Statistics for Computing
(CSC 502 0.0)

MSc in Computer Science

Prof. (Dr.) R.M. KAPILA RATHNAYAKA

B.Sc. Special (Math. & Stat.) (Ruhuna), M.Sc. (Industrial Mathematics) (USJ),
M.Sc. (Stat.) (WHUT, China),
Ph.D. (Applied Statistics, WHUT)

Model of Assessment

Continuous Assignments =40% (3 Assignments)

Final exam

Part I: 20 MCQ = 20%

Written paper (questions 3) = 40%

References :

- **Fundamentals of Descriptive Statistics using MINITAB (2020)**, R.M KAPILA THARANGA RATHNAYAKA, **ISBN :** 978-624-96333-0-8

- Borradaile, G. , Statistics for Earth Science, Data Springer.

- The Statistical Analysis of Experimental Data Book by John Mandel

- Statistical Analysis of Experimental Data: Springer Handbook of Experimental Solid Mechanics

What is Statistics? Where does this Data come from?

"Statistics is a way to get information from data"

Statistics

Data

Facts, especially numerical facts, collected together form references or information.

Information

Knowledge communicated concerning some particular fact.

Statistics is a *tool* for creating *new understanding* from a set of numbers.

What is Statistics?

Statistics is a discipline (scientific method) which is concerned with;

- Designing experiments and data collection,
- Summarizing information to aid understanding,
- Drawing conclusions from data, and
- Estimating the present or predicting the future.

Present Importance of Statistics…..

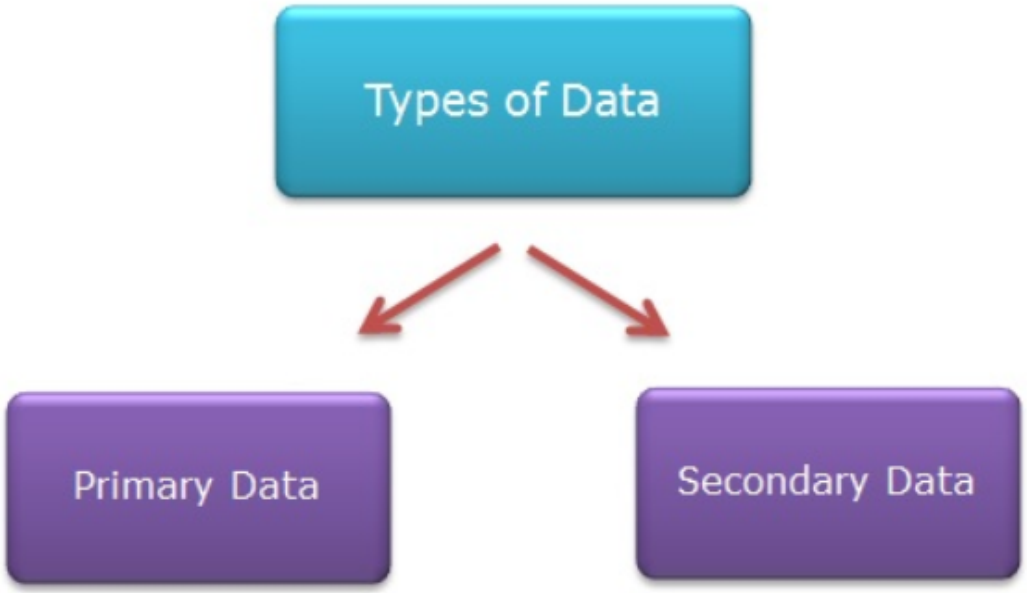
- Today, statistics has become an important tool in the work of many academic disciplines such as;
    - Medicine,
    - Psychology, Sociology,
    - Engineering
    - Physics and Chemistry

- Statistics is also important in many aspects of society such as business, industry and government.

Data Identification

Sources of Data

- The data that you collect may be either primary data or secondary data.

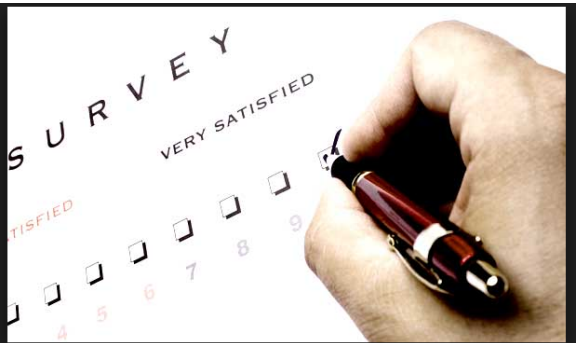


- It can be gathered by organizations using experiments or surveys, or by individual workers.
- The main difference between primary and secondary data is related to the way that the data itself is gathered.

Primary data

- When you create the data you want by yourself, it's called primary data.
- Data observed or collected directly from first-hand experience is called primary data. Also known as raw data.
- In primary data collection, you collect the data by yourself using methods such as ;

  Questionnaires
  Interviews
  Diaries
  Portfolios



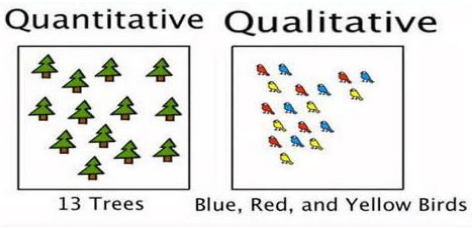
Collecting your own data-Primary research | Data Collection Services

Secondary data

Published data and the data collected in the past or other parties is called secondary data.

Some examples of Secondary sources

- Newspapers and popular magazine articles. (may also be Primary)
- Dictionaries and encyclopedias
- Organizational records



Data

Numerical

Made of numbers

Age, weight, number of children, shoe size

Categorical

Made of words

Eye colour, gender, blood type, ethnicity

Quantitative

Qualitative

Continuous

Infinite options

Age, weight, blood pressure

Discrete

Finite options

Shoe size, number of children

Ordinal

Data has a hierarchy

Pain severity, satisfaction rating, mood

Nominal

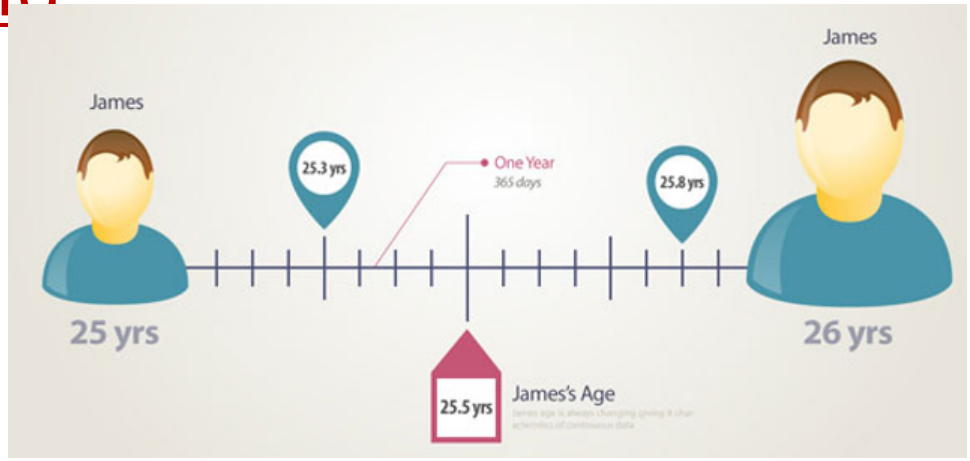
Data has no hierarchy

Eye colour, dog breed, blood type

Numerical Data (Data that is Numbers) : **Continuous Variables**

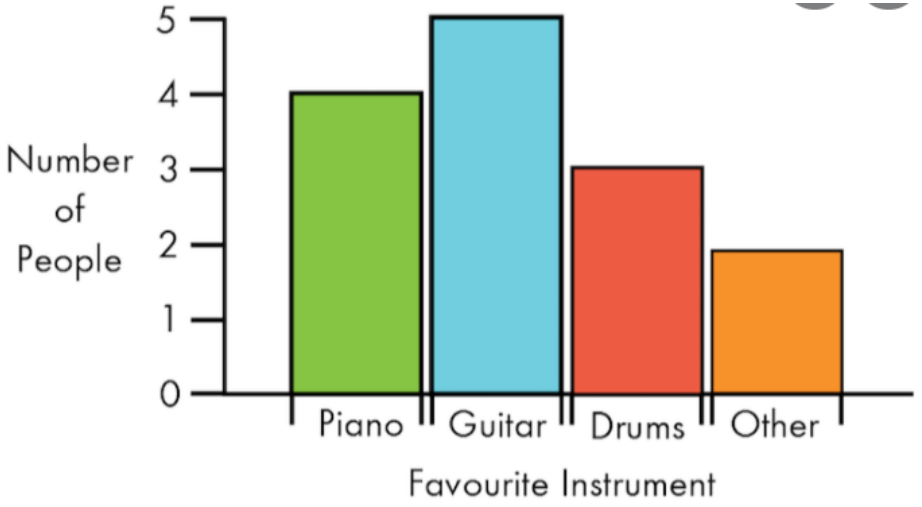
Continuous variables is a variable whose value is obtained by measuring

