

21CSC529T Inferential Statistics

Unit 1 : Introduction to Statistics

Syllabus :

Unit - 1: INTRODUCTION TO STATISTICS

Unit - 2: PROBABILITY

Unit - 3: ESTIMATION & HYPOTHESIS TESTING

Unit - 4: HYPOTHESIS TESTING -I

Unit - 5: HYPOTHESIS TESTING -II

Unit 1

- Role of statistics in Data science
- Different types of data. - Understanding Random variable, Numerical Variable and Categorical Variable
- Data Collection, Types and its formats
- Types of Sampling
- Descriptive Statistics - Measure of Central Tendency Mean, Median and Mode
- Measure of Dispersion Range, Quartiles, Standard Deviation, Variance
- Distribution of Data Skewness and Kurtosis
- Covariance and Correlation, Difference between covariance and correlation and its significance.

Agenda

- Introduction to Statistics
- Classification of Data
- Various Types of Sources
- Characteristics of Statistics
- Sampling and its types
- Descriptive Statistics
 - Measure of Central Tendency

‘STATISTICS’

“Statistics is not a body of substantive knowledge, but a body of methods obtaining knowledge”

-W.A.Wallis

Statistics

- ★ Thought to have been derived from the Latin word '*statisticum collegium*' meaning council of state
- ★ The definition was limited to the collection of economic and demographic data
- ★ In the 19th century the definition was broadened and included collection, summarization, and analysis of data
- ★ Definition: *Statistics is the branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It provides tools and methods for making sense of numerical data, allowing researchers and analysts to draw conclusions, make predictions, and inform decisions.*

Data In - Sights

"Data" has become a buzzing theme in today's world.

What is Data

Data is units of information in structured or unstructured format

Examples:

- Collection of relevant tweets
- Records of yield in a farm over a period of time
- Records of stock price every minute
- Records of performance of a sports person

“Collection of Facts”

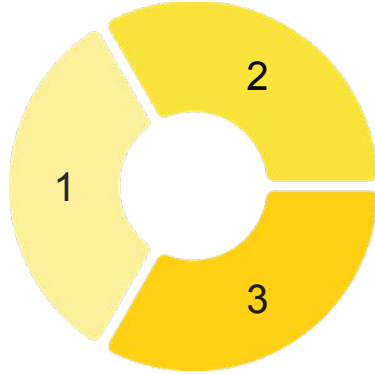
Kinds of Data?

- Sound
- Video
- Single character
- Number (integer or floating-point)
- Picture
- Boolean (true or false)
- Text (string)

Types of Data

Quantitative Data

Numerical and measurable data such as height, weight, or temperature.



Qualitative Data

Describes characteristics that cannot be measured numerically, like colors and labels.

Primary vs. Secondary Data

Primary data is firsthand, while secondary data is collected from existing sources.



