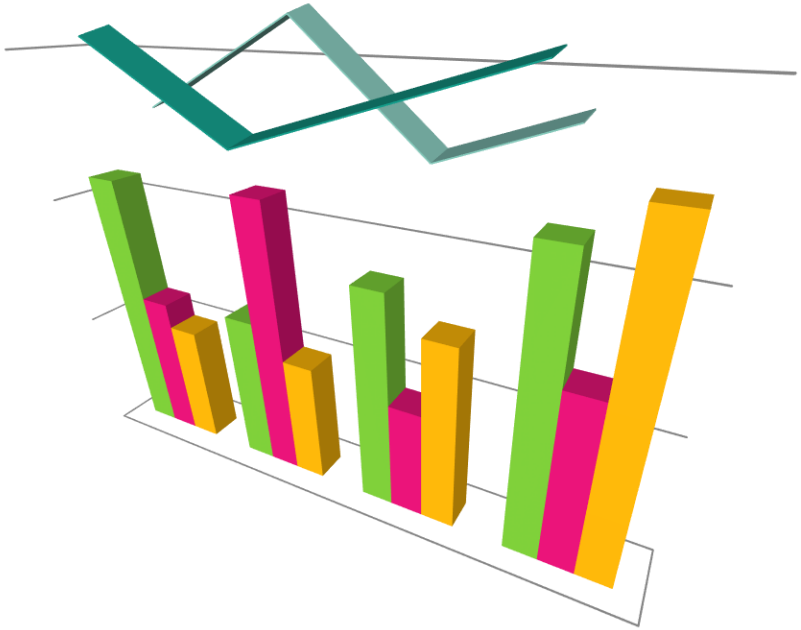
Amit Bhola

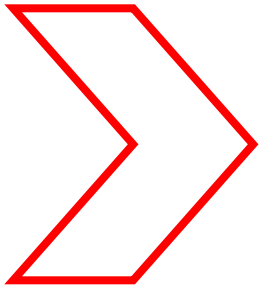
There are better ways of Sampling!

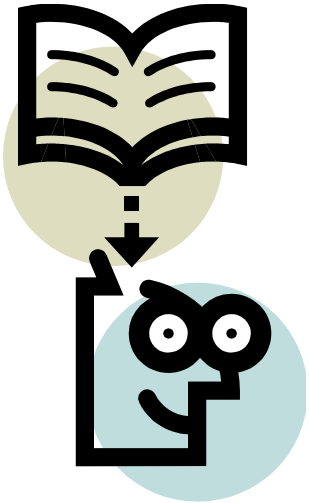


Source : Internet

GENERALIZATION







Representative Sample

Faithful Generalization

POPULATION

Population is '<u>observations</u>' or '<u>measurements</u>' of property under consideration

DO not confuse this with No. Of Objects.
(Objects may or may not be property under study)

	<u>Population</u>
(a) 'Height' of Plants in a garden	'Height'
(b) 'No. of Plants' in a garden	'No. of plants'

SAMPLE

A small part of a population

SAMPLING

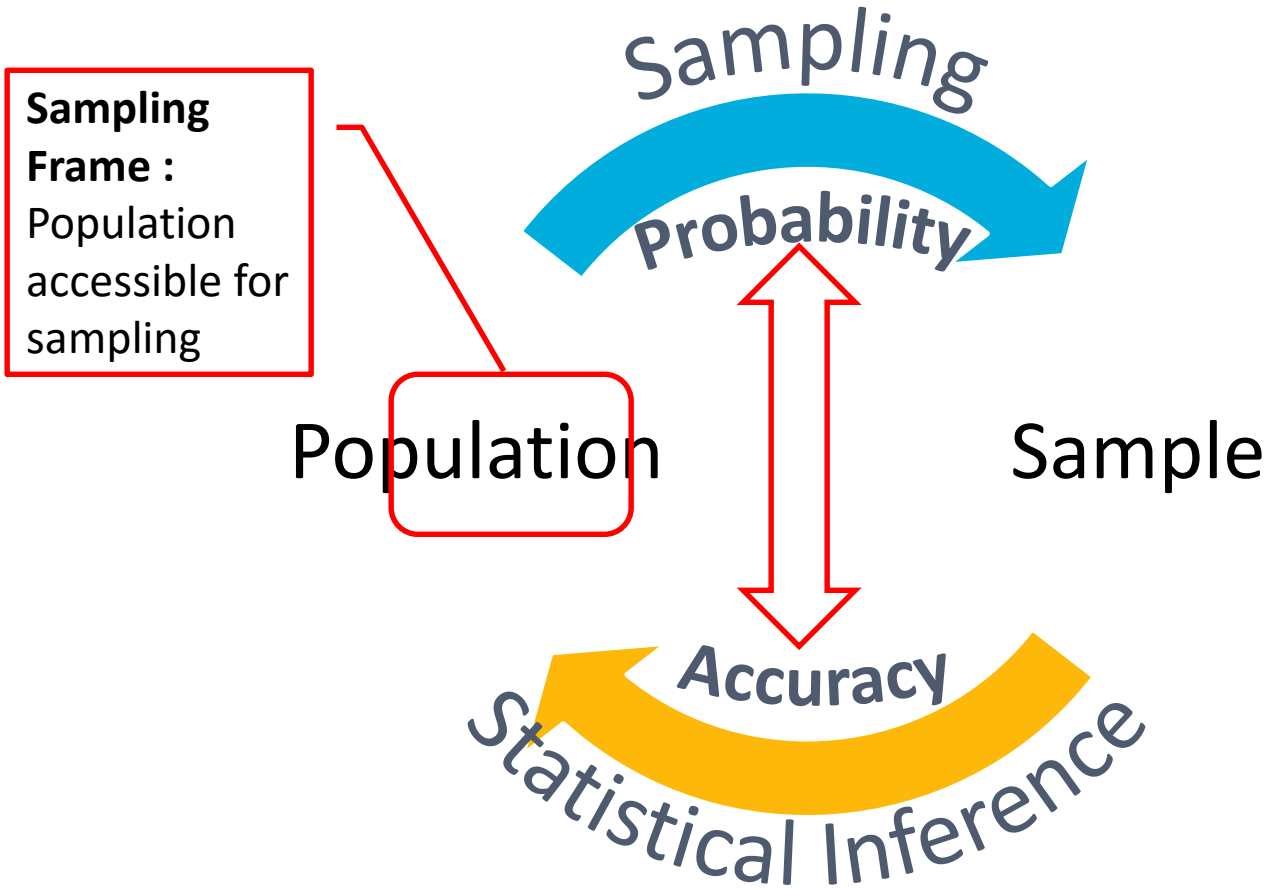
Process of obtaining samples

STATISTICAL INFERENCE

Aim of Sampling

Process of inferring facts about population from the results found in samples

POPULATION, SAMPLE & GENERALIZATION



SAMPLING METHODS & GENERALIZATION

PROBABILITY SAMPLING – Aim of Generalization

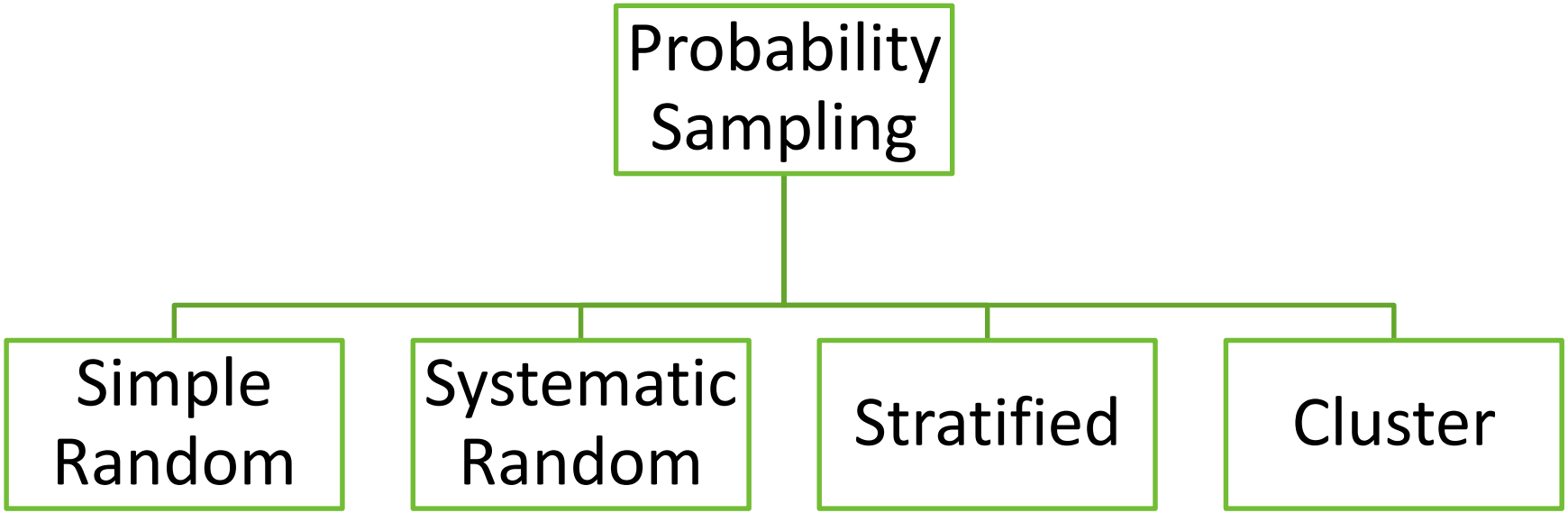
Focus

- Best effort is made to draw sample representative of population

NON PROBABILITY SAMPLING – When?