- ✓ height of students in class
- ✓ weight of students in class
- ✓ time it takes to get to school
- ✓ distance traveled between classes



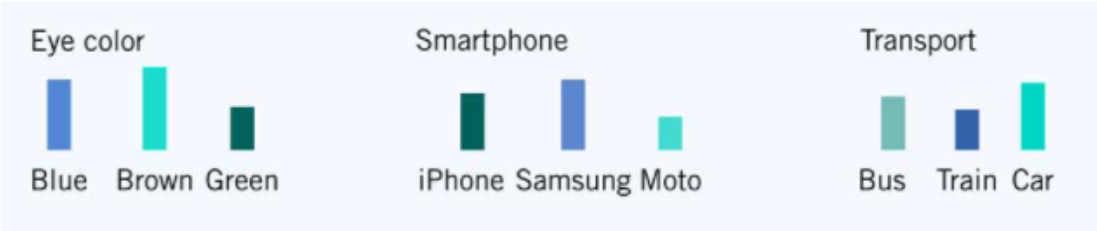
Numerical Data (Data that is Numbers) : **Discrete Variables**

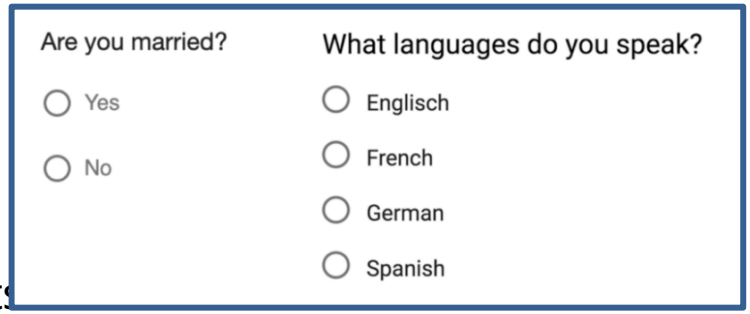
- A discrete variable is a variable whose value is obtained by counting.
- All continuous variables are numeric, but not all numeric variables are continuous.
- Examples:
  - number of students pre
  - number of red marbles
  - number of heads when
  - students' grade level



Categorical Data (Data that is not numbers) : **Nominal Variable**

- Sometimes there is no hierarchy in categorical data.
- If eye colour was coded
  - 0-- “Blue”
  - 1 --“Green”
  - 2 --“Brown”

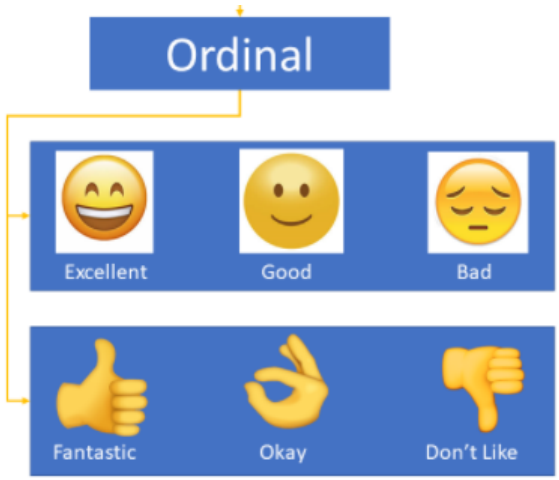




- we have to randomly choose which option gets which number.
- It doesn’t matter whether Blue eyes is zero, or one, or two, because there is no hierarchy in eye colour.

Categorical Data (Data that is not numbers) : Ordinal Variable

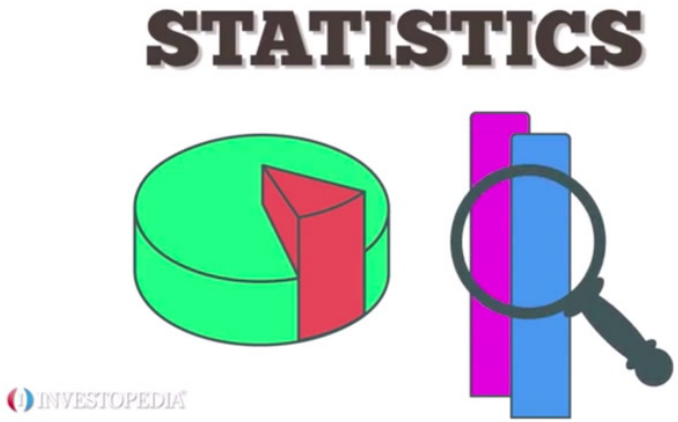
- Annoying surveys often ask you to answer with the options "Strongly Disagree", "Disagree", "Neutral", "Agree" or "Strongly agree".
- This data has a special structure, because if these are coded 0 "Strongly Disagree" to 4 "Strongly agree";
  - 0 = Strongly Disagree
  - 1 = Disagree
  - 2 = Neutral
  - 3 = Agree
  - 4 = Strongly



What Is Your Educational Background?