Sources of Data Collection

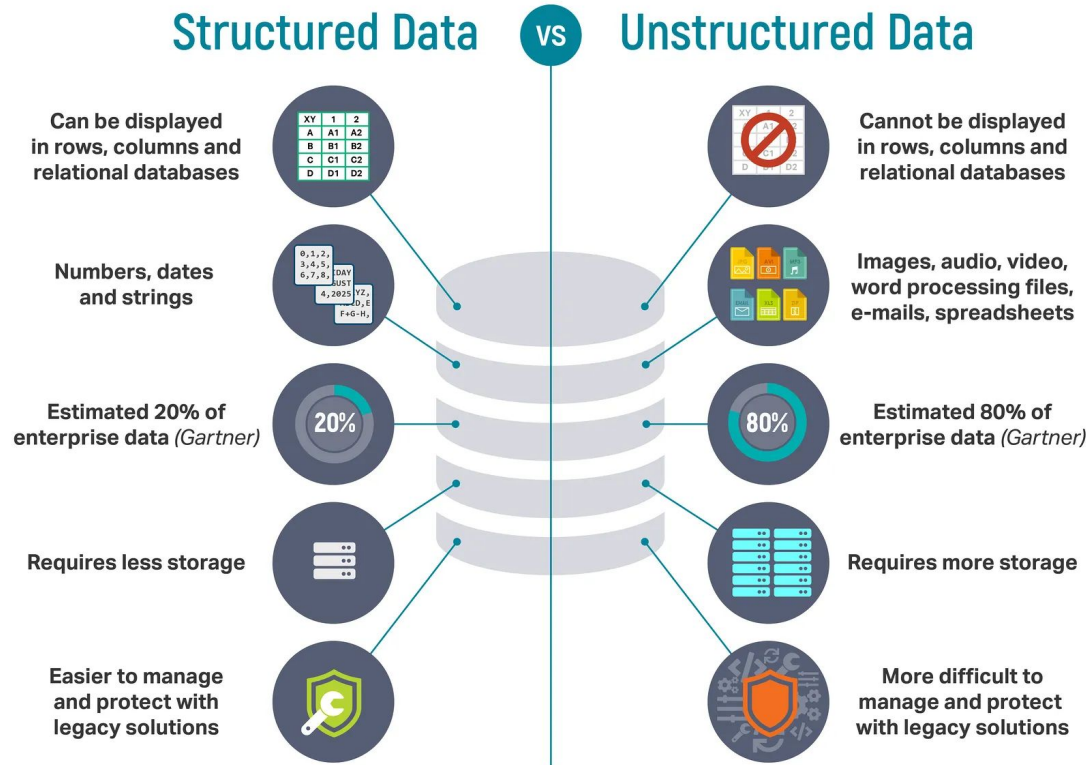
Primary Sources

- Data collected by user
- Raw data gathered first-hand from the source
- Objective for data collection is pre-defined

Secondary Sources

- Data is not collected by the user
- Data collected by someone else and is readily available
- Objective for the data collection could be different than the objective of the user

Difference between structured Vs Unstructured Data



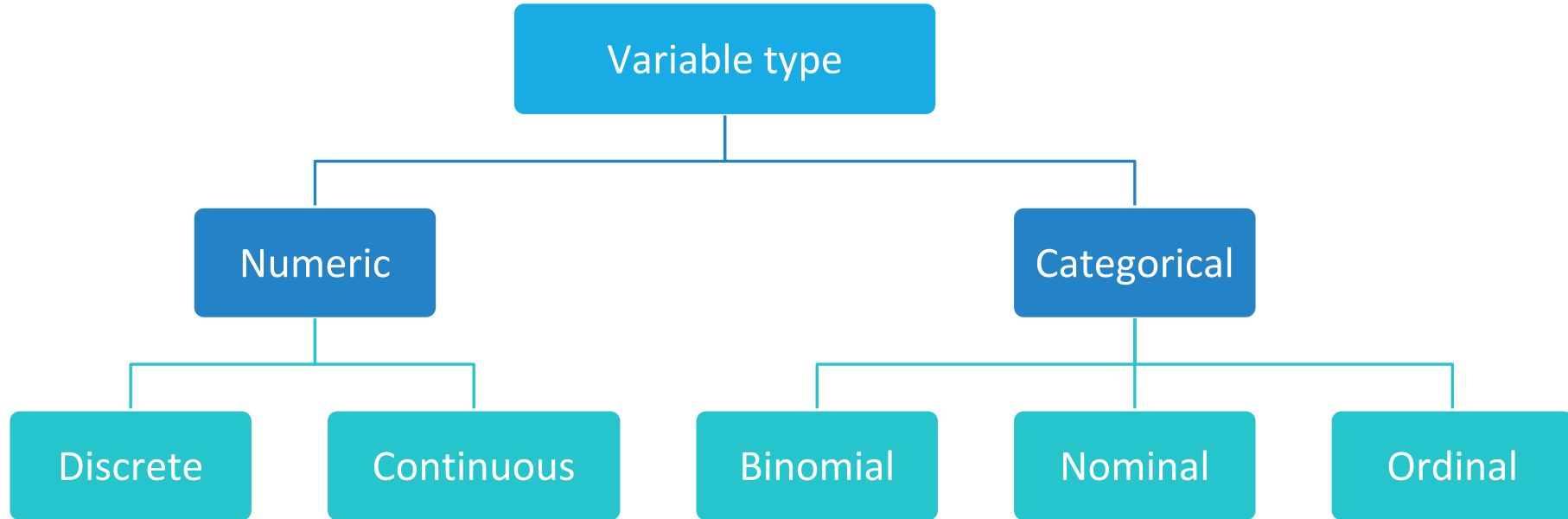
How to store Data?

- Data can be stored in temporary storage spaces called variables.
- A variable is a value that can change, depending on conditions or on information passed to the program.

How to work with variables?