- Generalization is not the aim
  - Qualitative study, Pilot study, Demonstration of population trait
- Probability sampling is infeasible
  - Inaccessible sampling frame, constraints of time, money, etc.
  - Initial study to be followed by probability sampling

TYPES OF PROBABILITY SAMPLING



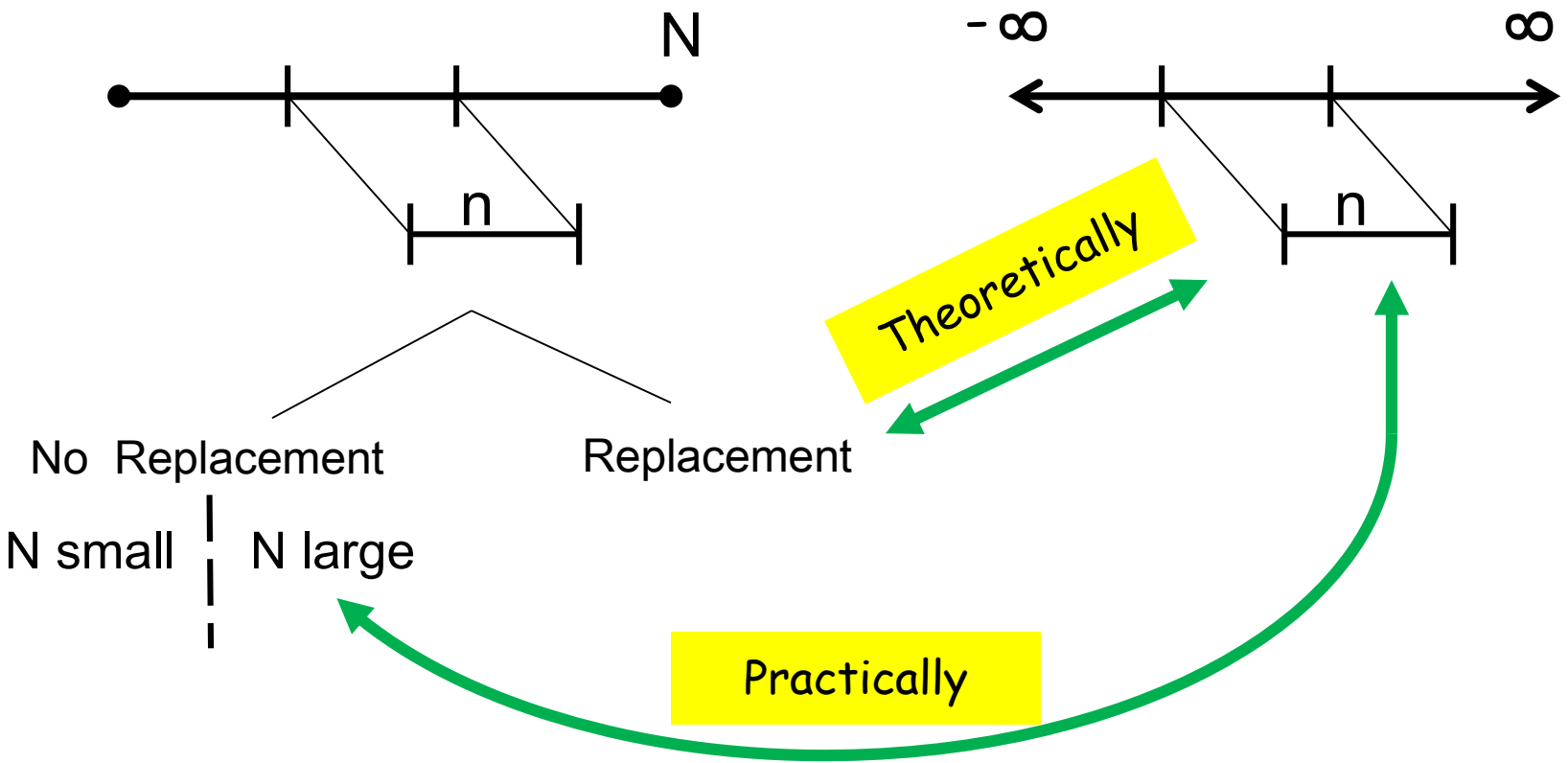
POPULATIONS & SAMPLE SIZE

EXAMPLE	Population	Sample
Estimate Avg. weight of college students by studying only 100	Finite N	Finite n
Estimate Head or Tail of a coin toss	Infinite ∞	Finite n

SAMPLING WITH REPLACEMENT

EXAMPLE	Replacement
Estimate Avg. weight of college students by studying only 100 different students	No
Estimate how many bolts in a bin of 500 are defective by : picking one bolt → checking it → returning it → …. (repeat say 20 times)	Yes

SAMPLING THEORY



RANDOMNESS

Being Random means being equally probable

A sample is random if there is **no bias** in selecting its n objects →
Each object has equal chance of getting selected

To effectively represent a population, a sample should be random

For getting a random sample of size n,
n random nos. should be obtained first

RANDOM NUMBER – MS EXCEL

=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(2,5)
=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(2,5)
=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(2,5)
=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(2,5)
=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(9,11)	=RANDBETWEEN(2,5)

10	9	9	5
10	11	11	2
11	11	10	5
9	10	10	5
9	9	10	4

BLINDING

Blinding is done to eliminate **psychological bias**

- **Single Blinding** : The participants (i.e. sample) are completely unaware of which group they are in and what intervention they are receiving until conclusion of the study.

- **Double Blinding** : Neither the participants nor the researcher knows to which group the participant belongs and what intervention the participant is receiving until the conclusion of study

SIMPLE RANDOM SAMPLING

STEP 1 : Obtain the approx size of population N

STEP 2 : Label the population items 1,2,… N

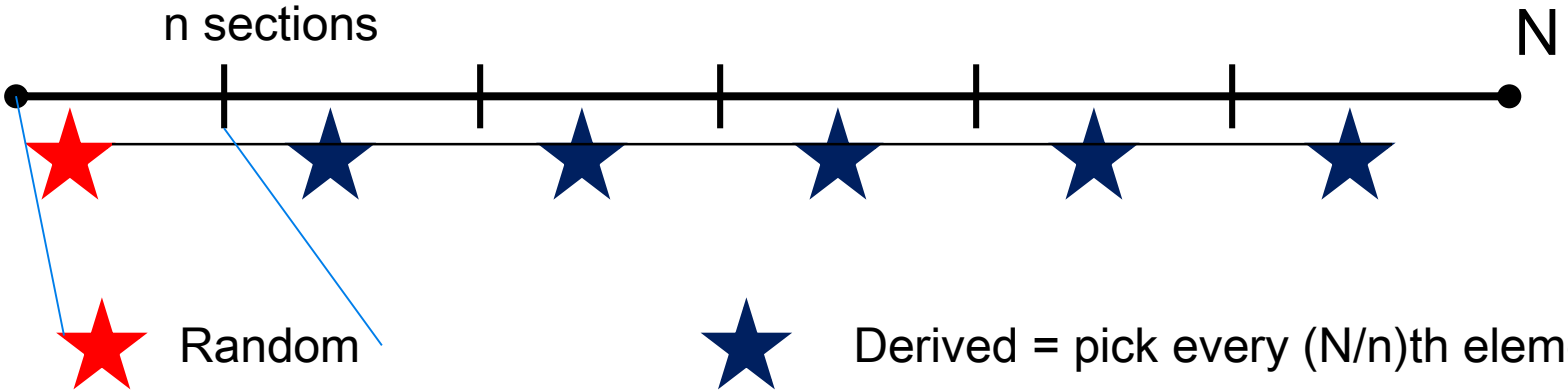
STEP 3 : Find n random nos.

STEP 4 : Select items labeled as nos. got in [3]