- 1 - Elementary
- 2 - High School
- 3 - Undegraduate
- 4 - Graduate

Basic Definitions and Concepts

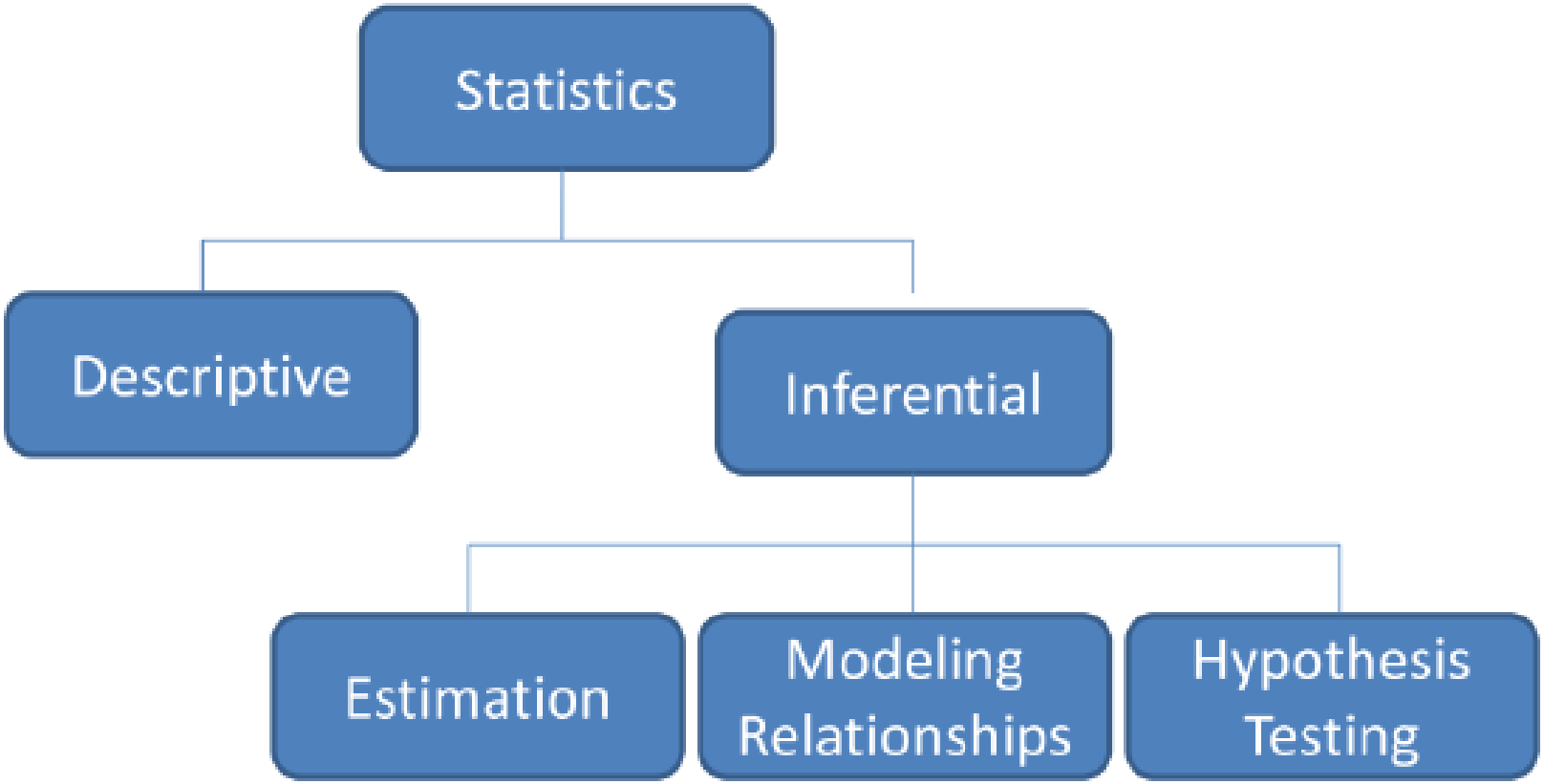


Kinds of Statistics

We can divide statistics in to two parts.

1. Descriptive statistics

2. Inferential statistics



Key Statistical Concepts…

1. Population

- ✓ A population is the set of all the individuals of interest in a particular study.

- ✓ It is an entire group of people or study elements, things or measurements having some common fundamental characteristics

- ✓ *Any actual or conceptual collection of individual items*, defined by stranded characteristics.

Example

- Advertisements for IT jobs in the Sri Lanka
- Songs from the VOICE Song Contest
- Undergraduate students in SUSL
- All countries of the world

- Mainly the term population can be divided in to two parts.

    1. Finite population

    2. Infinite population

- *Finite population*

    If a population consists of fixed number of values then it is said to be finite.

    Ex: Number of days per month.

- *Infinite population*

    If a population consists of an endless succession of values, it is said to be infinite.

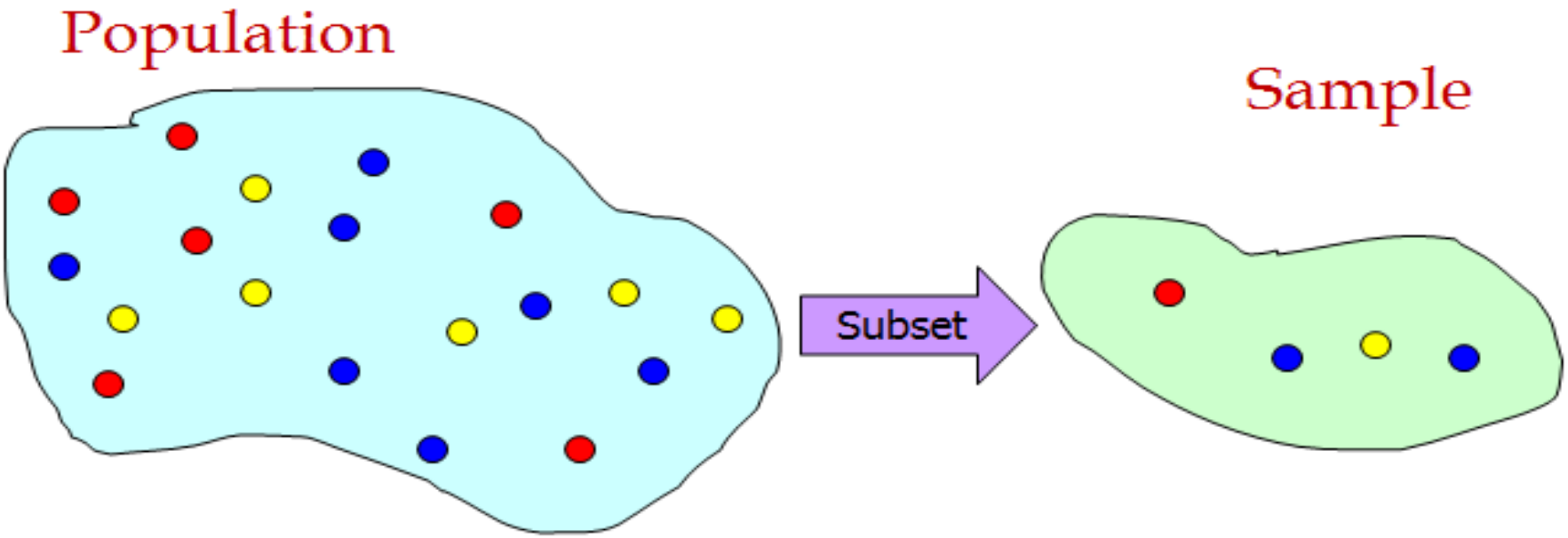
    Ex: Number of insects in a certain region.

