1. Sampling Techniques

Sampling refers to the process of selecting a group of individuals, items, or observations (sample) from a larger population so that the results obtained from studying the sample can be generalized to the entire population.
Sampling is used because studying an entire population (a census) is often time-consuming, costly, and sometimes impossible.

Key Concepts:

- **Population:** The entire group of individuals or items that we want to study.
- **Sample:** A subset of the population selected for observation and analysis.
- **Sampling Frame:** The list or database from which the sample is drawn.
- **Sampling Unit:** The basic unit containing the element to be sampled (e.g., individual, household, firm).

Types of Sampling Techniques

Sampling techniques are broadly classified into **two categories**:

A. Probability Sampling Methods

In probability sampling, every element in the population has a known and non-zero chance of being selected. This ensures objectivity and representativeness.

1. **Simple Random Sampling (SRS):**
   Simple Random Sampling is the most basic and widely used sampling method. In this method, every individual or item in the population has an equal chance of being selected. This ensures that the sample is unbiased and truly representative of the population.
   - **Methods:** Lottery method, random number tables, computer randomization.
   - **Example:** Selecting 100 customers randomly from a customer database of 5,000.

   Advantages: Eliminates bias, easy to understand.
   Disadvantages: Requires complete list of the population.

2. **Systematic Sampling:**
   Select every k[th] element from a list after choosing a random starting point.

$$k = \frac{N}{n}$$

   where N= population size, n= sample size.

   - Example: If N = 1,000 and n = 100, then k = 10 (select every 10th unit).
   - 10, 20, 30……1000

   In a warehouse of **600 items**, a quality inspector wants to select **50 items** systematically. Find $k$and the selection pattern if the random start is 5.

   Answer:

$$k = \frac{600}{50} = 12$$

   Random start = 5
   Selected items: 5, 17, 29, 41, 53, …, 593

   A supermarket recorded **900 customer bills**. The manager wants to check **60 bills** systematically. Random start is 3. Find $k$and the selection pattern.

   Answer:

$$k = \frac{900}{60} = 15$$

   Random start = 3
   Selected bills: 3, 18, 33, 48, 63, …, 888

   Advantages: Simple and convenient.
   Disadvantages: Risk of bias if there is a hidden pattern in the list.

3. **Stratified Sampling:**
   In stratified sampling, the population is divided into homogeneous groups (strata) based on specific characteristics such as age, gender, income, education, or region. Then, a random sample is drawn from each stratum in proportion to its size in the population.

- **Example:** An organization wants to survey employee satisfaction. Employees are divided into strata based on departments—Marketing, HR, Finance, and Operations. Random samples are drawn from each department to ensure all departments are represented.

Advantages: Ensures representation of all key subgroups.
Disadvantages: Requires detailed population information.

4. **Cluster Sampling:**
   In cluster sampling, the population is divided into heterogeneous groups (clusters), usually based on geographical or administrative boundaries. A few clusters are randomly selected, and either all units within those clusters are studied or a subsample is taken from each cluster.
   - **Example:** Selecting 5 cities out of 100 and surveying all households within those cities.

   Advantages: Cost-effective for large populations.
   Disadvantages: Higher sampling error compared to stratified sampling.

B. Non-Probability Sampling Methods

In non-probability sampling, the probability of selecting each unit from the population is unknown. It relies on the researcher's judgment, convenience, or accessibility rather than random selection.

1. **Convenience Sampling:**
   Convenience sampling involves selecting samples that are easily accessible or available to the researcher.
   - **Example:** Interviewing people at a shopping mall.
     Advantage: Economical and quick.
     Disadvantage: Biased and not representative.
2. **Judgment (Purposive) Sampling:**
   In purposive sampling, the researcher uses personal judgment and expertise to select the most suitable or representative units.
   - **Example:** Selecting key managers for an organizational study.
3. **Quota Sampling:**
   In quota sampling, the researcher divides the population into categories

(like gender, income group, or age) and then selects a fixed number (quota) of samples from each category, usually based on convenience rather than randomness.

- **Example:** Interviewing 50 males and 50 females to ensure gender balance.

4. **Snowball Sampling:**
   Snowball sampling is a method in which existing respondents refer or recruit future respondents. It is particularly useful for studying hidden or hard-to-reach populations.
   - **Example:** Studying drug users, refugees, or people with rare diseases, where participants help identify others with similar characteristics.

2. Sampling Distribution of the Sample Mean

When several random samples of the same size (n) are drawn from a population, each sample will produce its own sample mean ($\bar{X}$). If we plot the distribution of all these sample means, the resulting distribution is called the Sampling Distribution of the Sample Mean.

This concept is essential in inferential statistics, as it forms the basis for estimating population parameters and conducting hypothesis tests.

Properties:

1. **Mean of Sampling Distribution:**

$$E(\bar{X}) = \mu$$

   The mean of the sample means is equal to the population mean.

   Example:

   Suppose the population mean height of students (μ) = 170 cm.

   If we take several random samples of 50 students each and calculate their sample means ($\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, …),

   the average of all those sample means will also be 170 cm.