SIMPLE RANDOM SAMPLING

STEP 2 : Label the population items 1,2,… N

SYSTEMATIC RANDOM SAMPLING

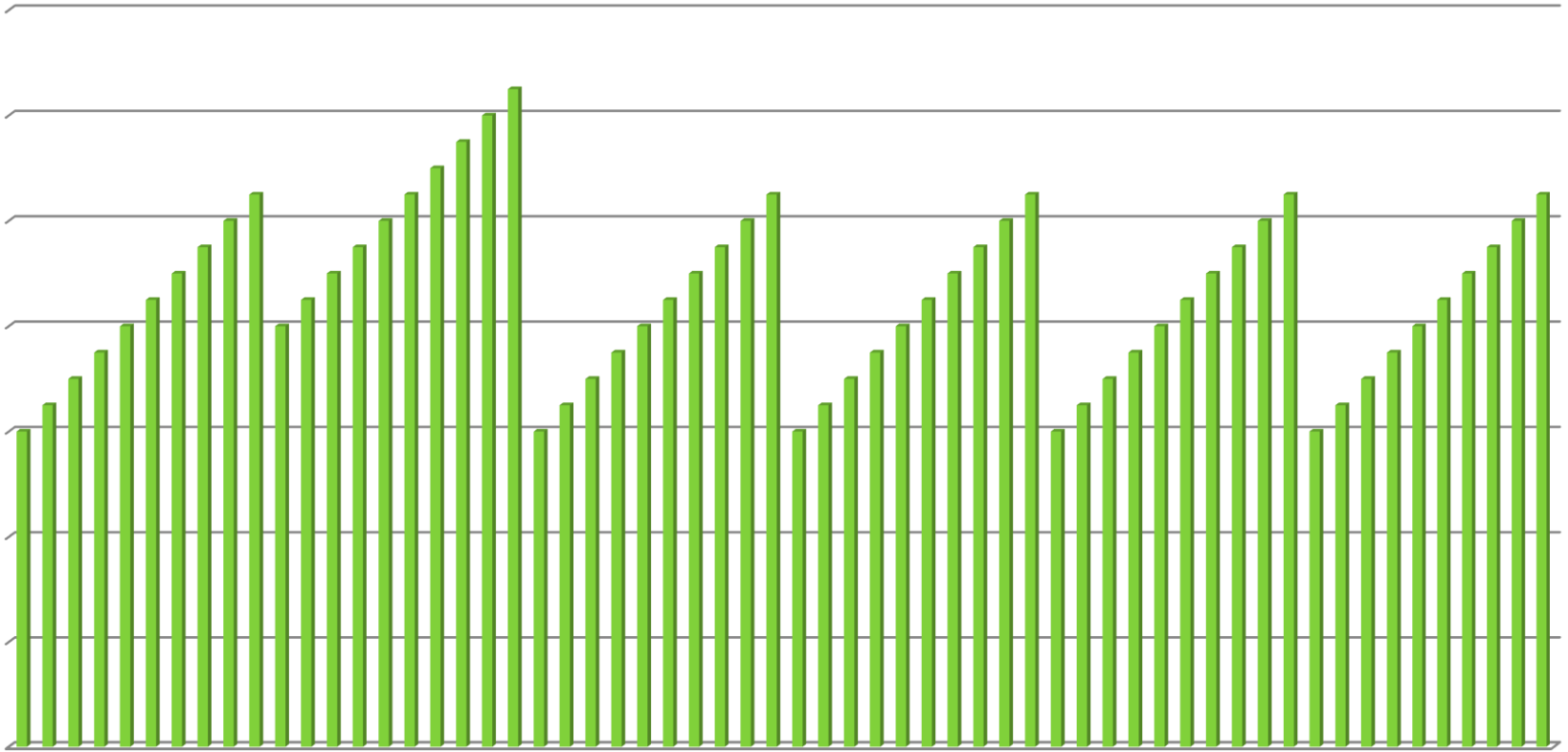


PROBLEM WITH SIMPLE / SYSTEMATIC

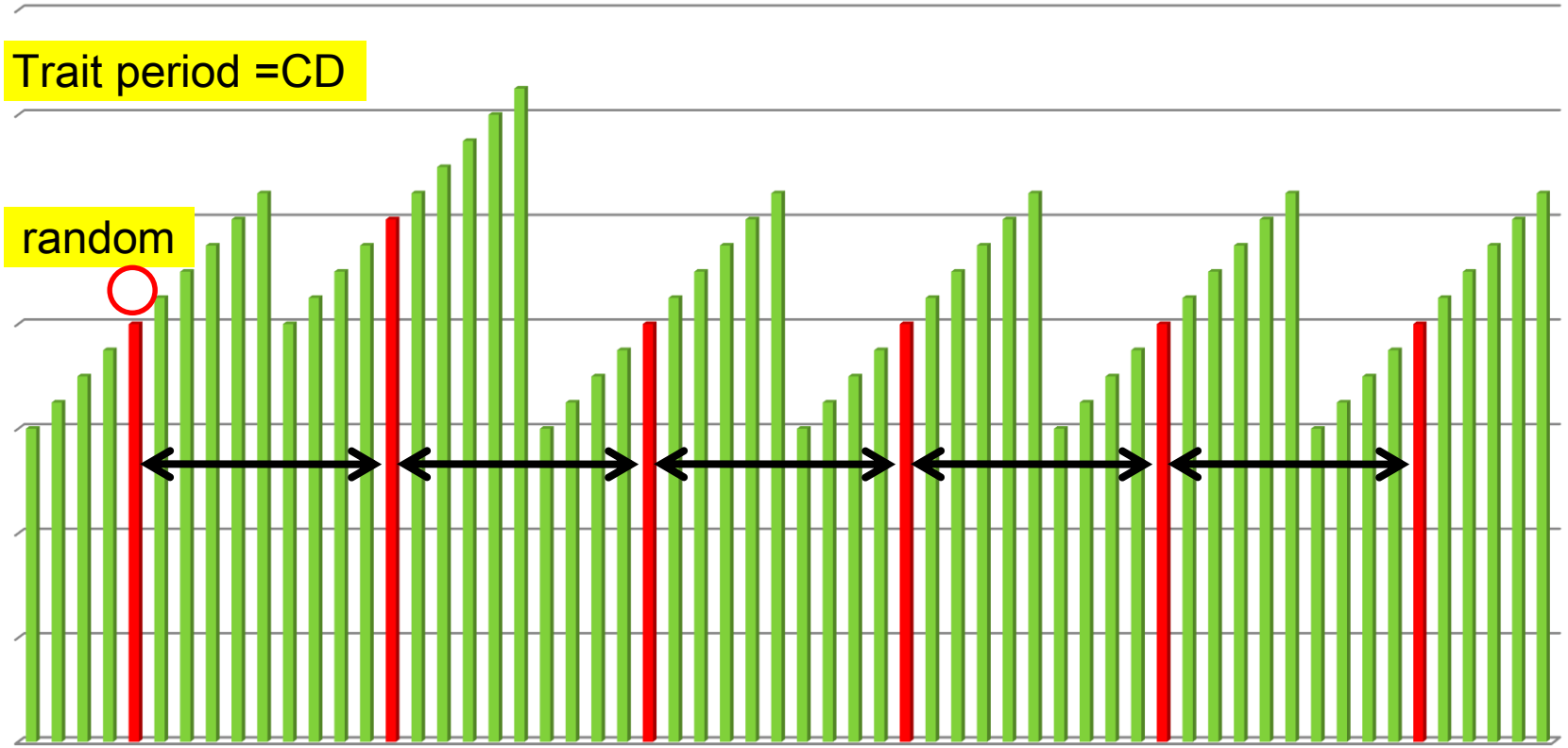
1. Need availability of complete list of population. For a large population, this may not be available! Sampling Frame Errors…

2. Although highly unlikely, Systematic sampling carries risk of collecting a poor sample if (A) there exist some periodic traits in the population, and at the same time (B) The period of the trait is a multiple of common difference!