- Create Variables
- Storing the values
- Retrieving and using the stored value

Understand the variable



Discrete Numerical Variable

- A variable which takes a numeric value is called a **numeric variable**
- Also known as **quantitative variable**
- A **discrete numeric variable** is a random variable which takes discrete values, i.e. values from the set of whole numbers only. It can take countably finite values

Examples:

1. Number of cars passing by a toll-gate every minute
2. Number of defective items in a box

Continuous Numerical Variable

A **continuous numeric variable** is a variable which can have infinite number of values within a range

Examples:

1. The amount of rainfall in millimeters
2. Price of a stock

Categorical Variable

- Categorical variable has two or more levels
- Also known as **qualitative variable**

Examples:

1. Colors: red, orange, yellow
2. Gender: male, female
3. Economic status: low, medium, high

Categorical Variable

Nominal Data:

- Nominal data has **no order** and has **two or more categories**
- It represents discrete units and are used to label variables
- Example: Housing type - Apartment, Bungalow, Penthouse

Binary Data:

- Binary data has **no order** and has **strictly two categories**
- Also known as dichotomous variable
- Example: Presence or absence of a disease

Categorical Variable

Ordinal Data:

- Ordinal data is **ordered** nominal data
- It represents discrete and ordered units
- Example: Education - Primary, Secondary, High School, College and University which are ordered



- At times data may appear to be numeric but is actually categorical
- Consider the adjoining table. The 'Gender' column has value 1 and 0
- The numeric values used to represent a categorical variable cannot be used for numerical computation

Name	Gender	Age
Ria	1	45
Arya	1	23
Sam	0	34
Joe	0	54
Harry	0	12
Sarah	1	43

Classification of Data

In general, there are four types of data. They are:

- Time Series Data
- Cross sectional data
- Pooled Data
- Panel Data

Time series data

- Time series data is the set of observations on a variable at different time points
- The data may be collected daily, weekly, monthly, or annually

Date	Prev.Close	Open.Price	High.Price	Low.Price
01-Jan-1996	408	407	407.9	405
02-Jan-1996	407.9	407	409	406.25
03-Jan-1996	406.25	409	409	409
04-Jan-1996	409	405	407	405
05-Jan-1996	406.3	401.5	401.5	401.5
08-Jan-1996	401.5	401.5	405	402
09-Jan-1996	404	404	400	395
10-Jan-1996	399.5	399.5	397	395
11-Jan-1996	396	399	405	396
12-Jan-1996	405	403	403	400

Sample data of daily
stock price of a
company



Cross-sectional data

Cross-sectional data are set of observations on two or more variables at the same time point

Produce in the 2019 in (quintals)			
Region	Rice	Wheat	Maize
Region 1	28.76	86.66	8.23
Region 2	84.03	65.45	41.94
Region 3	71.75	28.99	38.43
Region 4	14	68.87	7.04
Region 5	43.97	54.27	86.51

Sample data of a
farm produce in
five regions in the
year 2019

Pooled data

Pooled data is combination of both time series data and cross sectional data.

	Produce in the 2011		Produce in the 2012	
Appliance	Production	Profit (in '000 Rs)	Production	Profit (in '000 Rs)
Radio	280	13	121	8
Television	840	645	1456	5192
Washing Machine	775	2899	152	5706
Computer	876	6887	1005	6349
Air Conditioner	439	5427	328	8095

← Sample data of the
number of appliances
produced in an industry
for two years

Panel data

- Panel data is type of pooled data
- The same cross-sectional unit is surveyed over time
- Also known as longitudinal or micro-panel data

Appliance	Year	Production	Profit (in '000 Rs)
Radio	2010	280	13
Television	2010	840	645
Washing Machine	2010	775	2899
Computer	2010	876	6887
Air Conditioner	2010	439	5427
Radio	2011	234	123
Television	2011	835	645
Washing Machine	2011	564	2899
Computer	2011	874	6887
Air Conditioner	2011	435	5427

Sample data of the number of appliances produced and profits earned by a company over a period of two years

The real STATISTICS?



Which is real STATISTICS?

- ❖ Arjun secured 100/100 in Mathematics
- ❖ The average mark of students in the Mathematics course is 65



The reality?

- Statistics always involve numerical data, but not every piece of numerical data qualifies as statistics.
- For instance, the statement "Arjun secured 100/100 in Mathematics" is simply a numerical fact, not a statistical measure.
- Conversely, "The average mark of students in the Mathematics course is 65" represents a statistical idea because it summarizes data from a group.

Characteristics of Statistics

1. Aggregate of facts
2. Affected by several causes
3. Numerically expressed
4. Enumerated or estimated as accurately as possible
5. Collected in a systematic manner
6. Collected for a planned purpose
7. Comparability

1. Aggregate of Facts

- “Aggregate” - Cannot be a single numerical value
- “Facts” - Only something that can be measured
- Examples
 - Height of a student is 5 feet 4 inches 
 - Heights of students of a certain class 

2. Affected by Several Causes

- A statistic is affected by not a single but multiple causes
- This implies that we cannot study the effect of a particular factor in isolation
- Example
 - Sales of the latest model of an iPhone
 - Can be affected by the number of new features, price, supply, marketing, etc.