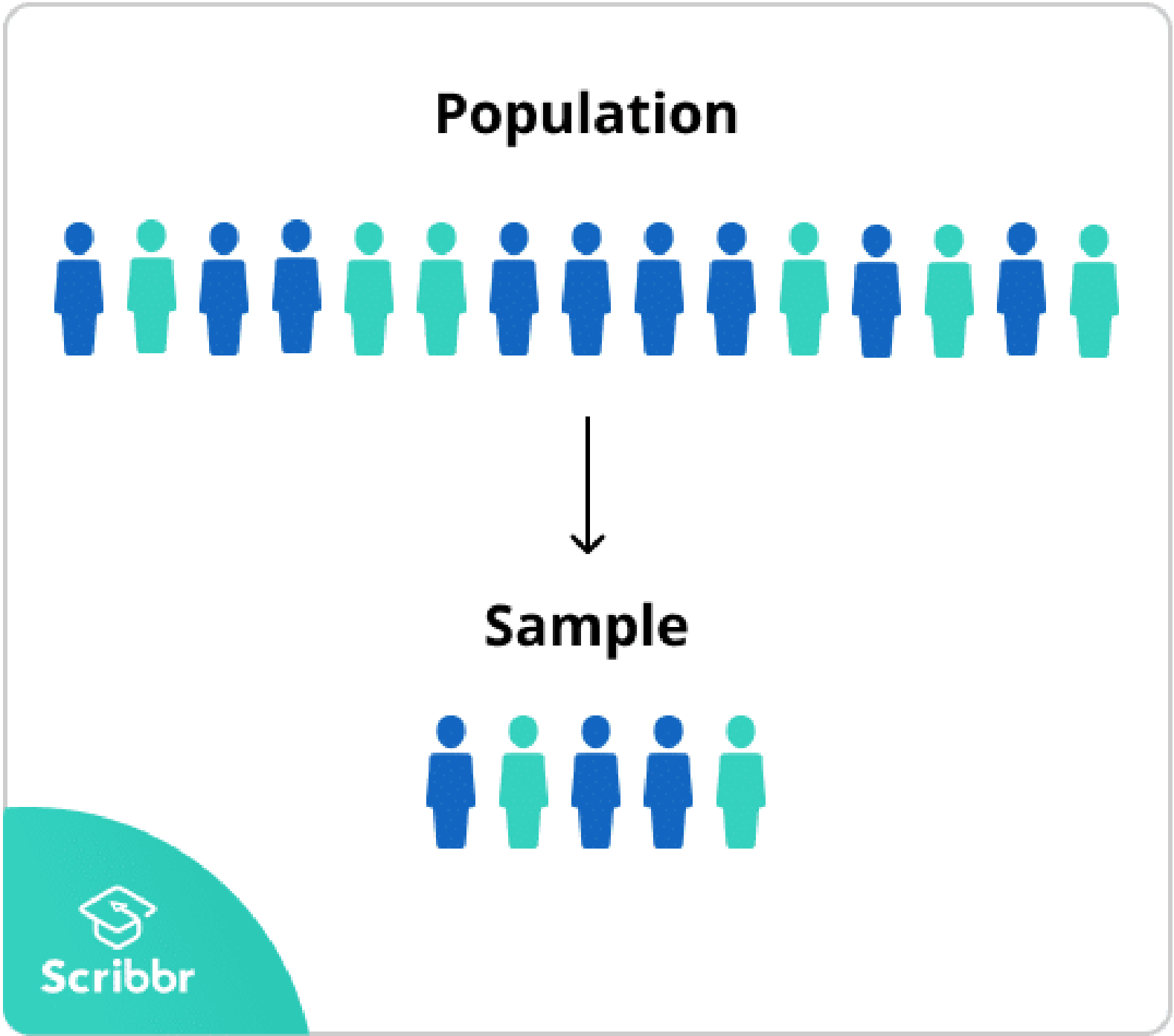
Key Statistical Concepts…

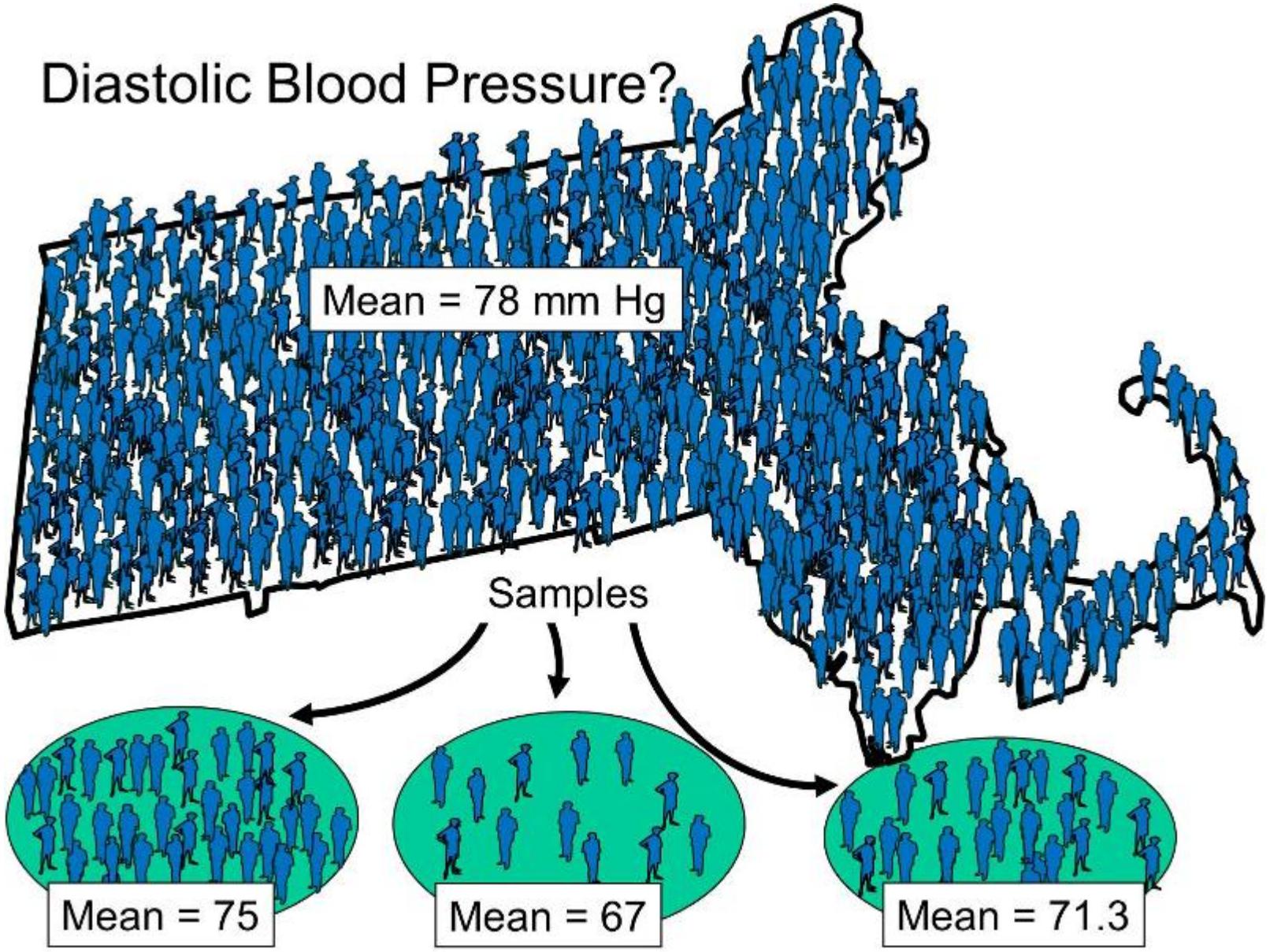
Sample

— A ***sample*** is a set of data drawn from the population. (A sample is a *small segment of the population*)

— Potentially very large, but less than the population.




Population
Sample
Scribbr


Diastolic Blood Pressure?
Mean = 78 mm Hg
Samples
Mean = 75
Mean = 67
Mean = 71.3

Reasons for sampling

- **Necessity**: Sometimes it's simply not possible to study the whole population due to its size or inaccessibility.

- **Practicality**: It's easier and more efficient to collect data from a sample.

- **Cost-effectiveness**: There are fewer participant, laboratory, equipment, and researcher costs involved.

- **Manageability**: Storing and running statistical analyses on smaller datasets is easier and reliable

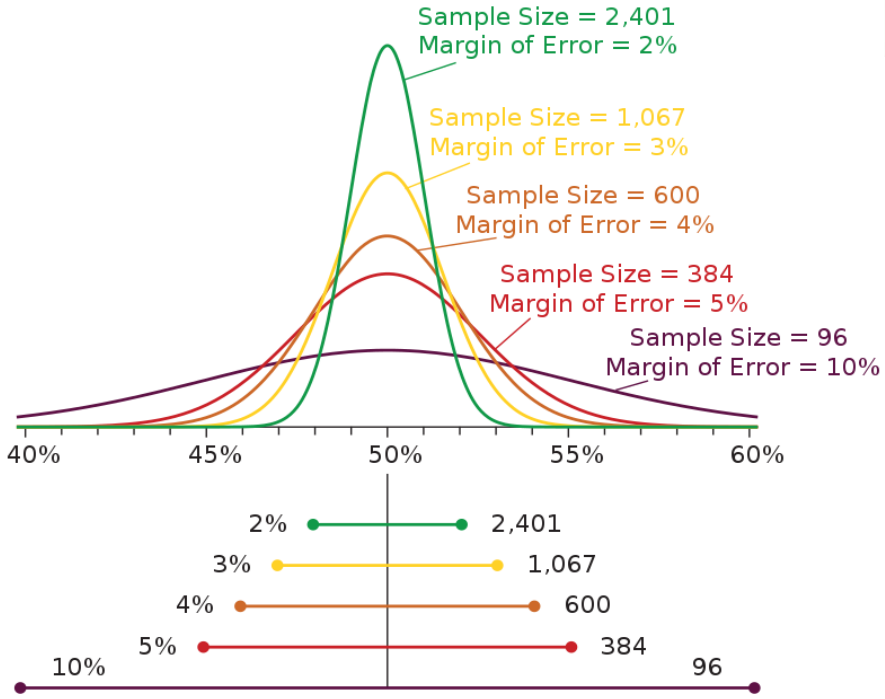
Collecting data from a sample

- When your population is
  - large in size,
  - geographically dispersed, or difficult to contact,

  it's necessary to use a sample.