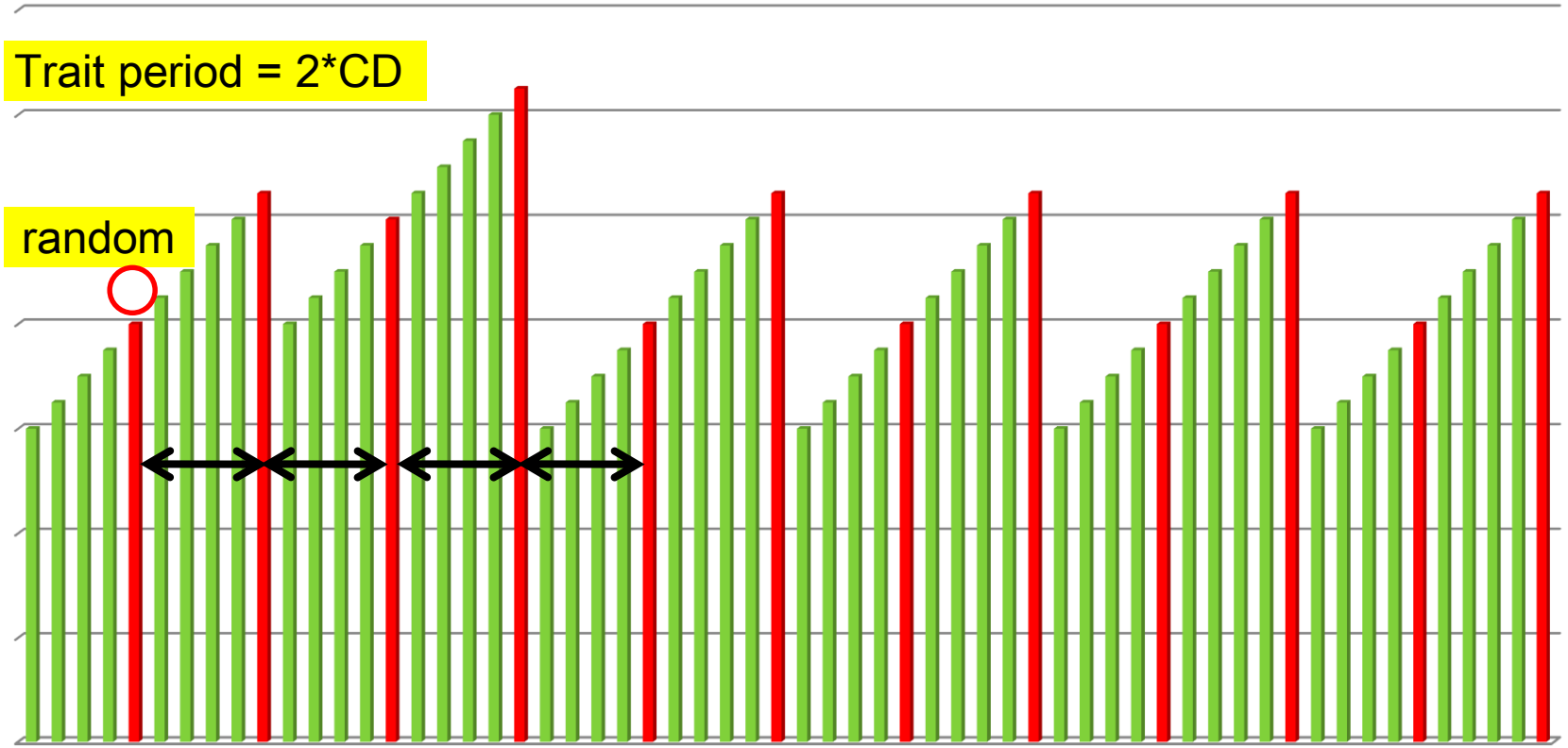
PROBLEM WITH SYSTEMATIC SAMPLING



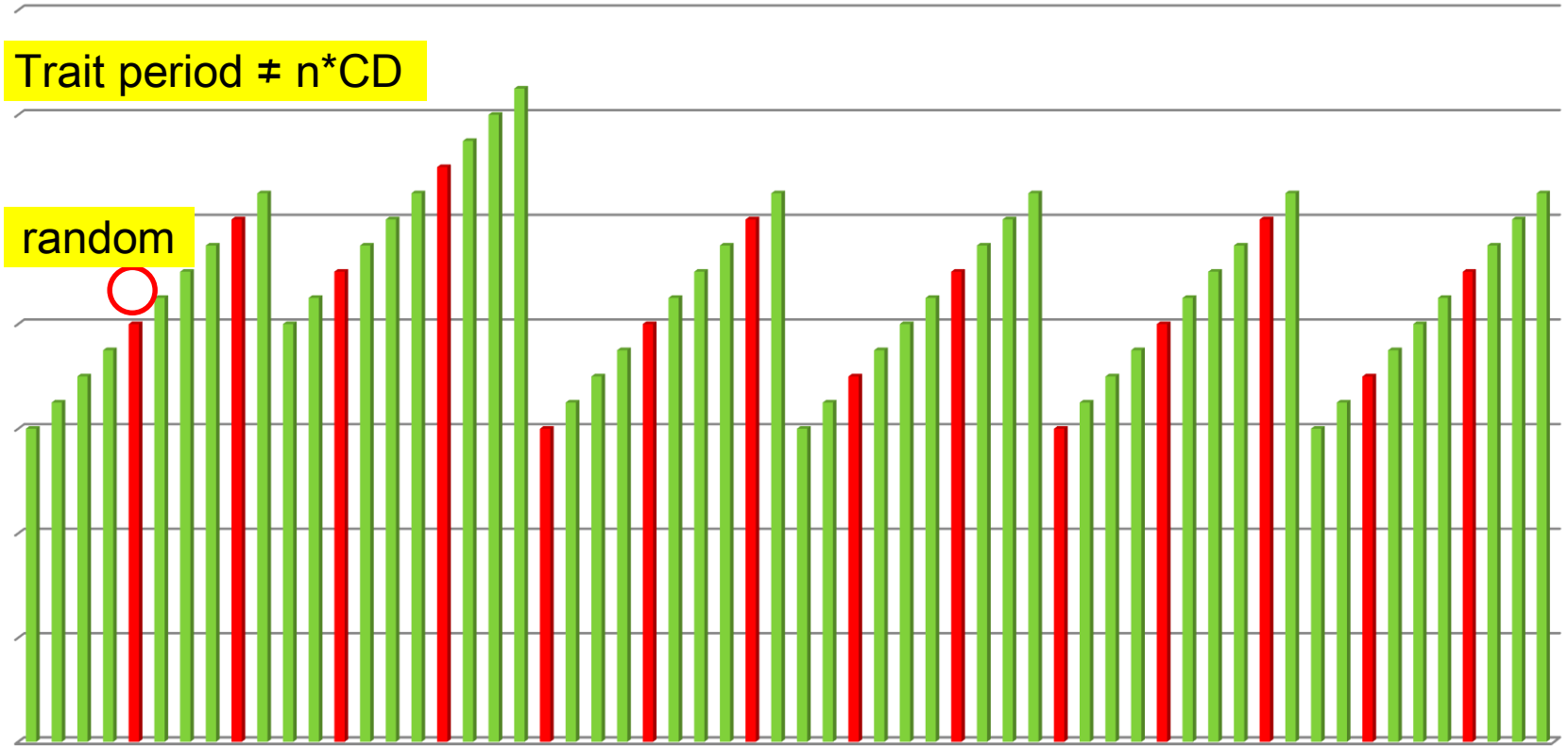
PROBLEM WITH SYSTEMATIC SAMPLING



PROBLEM WITH SYSTEMATIC SAMPLING

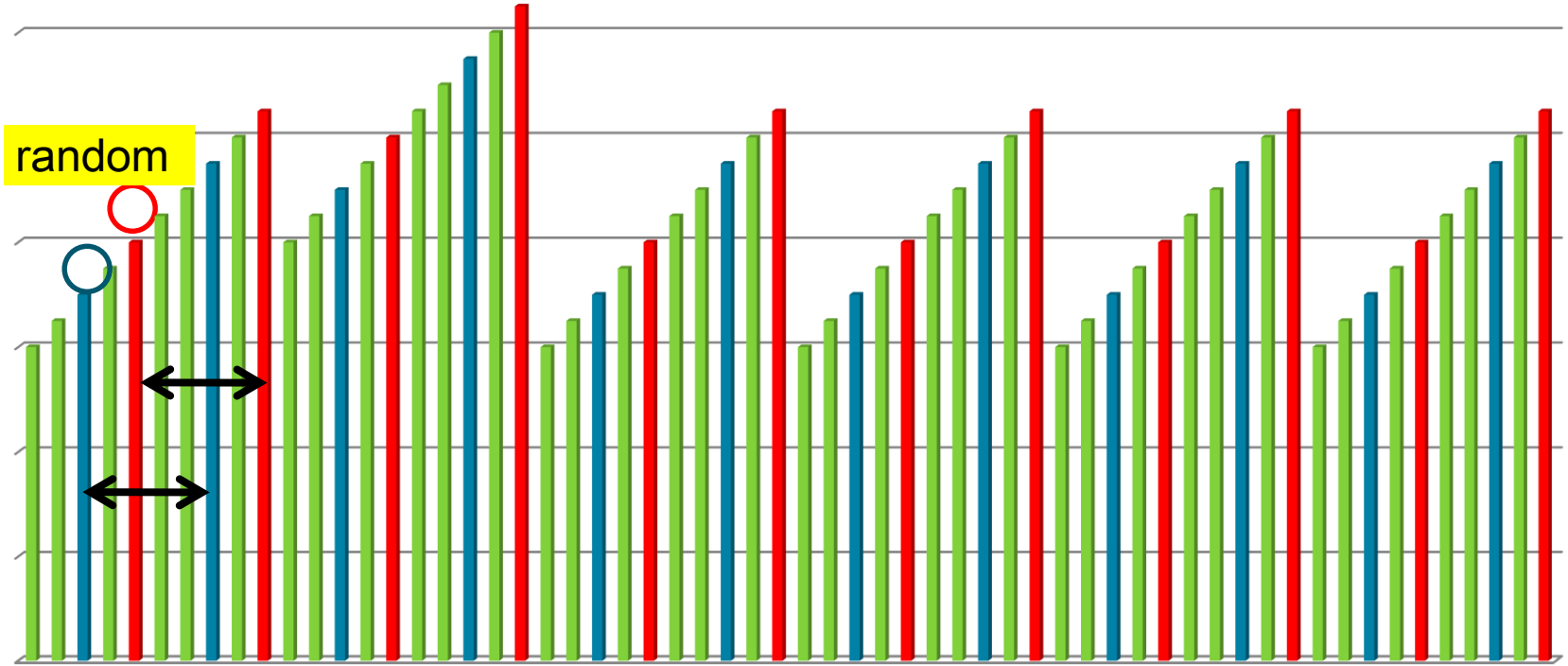


DEALING PERIODICITY PROBLEM – 1

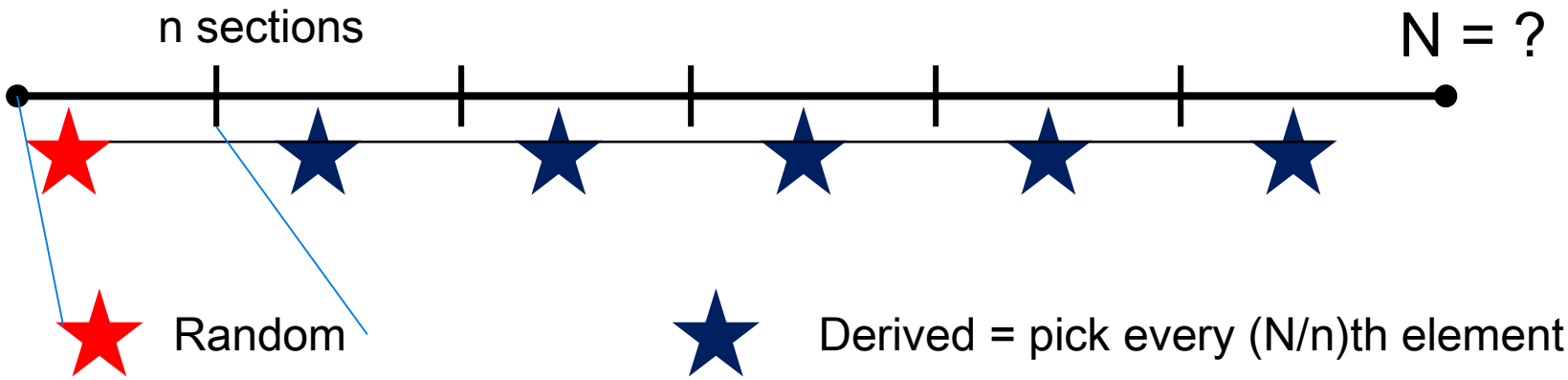


DEALING PERIODICITY PROBLEM – 2

Repeated sampling and combining two samples into one single sample



LINEAR SYSTEMATIC RANDOM SAMPLING

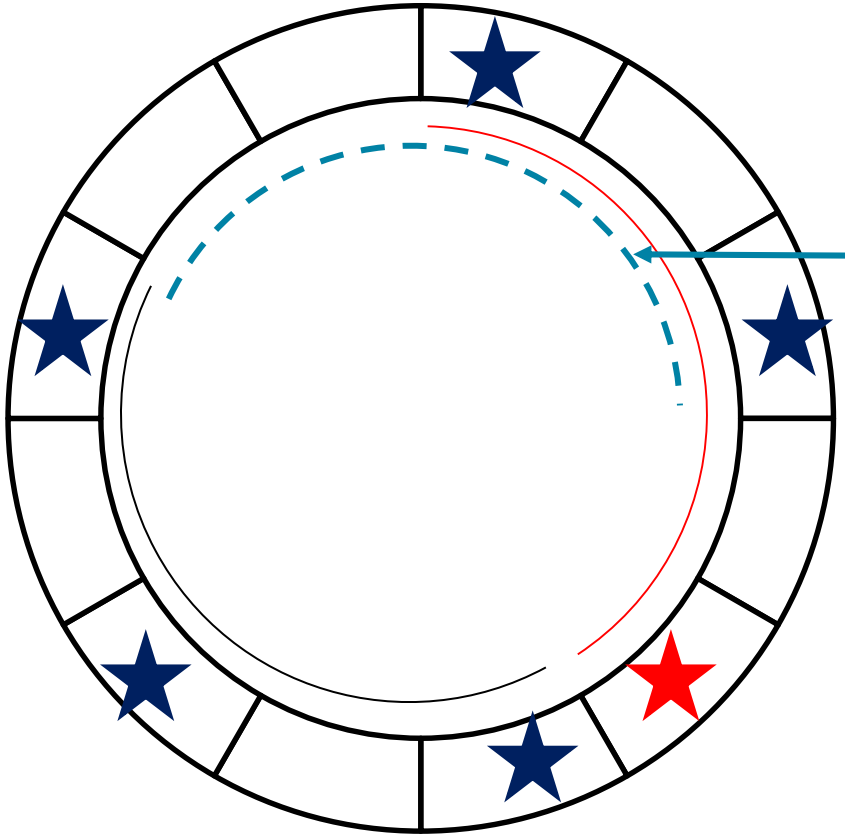


First Random is Chosen from 1st section (1 ~ N/n)