3. Numerically Expressed

- A statistic should take a quantitative form and expressed as a numerical figure
- This implies that qualitative phenomenon cannot be statistics
- Example
 - Happiness of employees in an organization ✗
 - Number of employees in an organization who say they are happy ✓

4. Enumerated or Estimated as Accurately as Possible

- Data collected by explicit counting- this will be exact
 - For example, the height of students in a class
 - Cannot leave out a single student
- Data collected by estimation when counting is not possible- accuracy varies
 - For example, average height of children aged 12-15 years in India
 - We cannot count every child in the age group thus we estimate using samples
 - Leaving out 10 children in the age group will not significantly affect the estimate

5. Collected in a Systematic Manner

- Unmethodical and disorganized data collection is undesirable
- It hampers the accuracy of the measurement
- It may lead to us making wrong conclusions

6. Collected for a Planned Purpose

- Before starting the process of data collection, the purpose should be clearly defined
- Resources and time will be wasted in collecting irrelevant data
- The relevant data may not be collected at all or its quality may be affected
- Example
 - We want to measure the number of Non-fiction books sold by a bookstore
 - Counting the number of fiction and comic books is irrelevant

7. Comparability

- Being able to compare different statistics or over is what makes them useful
- Comparability is essential for making inferences, establishing or detecting changes in trends
- Example
 - We measure the heights of students from two different classes
 - We can then compare the average height of each of the two

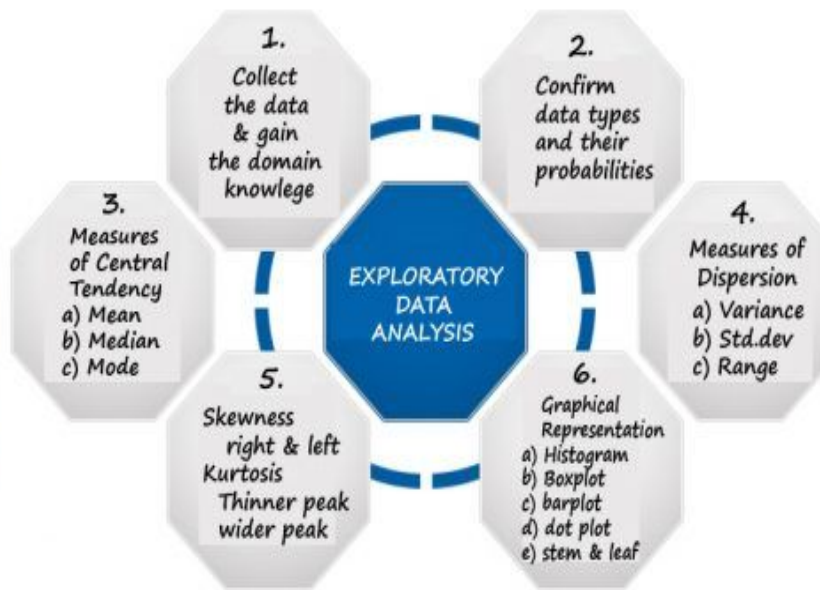
Functions of Statistics

1. Concrete Factual Presentations
2. Reduce complexity of raw data
3. Make comparison and draw conclusions
4. Helps in forming Hypothesis
5. Making Predictions
6. Useful for Decision Making

Limitations of Statistics

1. Limited to quantitative study
2. Aggregate Measurements
3. Homogeneity of data
4. Results are true only on Estimates and not accurate numbers
5. Leads to erroneous conclusions

Exploratory Data Analysis



- Understanding the data
- Identify Patterns and Relationships
- Generate Hypotheses
- Check Assumptions
- Prepare for Modeling

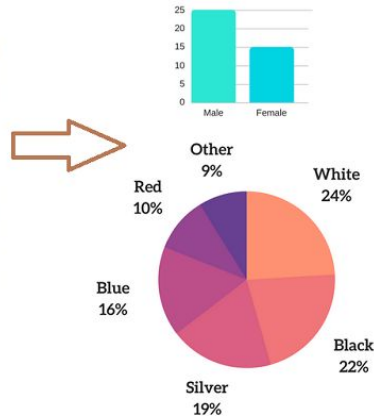
Data Analysis

Descriptive Analysis

- It is used to only describe the sample or summarize information about the sample
- Also known as Descriptive Statistics