- You can use sample data to make estimates or test hypotheses about population data.

$$\text{Sample Size} = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$



N = Population size

z = z-score

e = margin of error

p = standard of deviation

For Unknown or Very Large Populations

$$\text{Sample size} = \frac{z^2 \times p(1-p)}{e^2}$$

Example

- Determine the ideal survey size for a population size of 425 people. Use a 99% confidence level, a 50% standard of deviation, and a 5% margin of error.

- For 99% confide

- This means that
  - $N = 425$
  - $z = 2.58$
  - $e = 0.05$
  - $p = 0.5$

$$\text{Sample size} = \frac{\frac{2.58^2 \times 0.5(1-0.5)}{0.052}}{1 + \left(\frac{2.58^2 \times 0.5(1-0.5)}{0.052 \times 425}\right)} = 259.39$$

$$\text{Sample Size} = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$

	Confidence level = 95%			Confidence level = 99%		
	Margin of error			Margin of error		
Population size	5%	2,5%	1%	5%	2,5%	1%
100	80	94	99	87	96	99
500	217	377	475	285	421	485
1.000	278	606	906	399	727	943
10.000	370	1.332	4.899	622	2.098	6.239
100.000	383	1.513	8.762	659	2.585	14.227
500.000	384	1.532	9.423	663	2.640	16.055
1.000.000	384	1.534	9.512	663	2.647	16.317

The minimum sample size is 100

- Most statisticians agree that the minimum sample size to get any kind of meaningful result is 100.
- If your population is less than 100 then you really need to survey all of them.

A good maximum sample size is usually 10% as long as it does not exceed 1000

- A good maximum sample size is usually around 10% of the population, as long as this does not exceed 1000.
- For example, in a population of 5000, 10% would be 500.

Example:

- You want to study political attitudes in young people.
- Your population is the 30,000 undergraduate students in the Sri Lankan Universities.
- Because it's not practical to collect data from all of them, you use a **sample** of 300 undergraduate volunteers from three Sri Lankan universities
- This is the group who will complete your online survey.

Acceptance Sampling

- Acceptance sampling uses statistical sampling to determine whether to accept or reject a production lot of material.
- It is usually done as products leave the factory, or in some cases even within the factory.
- Inspection for acceptance purpose is carried out at many stagers in the manufacturing.
- They are generally two ways in which inspection is carried out:
  1. 100% inspection
  2. Sampling inspection

- In 100% inspection all the parts or products are subjected to inspection, whereas in the sampling inspection only a sample is drawn from the lot and inspected.

- Sampling inspection is more practical, quick and economical.

- The advantages of sample inspection are as follows.
  - The cost and time required for sampling inspection is quite less as compared to 100% inspection.
  - Smaller inspection staff is necessary.
  - Less damage to products because only few items are subjected to handling during inspection.
  - The lot is disposed off in shorter time so that scheduling and delivery are important.

Example:

- You want to study political attitudes in young people.

- Your population is the 10,000 undergraduate students in the SUSL.
- Because it's not practical to collect data from all of them, you use a **sample** of 300 undergraduate volunteers from three Faculties.

- This is the group who will complete your online survey.

SAMPLING

The process of selecting individuals for a study

Sampling Techniques

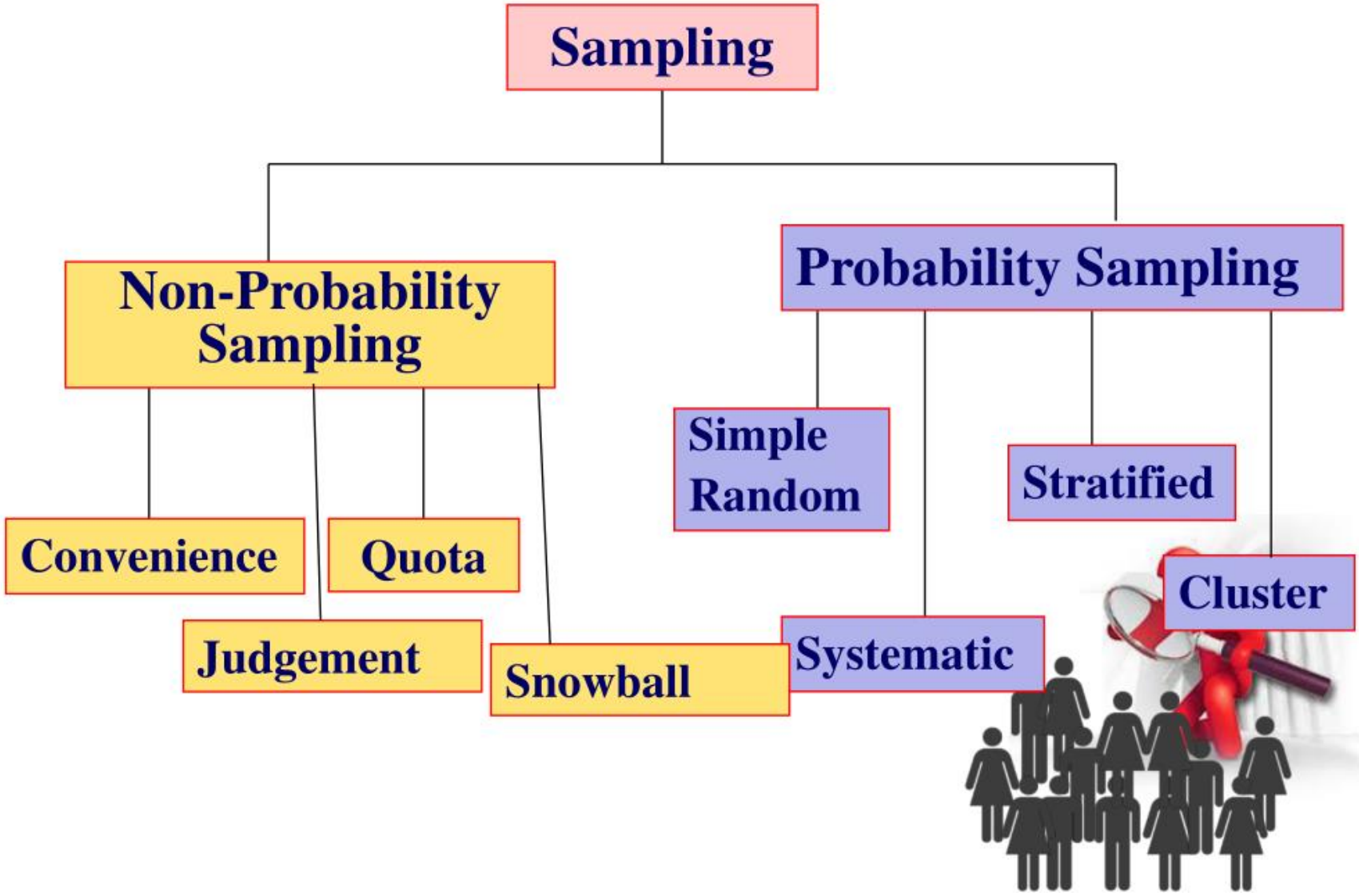
Sampling Methods can be classified into one of two categories:

- **Probability Sampling**: Sample has a known probability of being selected
- **Non-probability Sampling**: Sample does not have known probability of being selected as in convenience or voluntary response surveys

Probability Sampling

- In probability sampling it is possible to both determine which sampling units belong to which sample and the probability that each sample will be selected.

TYPES OF SAMPLING



The following sampling methods are types of **probability sampling**:

- **Simple Random Sampling (SRS)**
- **Stratified Sampling**
- **Cluster Sampling**
- **Systematic Sampling**
- **Multistage Sampling (in which some of the methods above are combined in stages)**

Simple Random Sampling

- Simple random sampling is a method of selecting n units out of the N.