CIRCULAR SYSTEMATIC RANDOM SAMPLING

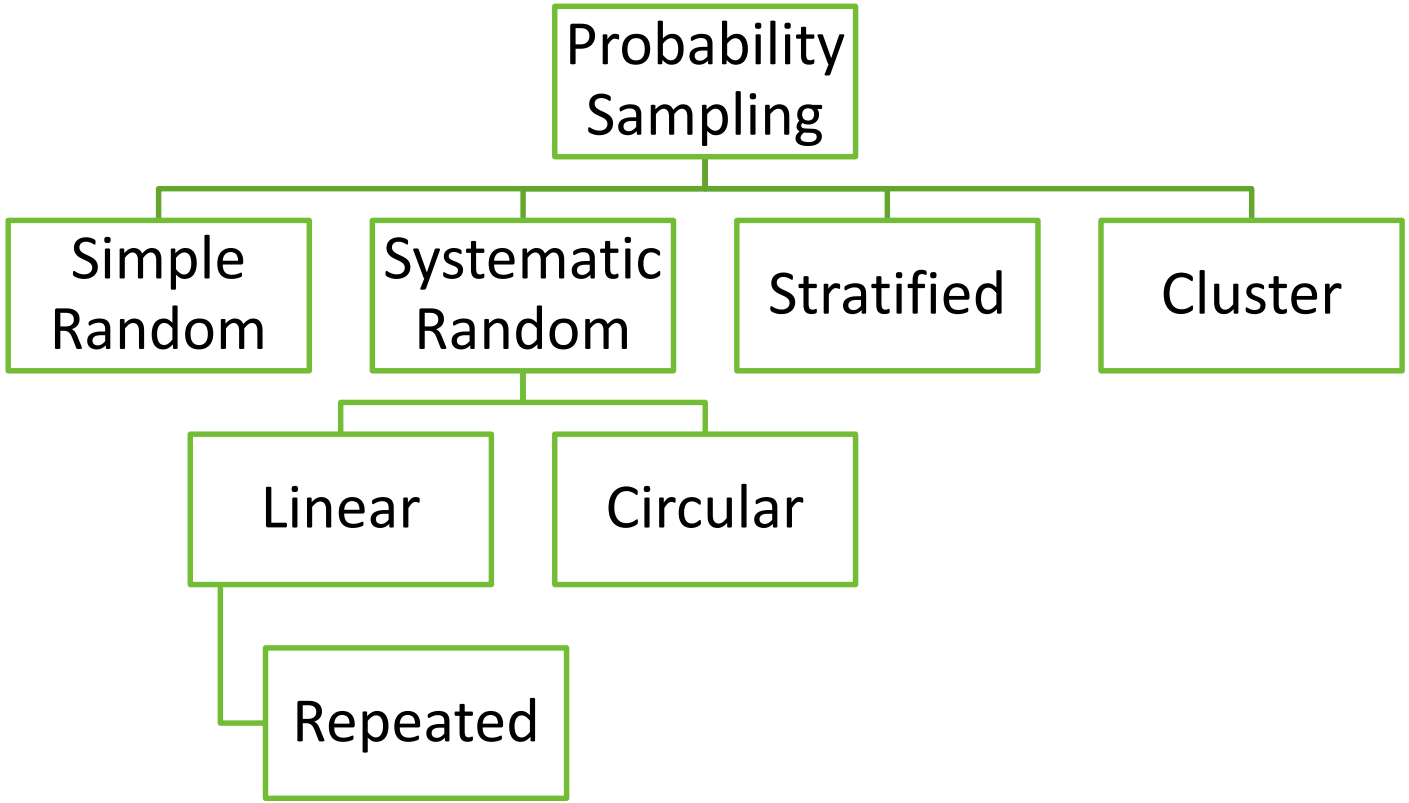
Random from 1 ~ N

Derived till n samples are obtained

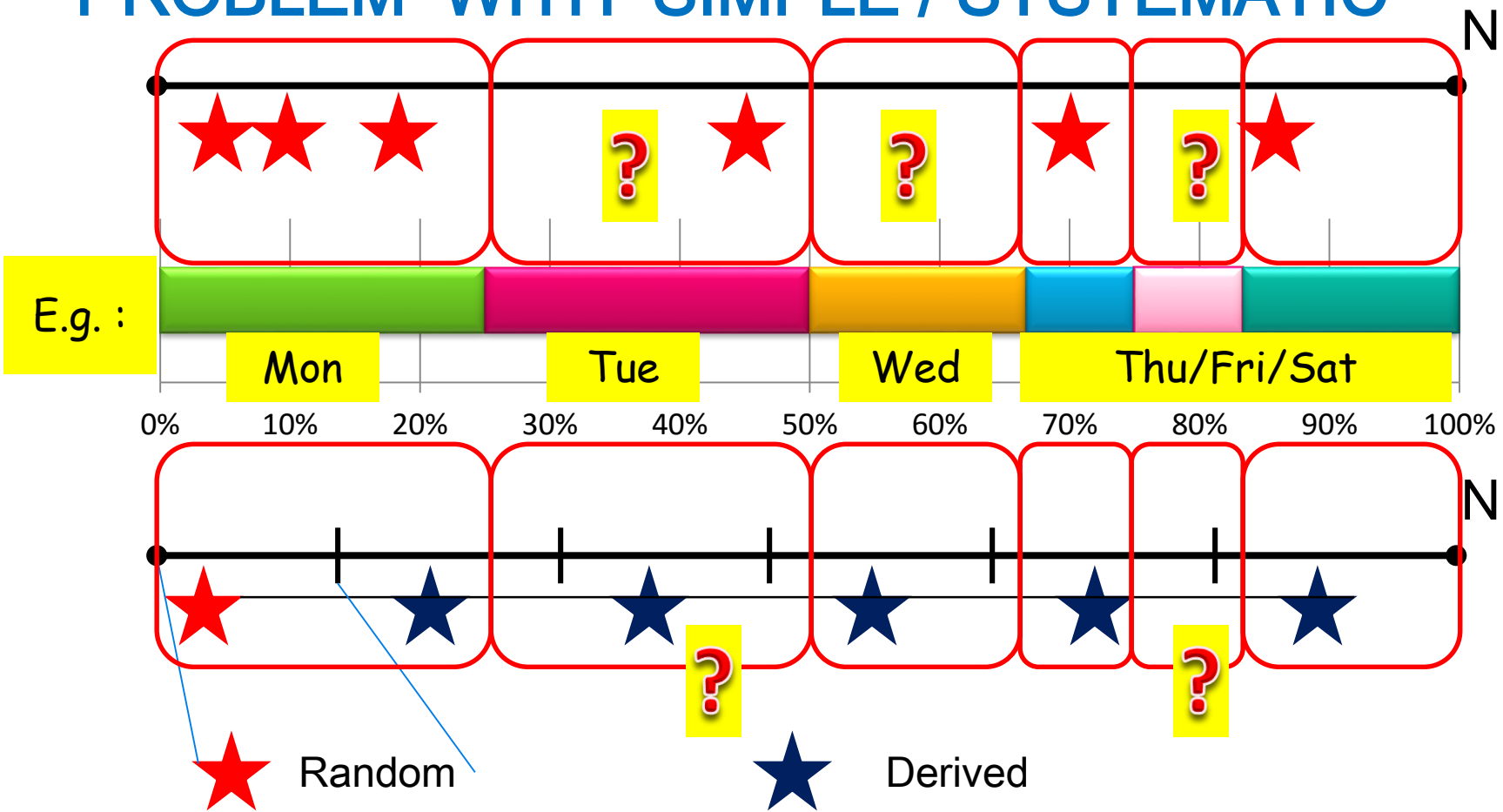


Selection is done with continuing counting at the end of the list

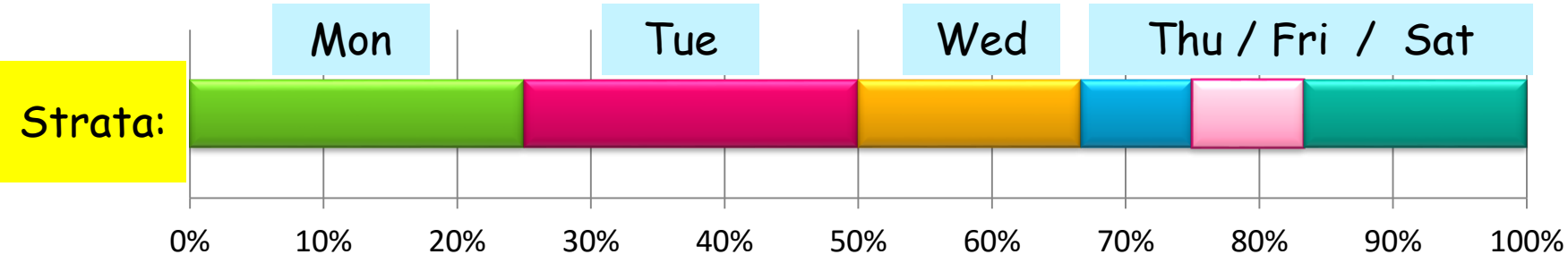
TYPES OF PROBABILITY SAMPLING



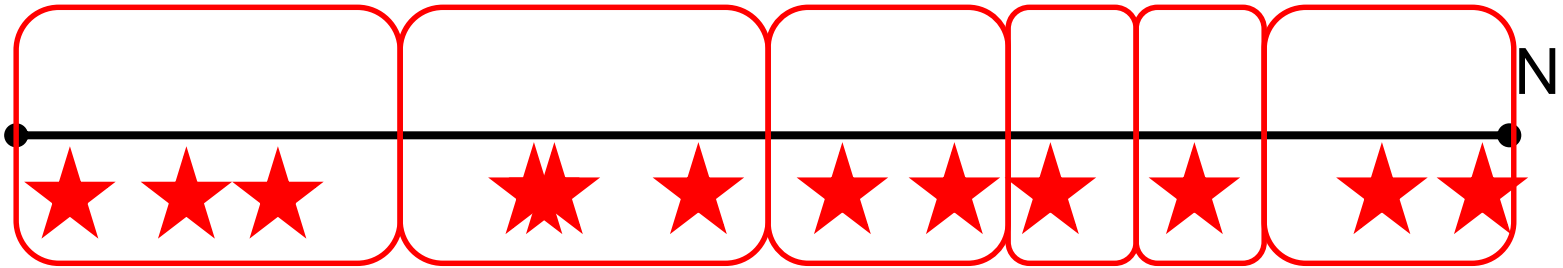
PROBLEM WITH SIMPLE / SYSTEMATIC



STRATIFIED SAMPLING

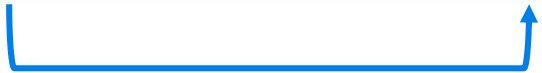


1. Divide N by category of Strata
2. Select random samples from each Strata – in same ratio as that Strata



STRATIFIED SAMPLING

Region	Population		Proportionate Stratified Sample	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	396	33%
District 2	600	1%	12	1%
District 3	12000	22%	264	22%
District 4	24000	44%	528	44%
Total	54600	100%	1200	100%



Proportionate
=>
Representative of Population

STRATIFIED SAMPLING – ADVANTAGES

1. Ensures the presence of each subgroup within the sample – better *representation* of population.

Especially useful for population with highly skewed strata eg. A:B::70:30

2. Permits analyses of within-stratum patterns and separate reporting of the results for each stratum.

STRATIFIED SAMPLING – DIFFICULTIES

1. Requires information on the proportion of the total population that belongs to each stratum.

2. More expensive, time-consuming, and complicated than simple random sampling.

3. In order to calculate sampling estimates, at least two elements must be taken in each stratum.

DISPROPORTIONATE STRATIFIED SAMPLING

Region	Population		Proportionate Stratified Sample	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	396	33%
District 2	600	1%	12	1%
District 3	12000	22%	264	22%
District 4	24000	44%	528	44%
Total	54600	100%	1200	100%

But considering the cost of the sampling, studies other than the Generalization may separately be performed at same time

Proportionate
=>
Representative of Population – Fine

DISPROPORTIONATE STRATIFIED SAMPLING

Region	Population		Disproportionate Stratified Sample	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	357	30%
District 2	600	1%	130	11%
District 3	12000	22%	238	20%
District 4	24000	44%	475	39%
Total	54600	100%	1200	100%

Example 1 :-
Analysis of variation **within** strata

Dis-Proportionate
to
Represent all strata **sufficiently**

DISPROPORTIONATE STRATIFIED SAMPLING

Region	Population		Disproportionate Stratified Sample Using Equal Allocation	
	Frequency	Percent	Frequency	Percent
District 1	18000	33%	300	25%
District 2	600	1%	300	25%
District 3	12000	22%	300	25%
District 4	24000	44%	300	25%
Total	54600	100%	1200	100%