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

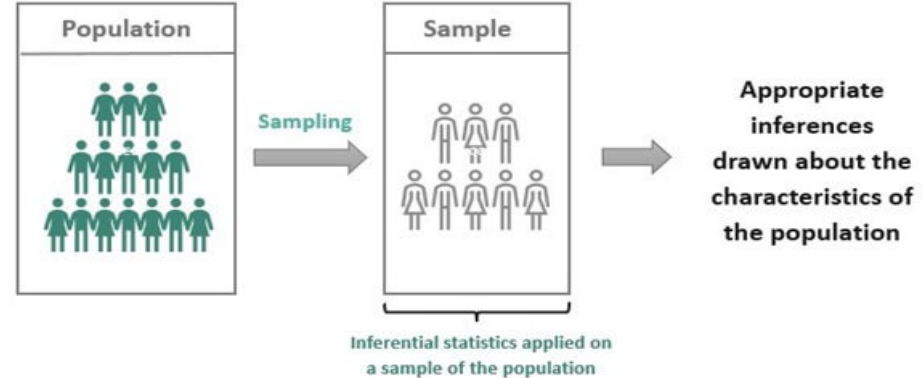
RAW DATA



Descriptive Statistics

Inferential Analysis

- It is used to make inferences and generalizations about the broader population.
- Also known as Inferential Statistics

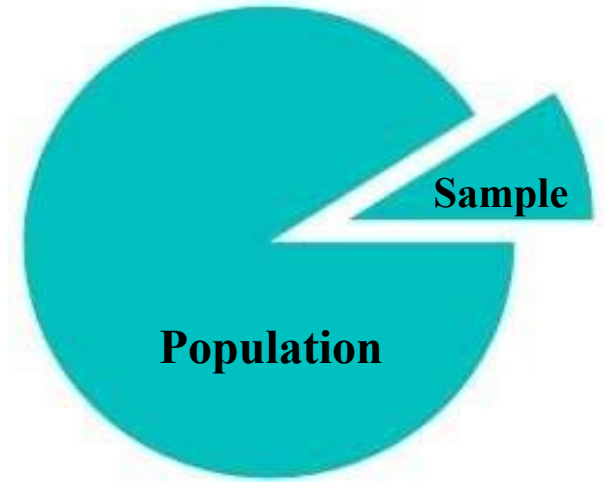


Population & Sample

- **Population** is a collection of all the individuals or objects
- **Sample** is a subset of the population which is a representative of the population

Types of Sampling

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling
- Convenience Sampling



Sampling

- **Sampling** is the process of selecting a subset of individuals, items, or observations from a larger population to make inferences or predictions about the entire population.
- The goal of sampling is to gather data in a way that accurately reflects the population as a whole
- Sampling save time, effort, and resources compared to studying the entire population.

Need to draw a sample

Suppose a company producing electric bulbs wants to know the average life of a bulb. If the details of all the bulbs are available,

then this information is regarded as the population.

It would be challenging for the company to test each and every bulb produced. In such a scenario, the company would draw a sample from the produced bulbs to test.