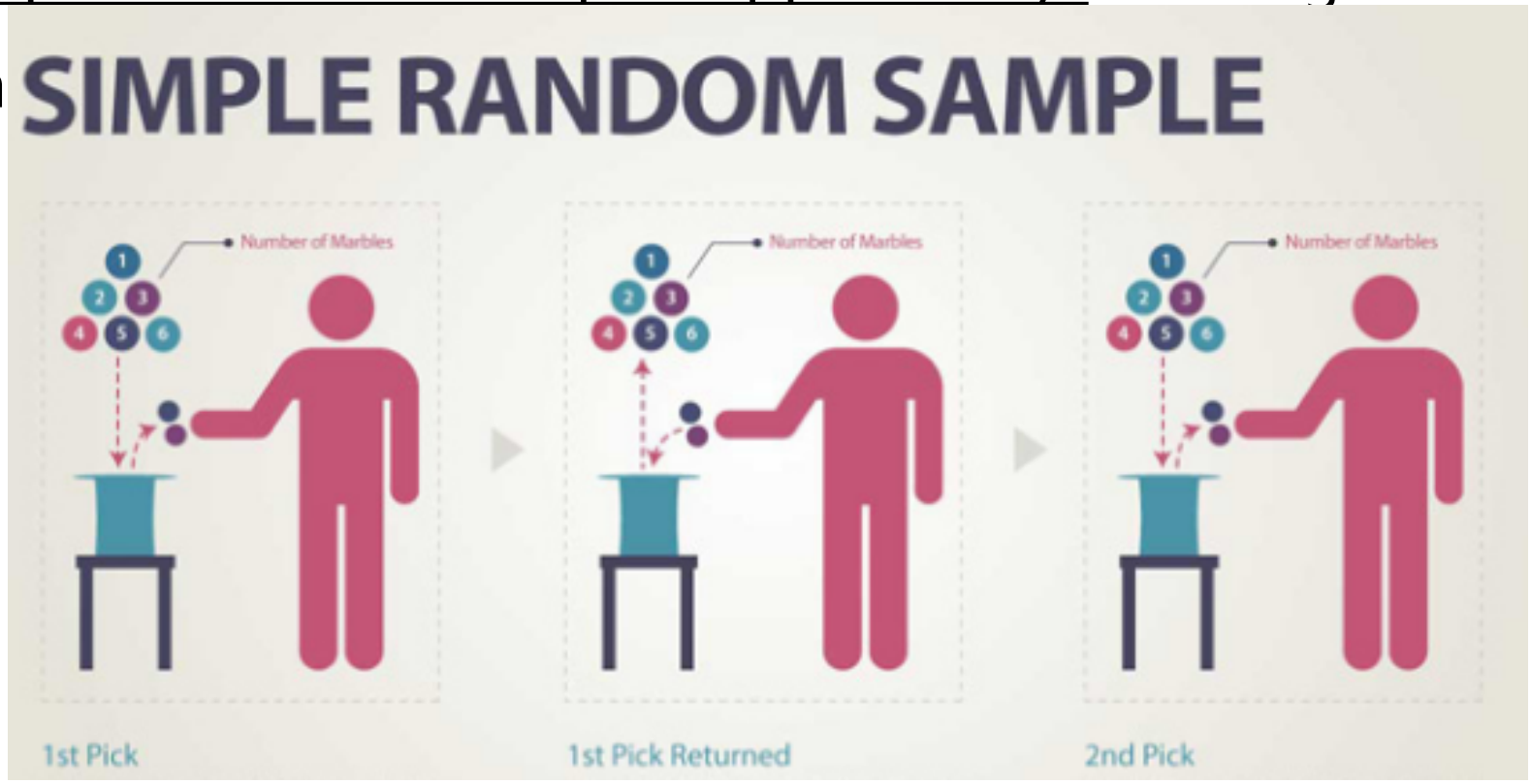
- Such that every one of the N_{C_n} distinct samples has an equal chance of being drawn.

- In practice, a simple random sample is drawn unit by unit.

- The units in the population are numbered from 1to N.

Random Sample

It is a sample chosen in a very specific way and has been selected in such a way that every element in the population has an equal opportunity of being included in



- A series of random numbers between 1and N is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table.

TABLE 1 - RANDOM DIGITS

11164	36318	75061	37674	26320	75100	10431	20418	19228	91792
21215	91791	76831	58678	87054	31687	93205	43685	19732	08468
10438	44482	66558	37649	08882	90870	12462	41810	01806	02977
36792	26236	33266	66583	60881	97395	20461	36742	02852	50564
73944	04773	12032	51414	82384	38370	00249	80709	72605	67497
49563	12872	14063	93104	78483	72717	68714	18048	25005	04151
64208	48237	41701	73117	33242	42314	83049	21933	92813	04763
51486	72875	38605	29341	80749	80151	33835	52602	79147	08868
99756	26360	64516	17971	48478	09610	04638	17141	09227	10606
71325	55217	13015	72907	00431	45117	33827	92873	02953	85474
65285	97198	12138	53010	94601	15838	16805	61004	43516	17020
17264	57327	38224	29301	31381	38109	34976	65692	98566	29550
95639	99754	31199	92558	68368	04985	51092	37780	40261	14479
61555	76404	86210	11808	12841	45147	97438	60022	12645	62000
78137	98768	04689	87130	79225	08153	84967	64539	79493	74917

Population	Sample
Advertisements for IT jobs in the Sri Lanka	The top 50 search results for advertisements for IT jobs in the Sri Lanka on January 1, 2021
Songs from the VOICE Song Contest	Winning songs from the VOICE Song Contest that were performed in English
Undergraduate students in the Sri Lanka	300 undergraduate students from three Sri Lanka universities who volunteer for your psychology research study
All countries of the world	Countries with published data available on birth rates and GDP since 2000

Stratified Random Sampling

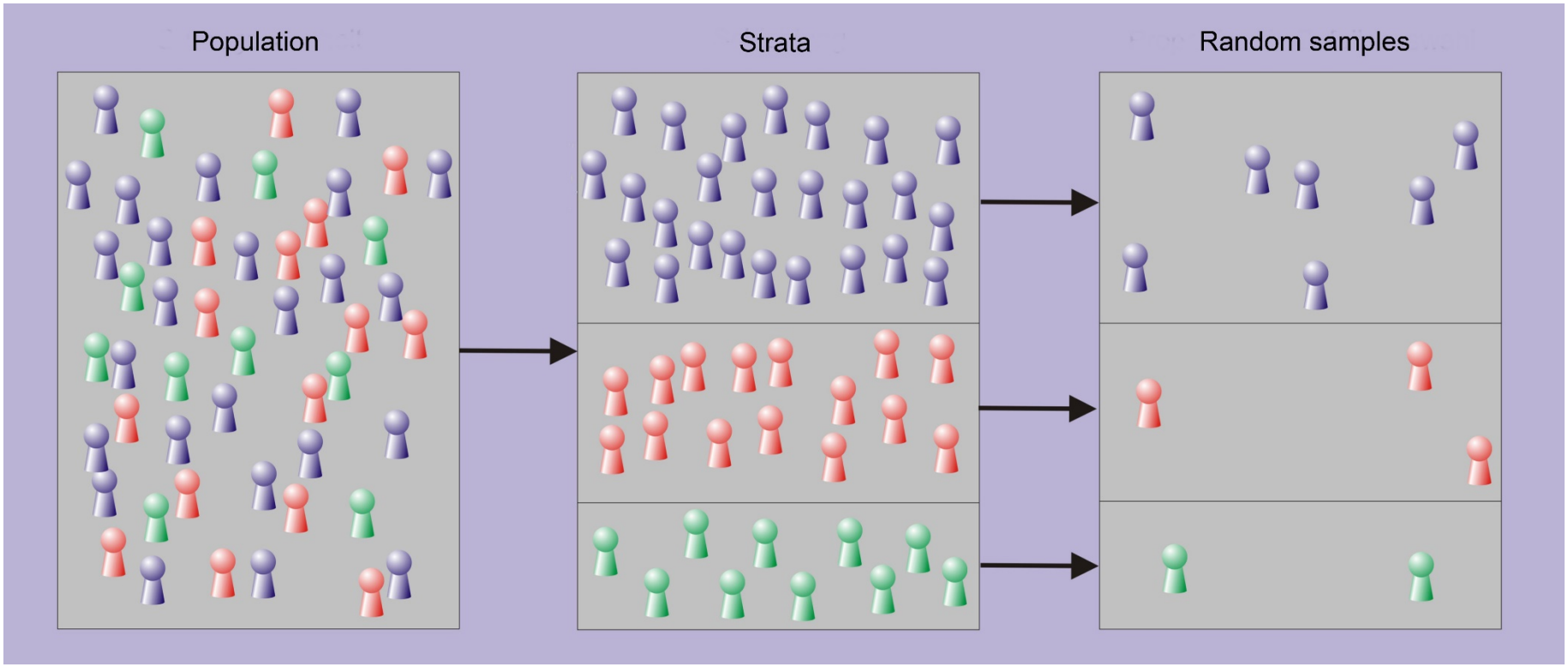
- If the population is not homogeneous, and if it is made up of homogeneous parts, then Population of N units is first divided into subpopulations of $N_1, N_2, N_3, \ldots, N_L$ units.