Example 2 :-
Analysis of variation **among** strata

Dis-Proportionate
to
Represent all strata **equally**

DISPROPORTIONATE STRATIFIED SAMPLING

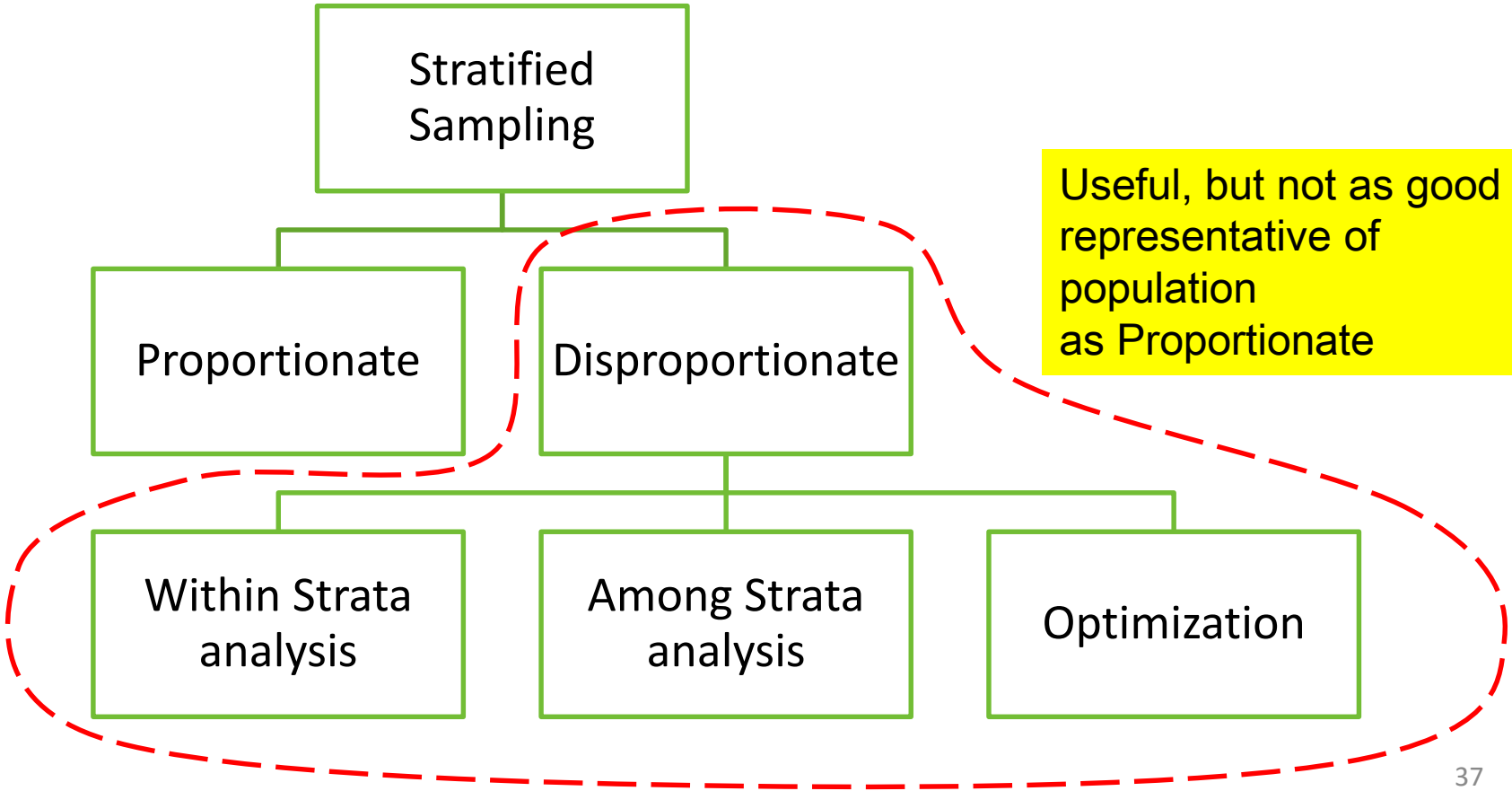
Regions (1)	Population Frequency (2)	Population Percent Distribution (3)	Data Collection Cost Per Unit (j) (4)	Variability (s) (5)	$\frac{s}{\sqrt{j}}$ (6)	Sample Size Optimizing Costs (7)	Sample Size Optimizing Variability (8)	Sample Size Optimizing Costs and Variability (9)
District 1	18000	33%	$18	4.3	1.014	300	190	203
District 2	600	1%	$10	6.4	2.024	538	282	405
District 3	12000	22%	$39	9.4	1.505	138	415	302
District 4	24000	44%	$24	7.1	1.449	224	313	290
Total	54600	100%				1200	1200	1200

Example 3 :-
Optimization of Cost and/or Precision

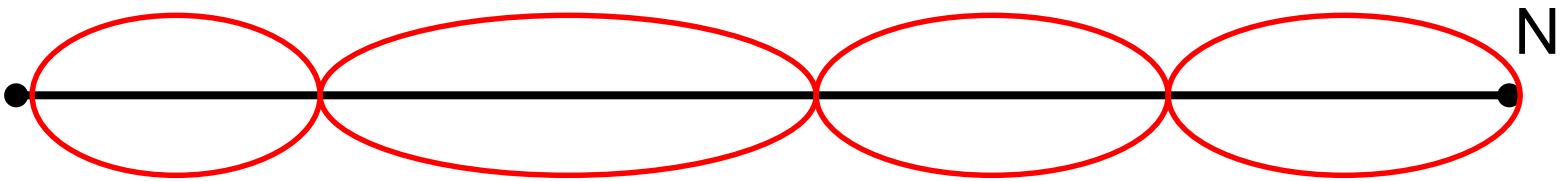
j	1/j	1200
18	0.055556	299.1371
10	0.1	538.4468
39	0.025641	138.0633
24	0.041667	224.3528

s	1200
4.3	189.7059
6.4	282.3529
9.4	414.7059
7.1	313.2353

TYPES OF STRATIFIED SAMPLING



CLUSTER SAMPLING

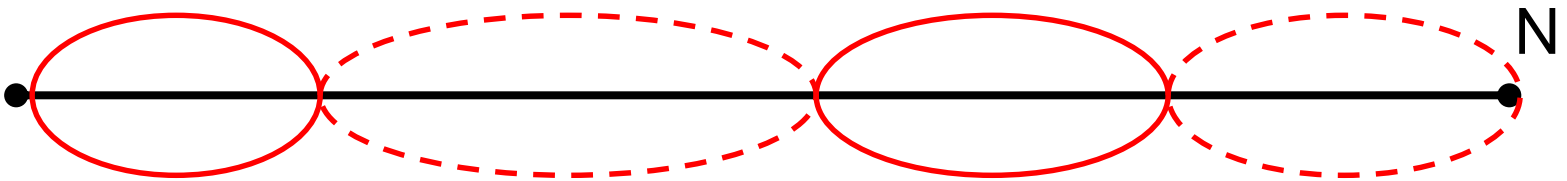


STEP 1 : Divide N into homogenous clusters

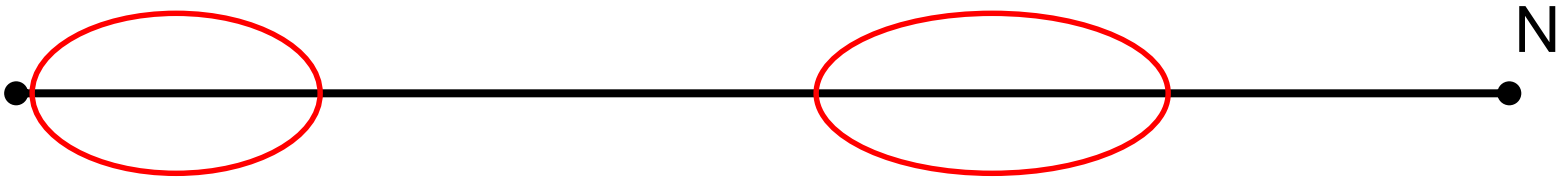
(Clusters : different from each other but same within)

Eg. Subjects , Districts , Offices-shops-showrooms

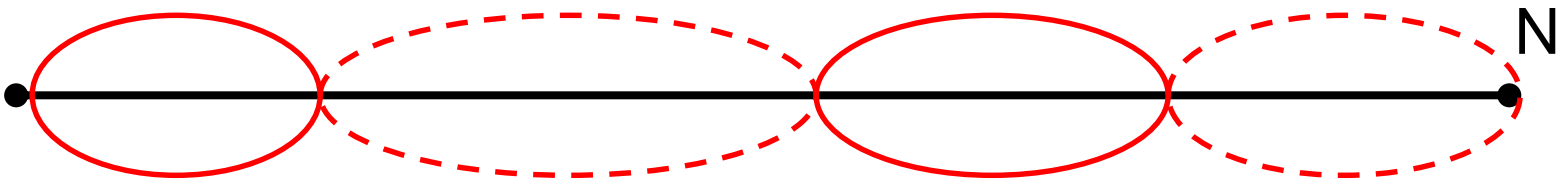
CLUSTER SAMPLING



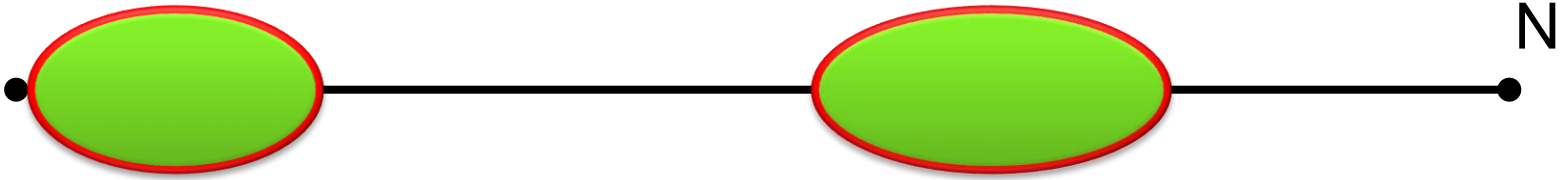
STEP 2 : Choose some clusters randomly



CLUSTER SAMPLING



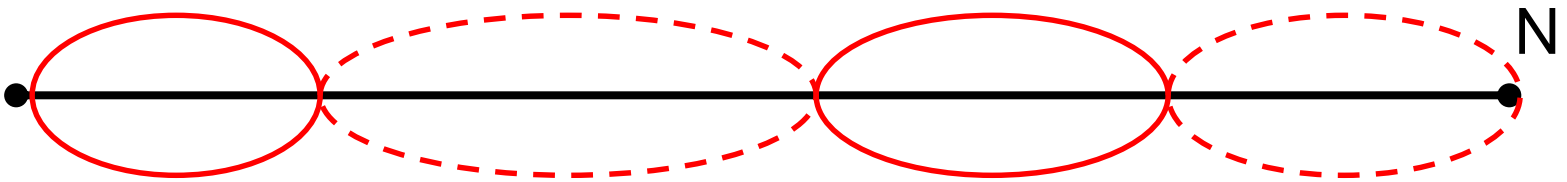
STEP 3 : Recommendation but not must, select whole of the selected clusters.