SAMPLING

Target Population



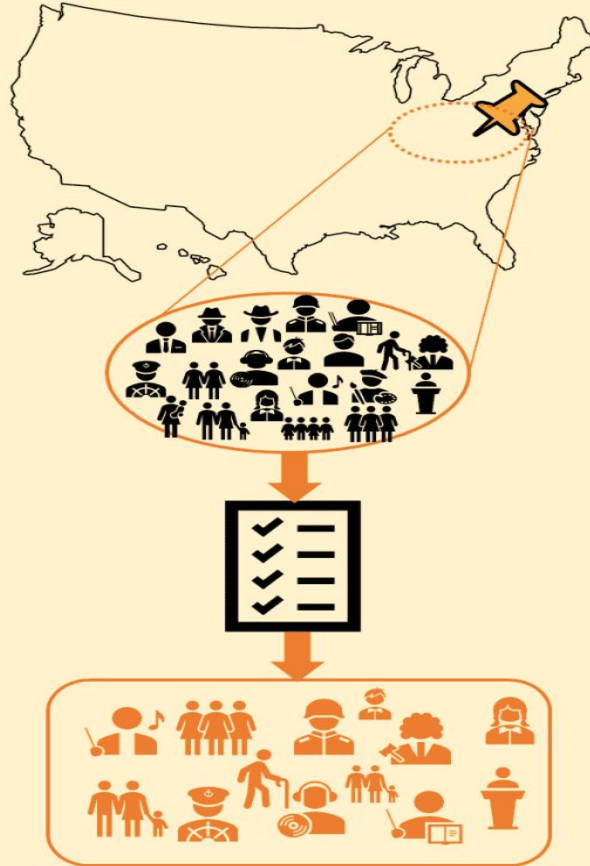
Study Population



Sampling frame



Sample



Generalization
of results



Inference



Reliable and
Valid data



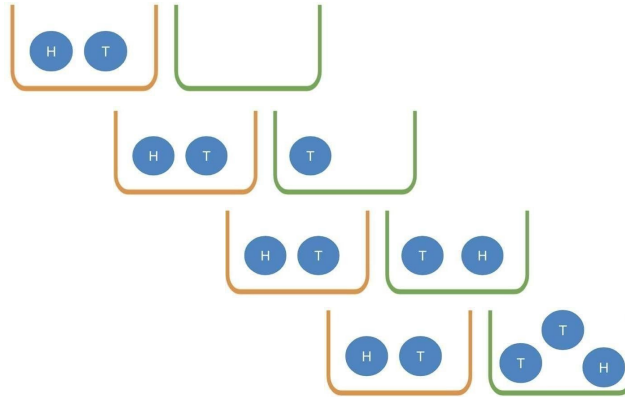
Sample
statistic

Significance of Sampling

1. Resource Optimisation
2. Improved Data Accuracy
3. Facilitates Generalization
4. Ethical Considerations
5. Enhances Feasibility

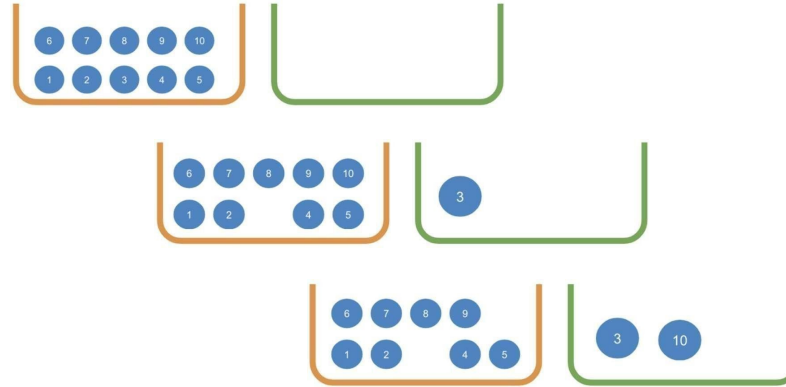
Sample with replacement

The unit selected to be in the **sample** is returned to the **population**.