- Subpopulations are non-overlapping.

- $N_1 + N_2 + N_3 + \ldots + N_L = N$

- Subpopulations are called Strata.

- To obtain the full benefit of the stratification, N must be known.


Population
Strata
Random samples

- When the strata have been determined, a sample is drawn from each , the drawings being made independently in different strata.

- The sample sizes within the strata are denoted by $n_1, n_2, \ldots, n_L$, respectively.

- If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling.

Example:

- A school offers three subject streams science, arts and commerce. Number of students in the above streams are 135, 45, and 90 respectively.

- Explain how do you use stratified random sampling to select 30 students?

Systematic Sampling

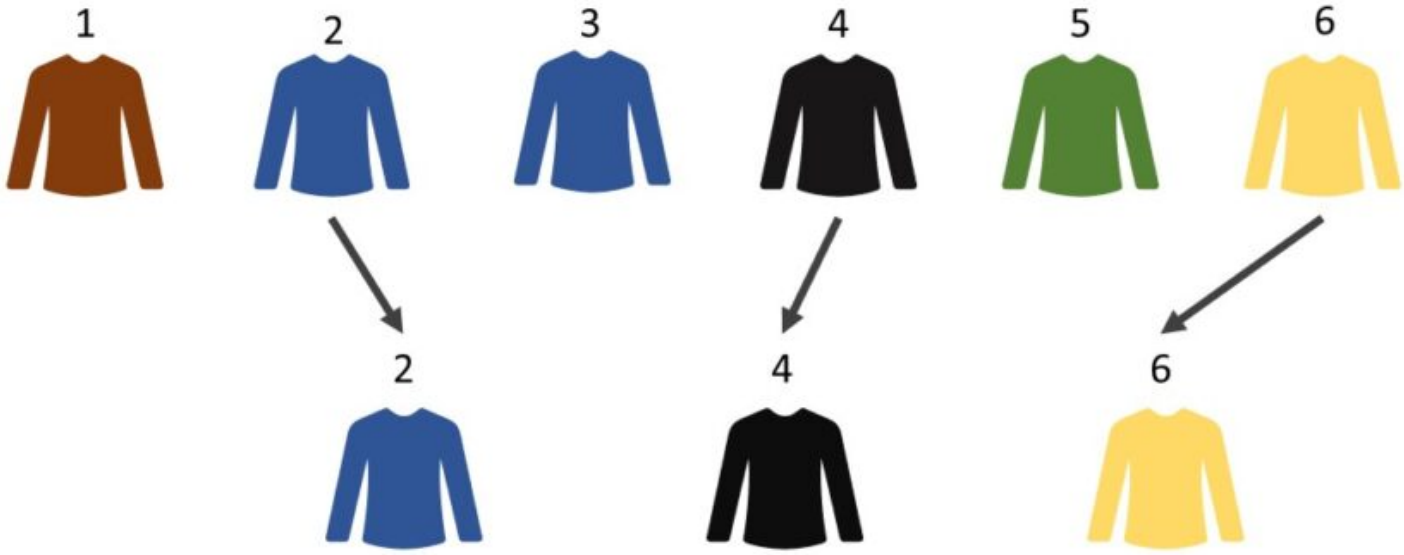
- For a homogeneous population, if the list of all item is not available, it is not possible to apply simple random sampling method.

- Here, it can be applied systematic sampling technique.

- Suppose that the N units in the population are numbered 1 to N in some order.

Systematic Sampling Process

Here, we select every 2nd T-shirt from the lot



- To select a sample of n units, we take a unit at random from the first k units and every kth unit thereafter.

- k can be guessed or can be the ratio of "population size / sample size"

  if the population size is known.

Examples:

- If k is 15 and the first unit drawn is number 13, the subsequent units are numbers $13, 28, 43, 58$, and so on.

- There are 150 students in a class. Obtain a systematic sample of size 20 by determining suitable sample point.

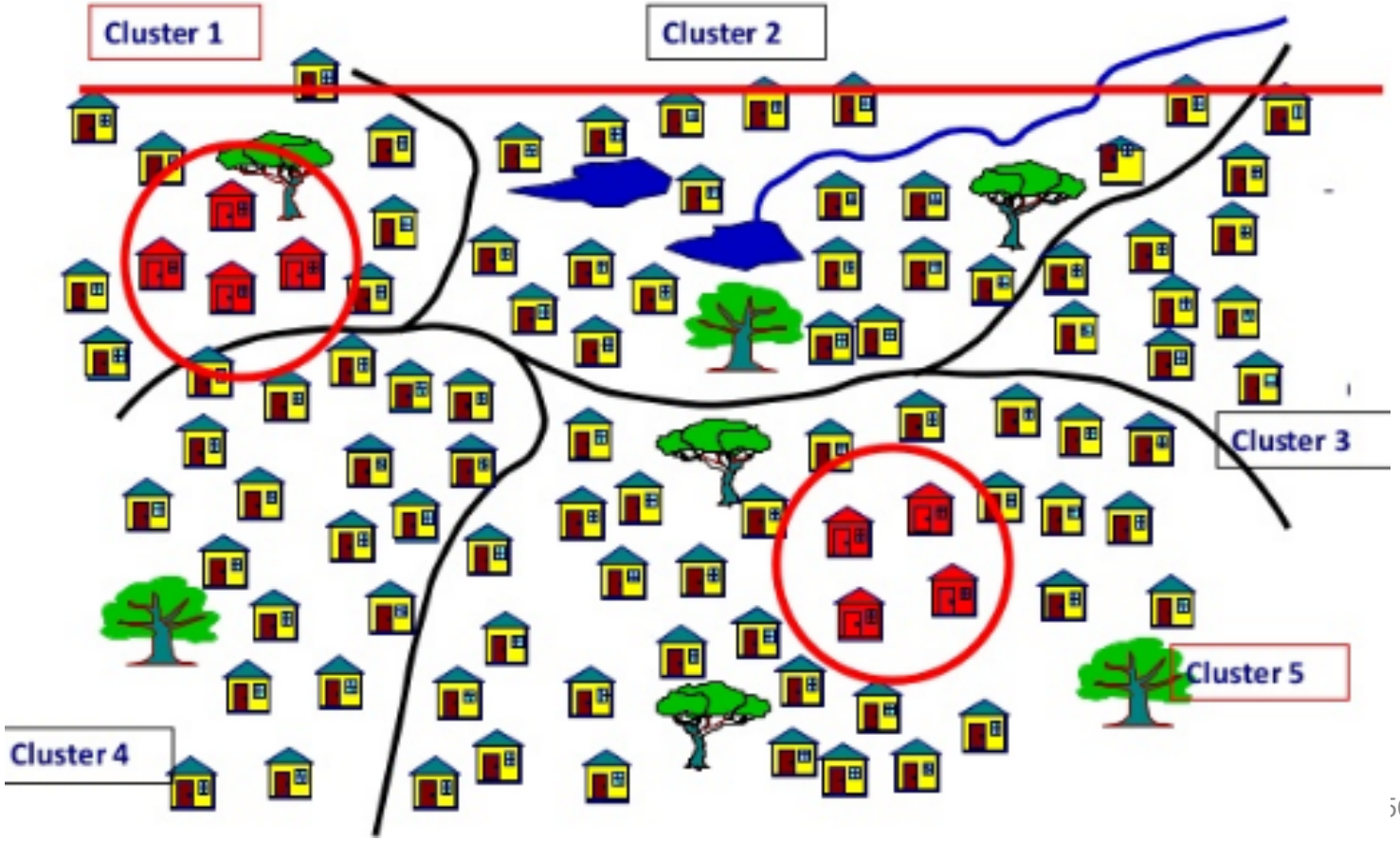
- $k = \frac{150}{20} = 7.5 \sim 7$, let the sample point is 2. then the sample is $2, 9, 16, 23, 30, \ldots.$

- The selection of first unit determine the whole sample.