(STEP 4) : Sampling within clusters may be done

Multi-Stage Sampling…

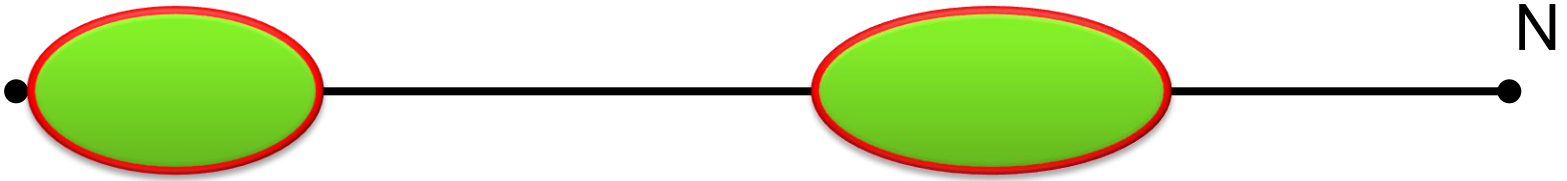
CLUSTER SAMPLING



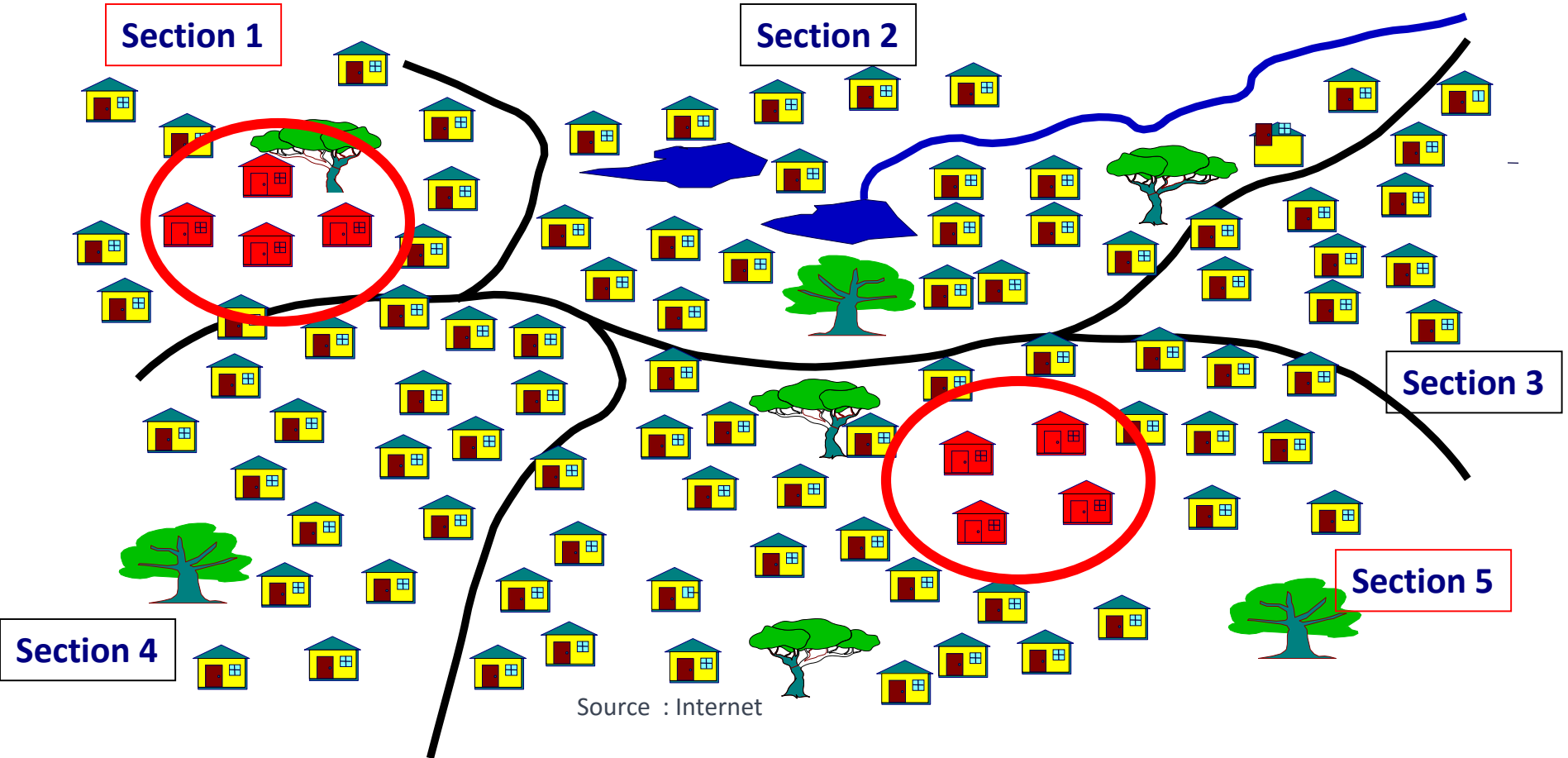
Usual motive is to avoid the high cost of a geographical survey

Accuracy is not as good as Stratified, but in comparison to No Survey at all (due to high cost), it is better to do Cluster Sampling

E.g. TRPs , IQS . . .



CLUSTER SAMPLING



Source : Internet

CLUSTER SAMPLING vs. STRATIFIED

1. In stratified random sampling, **all the strata** of the population are sampled while in cluster sampling, **only a part of** clusters are sampled.

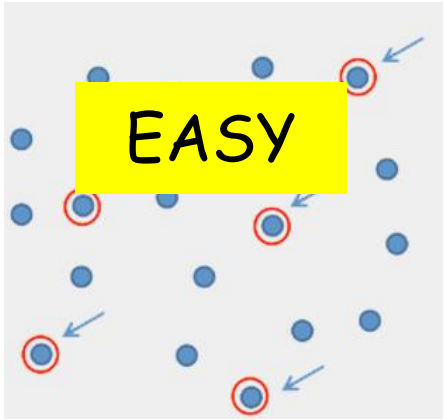
2 With stratified sampling, the best survey results occur when elements within strata are internally **homogeneous**. However, with cluster sampling, the best results occur when elements within clusters are internally **heterogeneous**.

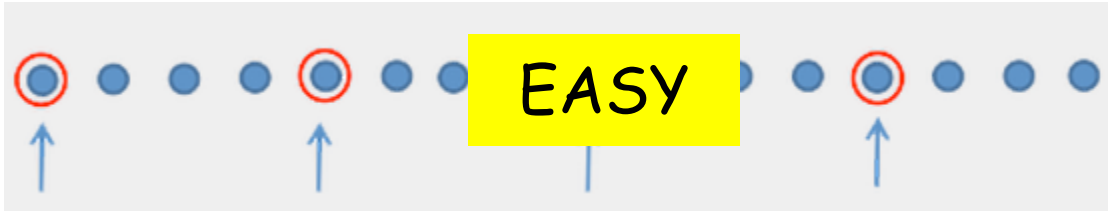
CLUSTER SAMPLING – DISADVANTAGES

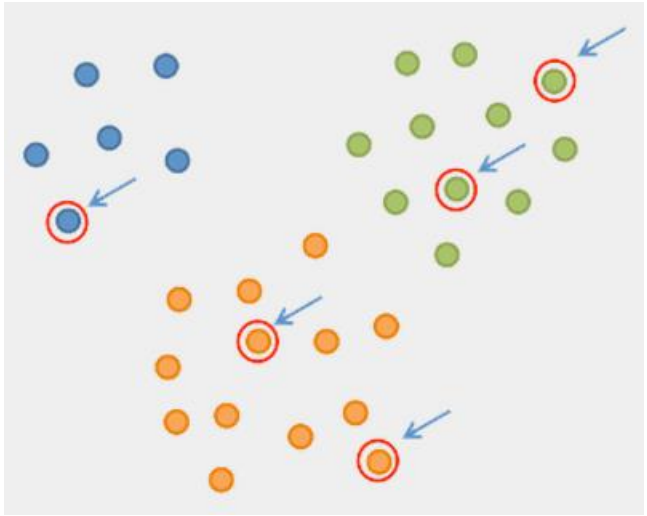
Statistically, it is the **least precise** compared to the Simple, Systematic and Stratified sampling.

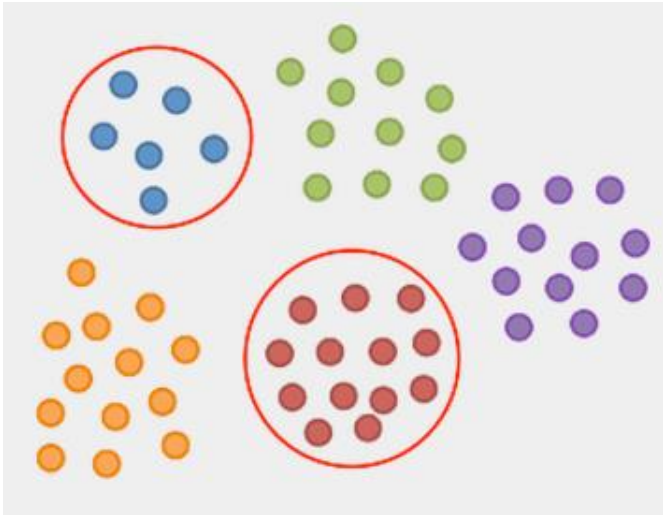
There is tendency for the clusters to display similar characteristics within themselves – especially so in case the clusters are regional