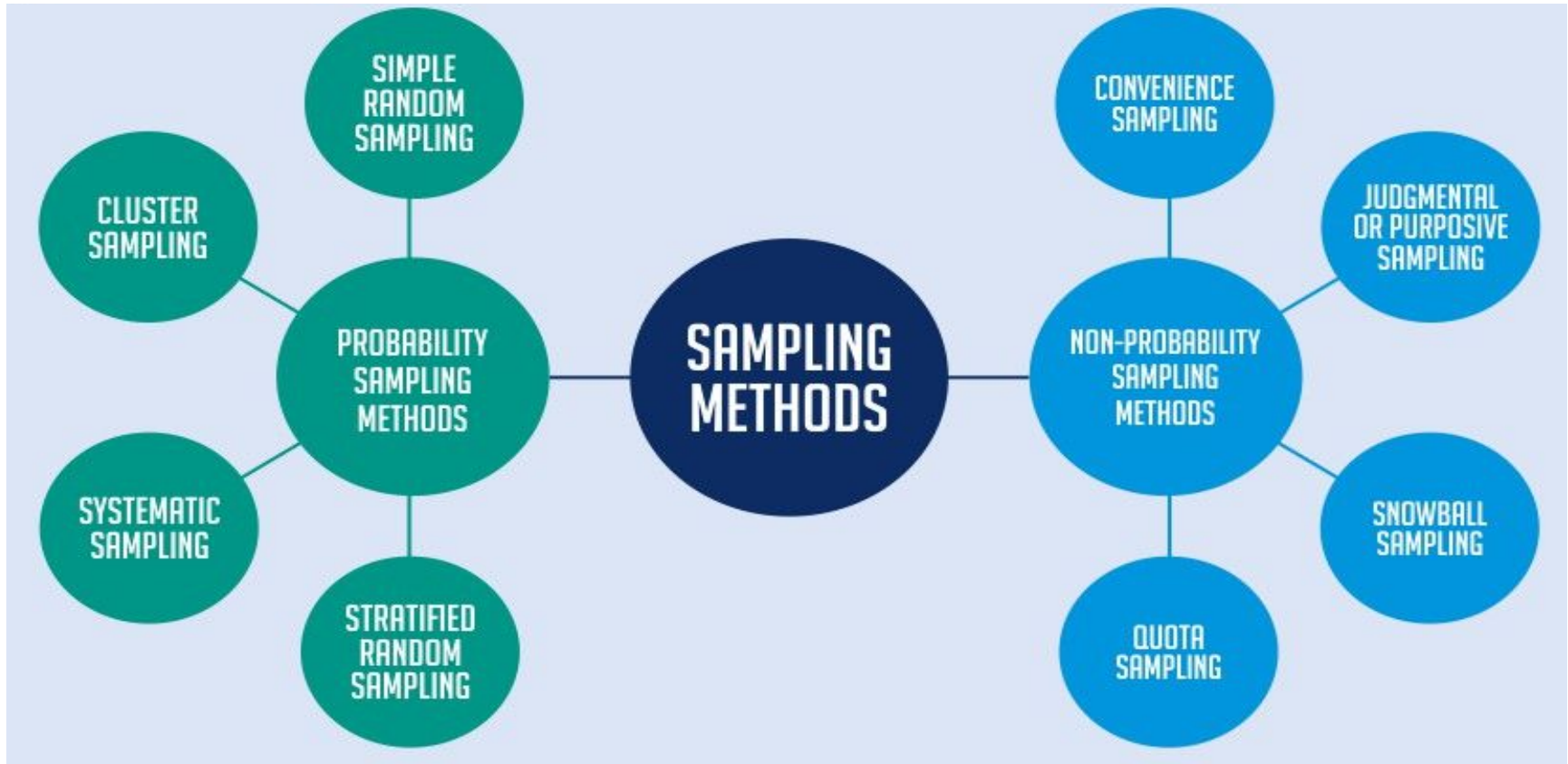


Sample without replacement

The unit selected to be in the **sample** is not returned to the **population**.



Types of Sampling:



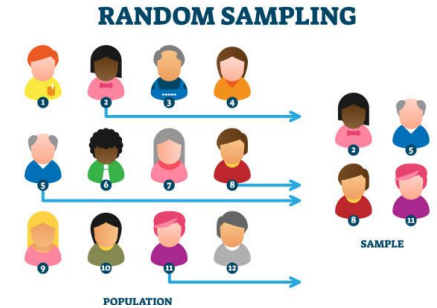
Probability Sampling

- Probability sampling is a sampling technique where a researcher selects a few criteria and chooses members of a population randomly.
- All the members have an equal opportunity to participate in the sample with this selection parameter.

Types of Probability Sampling

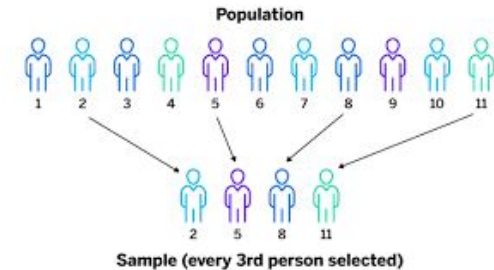
Simple Random Sampling:

- Every member of the population has an equal chance of being selected.
- Respects the principle of independence
- Example: random sample of 5 students from a class of 50 students



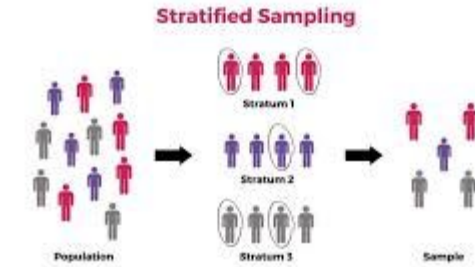
Systematic Sampling:

- Selects every k^{th} individual from a list or sequence.
- Example: Surveying every 3rd person entering a store.



Stratified Sampling:

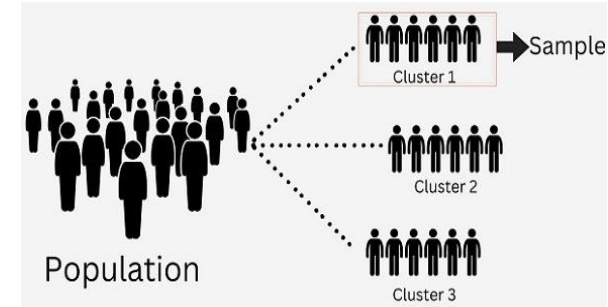
- The population is divided into strata (groups) based on a characteristic (e.g., age, gender), and a random sample is taken from each stratum.
- In Stratified sampling we assume strata is homogenous.
- Example: Ensuring representation of different age groups in a survey.