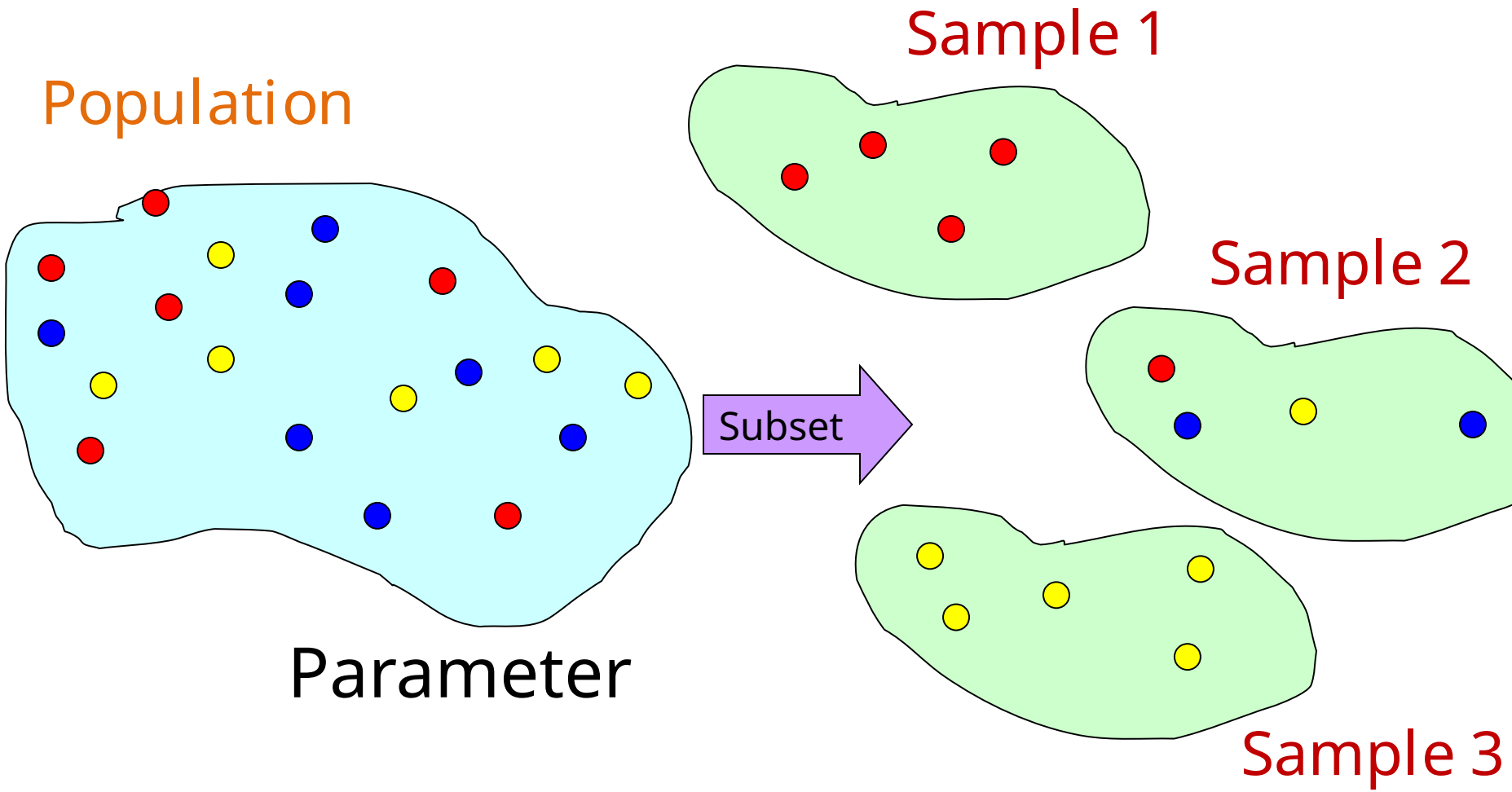
Cluster Sampling

- If the population is made up of similar subgroups, such subgroups are known as clusters.

Example:



Key Statistical Concepts…

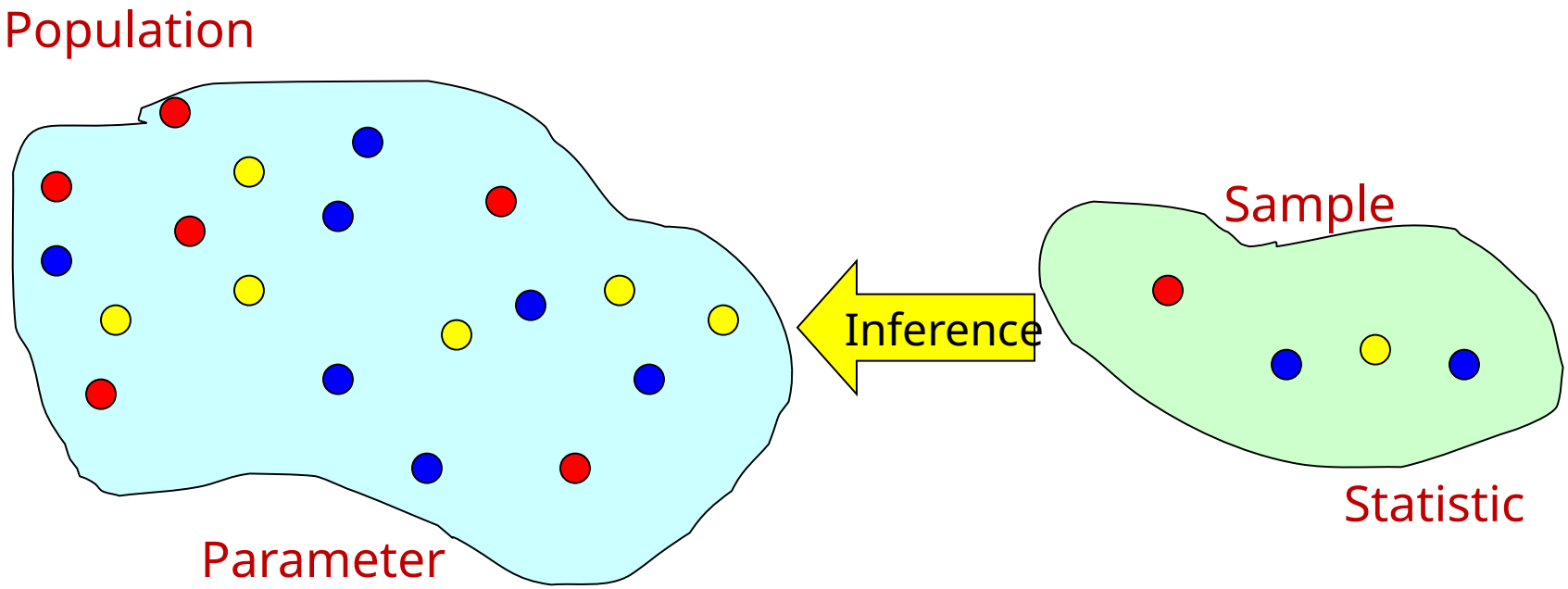


Population has Parameters,

Samples have Statistics.

Statistical Inference…

Statistical inference is the ***process*** of making an estimate, prediction, or decision about a population based on a sample.



- Inferential statistics is used to draw conclusions or inferences about characteristics of ***populations*** based on data from a ***sample***.