COMPARISON OF SAMPLING TECHNIQUES

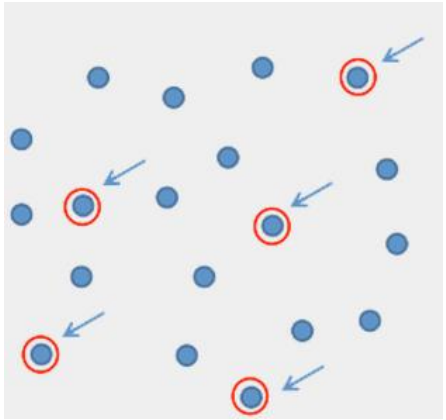


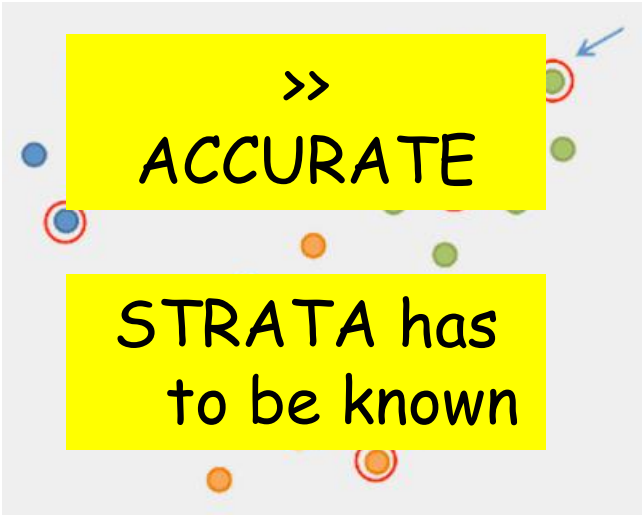


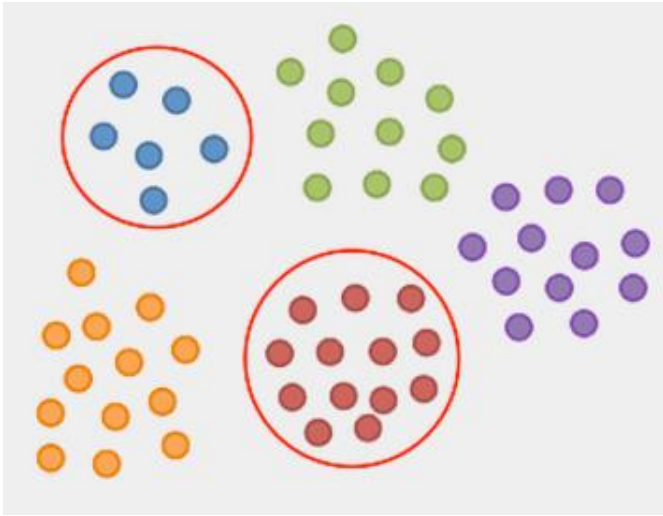




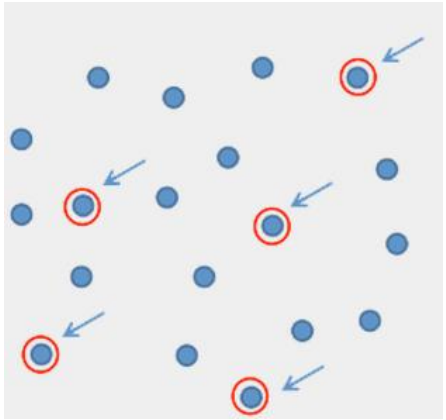
COMPARISON OF SAMPLING TECHNIQUES

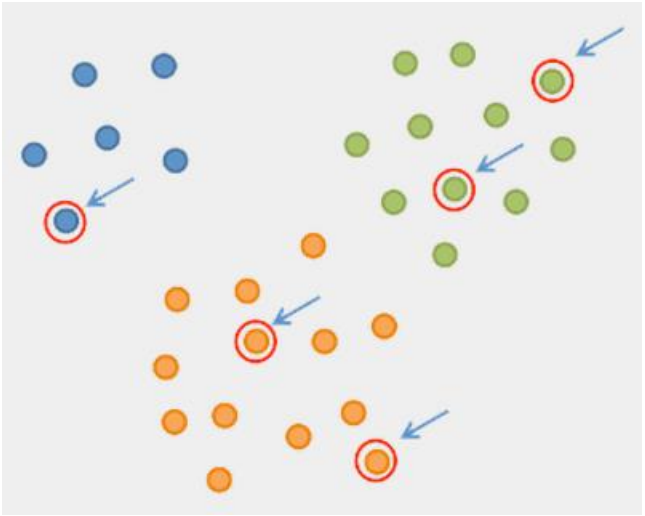


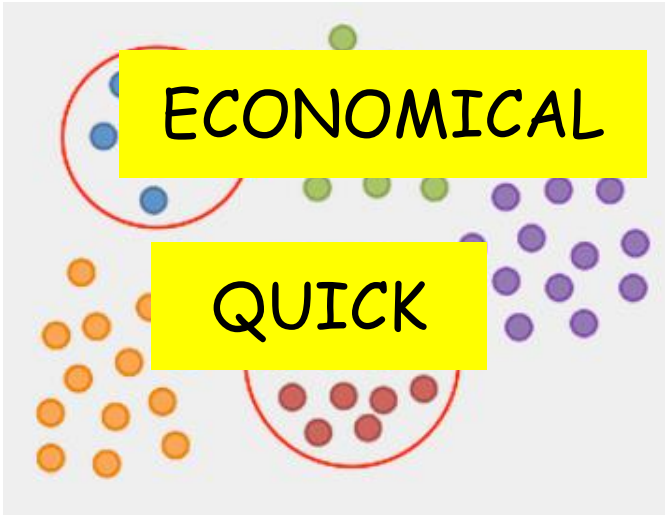




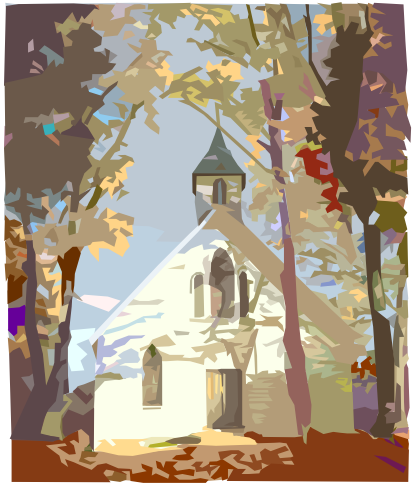
COMPARISON OF SAMPLING TECHNIQUES



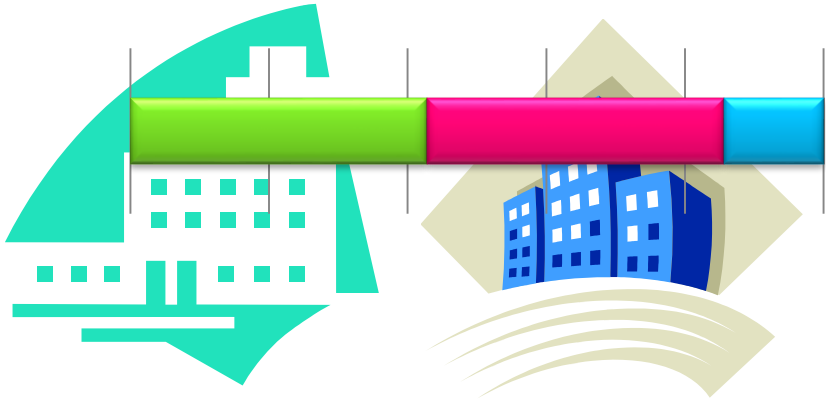




SELECTION OF SAMPLING TECHNIQUES

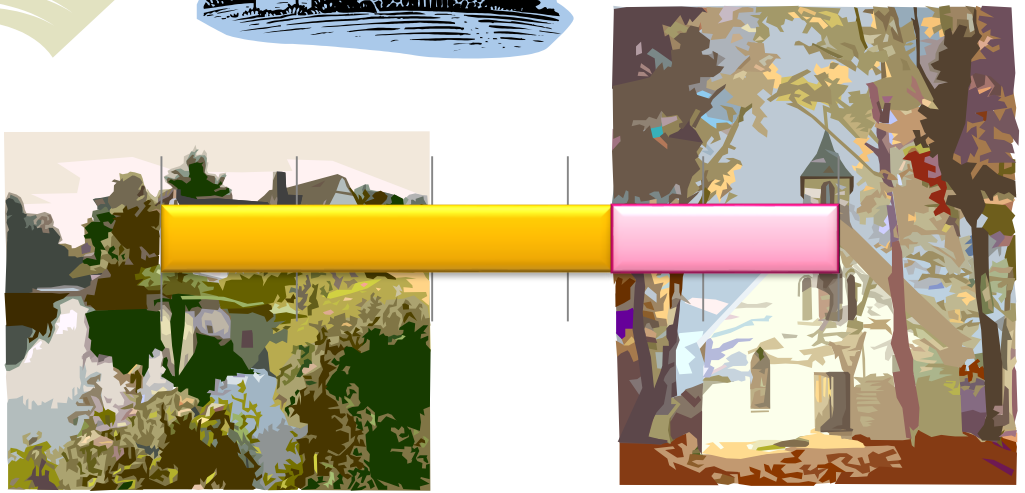


SELECTION OF SAMPLING TECHNIQUES

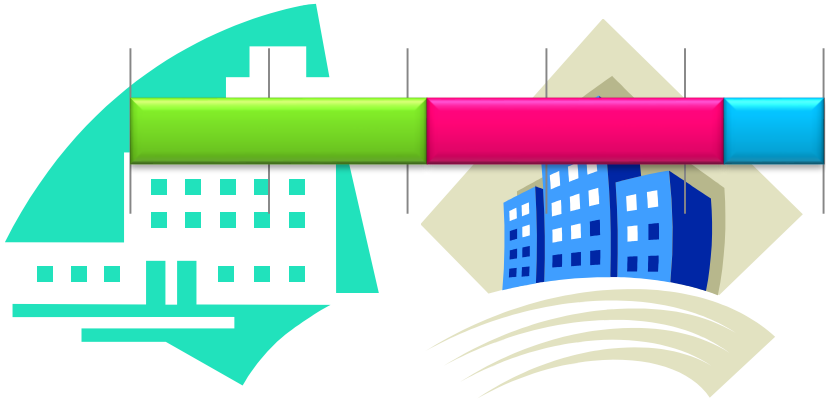


Merit based Sections

Gender based Sections

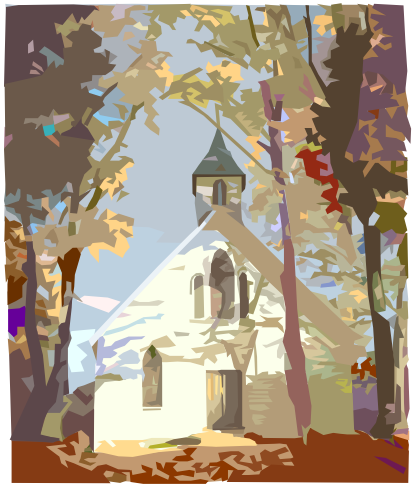


SELECTION OF SAMPLING TECHNIQUES



Merit based Sections

No particular criteria



CASE STUDY

THERE ARE AROUND 8,000 FIRMS ACROSS INDIA PROVIDING A CAB SERVICE.

AN AUDIT IS TO BE PLANNED TO CHECK THE CONFORMANCE OF TAX PAYMENTS.

ZONE	OPERATORS
NORTH	4627
SOUTH	3423
CENTRAL	1488
EAST	891
WEST	2396

REVENUE	NO. OF FIRMS
< 1 CRORE	4221
1 CR – 5 CR	3217
5 CR – 50 CR	770
50 CR – 250 CR	145
≥ 250 CR	13

CASE STUDY

HETROGENEOUS WITHIN

GEOGRAPHICAL

~~MUTUALLY EXCLUSIVE~~

SKEWED PROPORTIONS

ZONE	OPERATORS
NORTH	4627
SOUTH	3423
CENTRAL	1488
EAST	891
WEST	2396

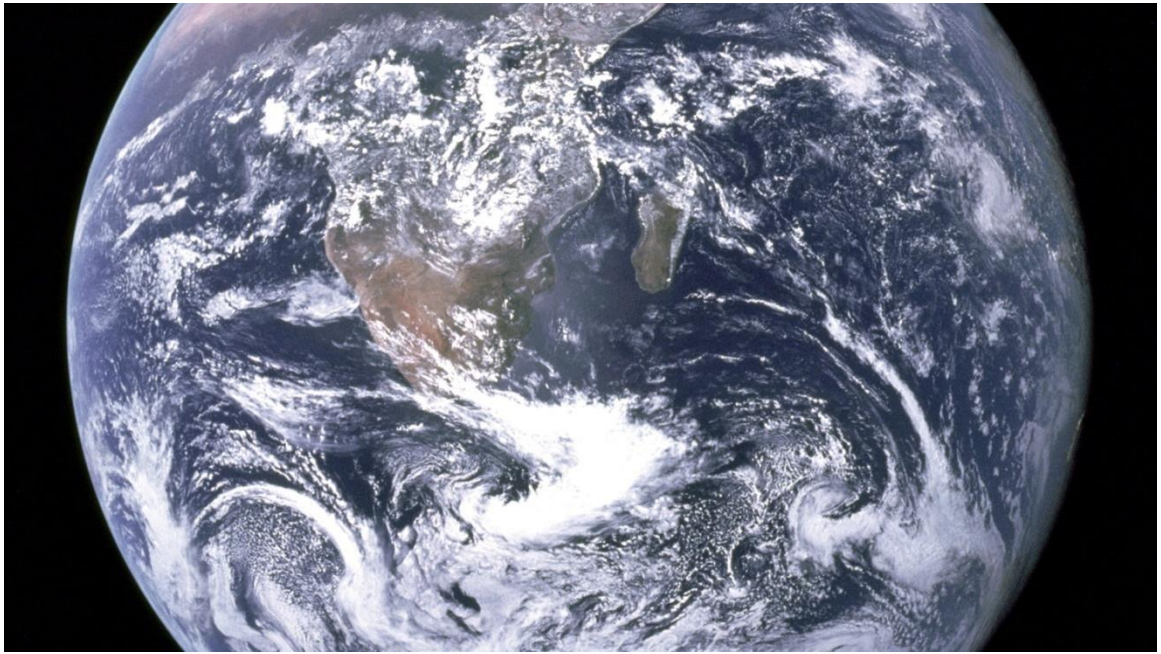
~~HETROGENEOUS WITHIN~~

PERIODIC TRAITS – NO

MUTUALLY EXCLUSIVE

SKEWED PROPORTIONS

REVENUE	NO. OF FIRMS
< 1 CRORE	4221
1 CR – 5 CR	3217
5 CR – 50 CR	770
50 CR – 250 CR	145
≥ 250 CR	13



Thanks

Note from Author

Much, if not all of this text uses content from internet and textbooks. Copyrights if any belong to respective holders.
The purpose of this compilation is only for education, and it is understood that it is not permitted to use it for commercial purposes.