Types of Probability Sampling

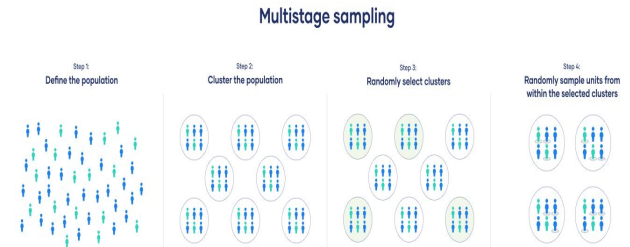
Cluster Sampling:

- The population is divided into clusters (e.g., geographical areas), and a random sample of clusters is selected, with all individuals within those clusters surveyed.
- Naturally formed group with diverse samples within group.
- Example: Selecting certain schools and surveying all students within those schools.



Multistage Sampling:

- Combines several sampling methods, typically involving selecting samples in stages, e.g., selecting clusters first, then using simple random sampling within each cluster.
- Example: First selecting cities, then schools within those cities, and then students within those schools.



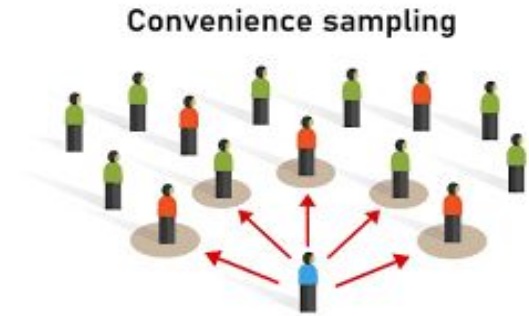
Non-Probability Sampling:

- In non-probability sampling, not every member of the population has a chance of being included, and the selection is often based on convenience or judgment rather than randomness.
- This method is often used when probability sampling is impractical.

Types of Non-Probability Sampling:

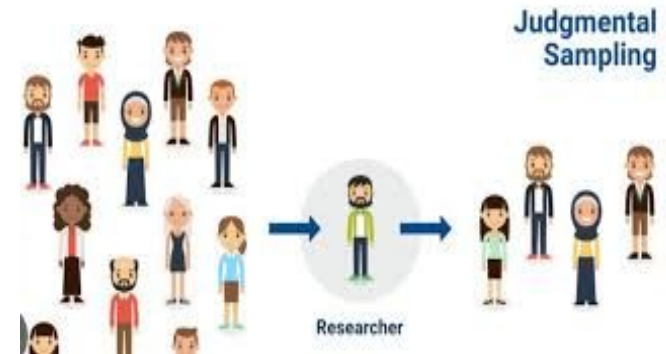
Convenience Sampling:

- The sample is taken from a group that is easy to access.
- Example: Surveying people at a mall because it's convenient.



Judgmental or Purposive Sampling:

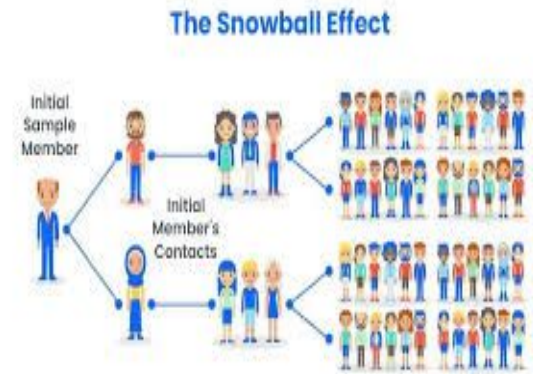
- The researcher selects the sample based on their judgment of what is most representative or typical.
- Example: Choosing experts in a field for a study.



Types of Non-Probability Sampling:

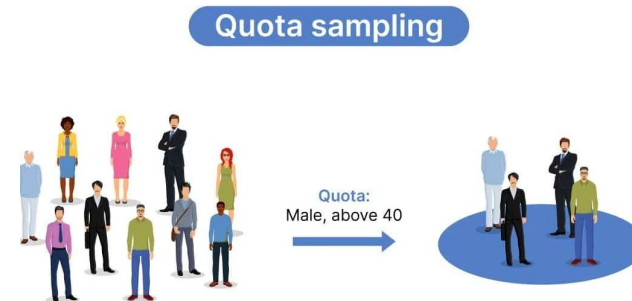
Snowball Sampling:

- Existing study subjects recruit future subjects from among their acquaintances.
- Example: A study on a rare disease where current participants refer others with the same condition.



Quota Sampling:

- The researcher ensures that the sample reflects certain characteristics of the population, but selection within those quotas is non-random.
- Example: Ensuring a sample includes a specific percentage of males and females, but selecting individuals conveniently.



- **Random error**

Random errors are unsystematic, unpredictable, and just as likely to over-as under-estimate the “true” score.

Difference between sample mean and true mean

Can be reduced by selecting large sample size

- **Non - Random error**

Non-random errors are systematic in over or under-estimating the true score

It is error in sampling process while collecting data

1. Sample bias
2. Completion rate
3. Interviewer effect