2. **Standard Error of the Mean (SEM):**

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where

- **σ** = population standard deviation
- **n** = sample size

The Standard Error measures the dispersion or variability of the sample means around the population mean. As the sample size (n) increases, the denominator becomes larger, which reduces the standard error.
Thus, larger samples provide more precise estimates of the population mean.

Example 1:
If a population has σ = 20 and we take a sample of size n = 25,

$$\sigma_{\bar{X}} = \frac{20}{\sqrt{25}} = \frac{20}{5} = 4$$

Hence, the standard deviation of the sampling distribution (the SEM) is 4.

3. Sampling Errors

A **sampling error** is the **difference between the sample statistic (e.g., sample mean $\bar{X}$)** and the **actual population parameter (μ)**.
It arises because a **sample represents only a portion of the population**, not the entire population.

$$\text{Sampling Error} = \bar{X} - \mu$$

Explanation:
Even though samples are drawn randomly, no two samples are exactly the same. Each sample may yield a slightly different estimate of the population mean, creating a **sampling error**.
The smaller the sampling error, the **more accurate** the sample is as a representation of the population.

Example 1:
The population mean income (μ) of all households in a city is ₹60,000. A researcher randomly selects a sample of 100 households and finds the sample mean ($\bar{X}$) = ₹58,500.

$$\text{Sampling Error} = 58{,}500 - 60{,}000 = -₹1{,}500$$

Thus, the sample estimate is ₹1,500 lower than the population mean.

Example 2:
Suppose the average weight (μ) of apples in an orchard is 120 g. A sample of 50 apples shows a mean weight ($\bar{X}$) of 118 g.

$$\text{Sampling Error} = 118 - 120 = -2g$$

This difference arises due to the chance variations in sample selection.

Key Points about Sampling Error:

- It is **unavoidable** when using samples instead of the entire population.
- It can be **reduced** by increasing the sample size.
- It is **random** and can be either positive or negative.
- It **does not imply a mistake**—it's a natural part of sampling.

Types of Errors:

A. Sampling Errors:

1. **Random Errors:**
   Occur by chance; vary from one sample to another.
2. **Systematic Errors (Bias):**
   Result from faulty sampling design, improper selection, or non-randomization.

B. Non-Sampling Errors:

1. **Measurement Errors:** Wrong recording or misunderstanding of data.
2. **Non-Response Errors:** Missing information due to unavailability or refusal.

3. **Processing Errors:** Mistakes in coding or tabulation.
4. **Coverage Errors:** Excluding certain groups unintentionally.

Ways to Minimize Sampling Errors:

- Use random sampling methods.
- Increase sample size.
- Train survey personnel properly.
- Conduct pilot surveys to refine questionnaires.

4. Central Limit Theorem (CLT)

The Central Limit Theorem states that when a large number of independent random samples are taken from a population with any distribution having a mean (μ) and a finite standard deviation (σ), the **distribution of sample means** ($\bar{X}$) tends to be **normal**, with mean μ and standard deviation $\sigma/\sqrt{n}$.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

For large samples (n ≥ 30), this approximation holds even if the population itself is not normally distributed.

Importance:

- It allows the use of normal distribution theory for inferential statistics.
- Enables construction of confidence intervals and hypothesis testing.
- Simplifies analysis of non-normal data.

Example:

A population has μ = 50 and σ = 10. For samples of size n = 100:

$$\sigma_{\bar{X}} = \frac{10}{\sqrt{100}} = 1$$

Thus, $\bar{X}$will be approximately normally distributed with mean = 50 and SD = 1.

1. Estimators and Their Properties

Meaning of Estimator

An **estimator** is a **statistical rule, measure, or formula** used to estimate an **unknown population parameter** using information from a **sample**.
The **numerical value** computed from the sample is called an **estimate**, while the **formula or method** used to compute it is called an **estimator**.

Examples:

Population Parameter	**Symbol**	**Estimator**	**Formula**
Population Mean	μ	Sample Mean ($\bar{X}$)	$\bar{X} = \frac{\sum X_i}{n}$
Population Proportion	p	Sample Proportion ($\hat{p}$)	$\hat{p} = \frac{x}{n}$
Population Variance	σ^2	Sample Variance (s^2)	$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$

Types of Estimators

1. **Point Estimator:**
   Provides a single best estimate of a population parameter.
   Example:

   If a sample of 50 students has an average height ($\bar{X}$) of 165 cm, then 165 cm is a point estimate of the population mean height (μ).

2. **Interval Estimator:**
   An interval estimator provides a range of values (confidence interval) within which the true population parameter is expected to lie, with a certain level of confidence (e.g., 95%).
   Example: If the average test score of a sample is 150, and the margin of error is ±5,

   then the 95% confidence interval for the population mean (μ) is: $145 \leq \mu \leq 155$

Properties of a Good Estimator

A good estimator should satisfy the following properties:

1. Unbiasedness

An estimator is **unbiased** if the expected value of the estimator equals the true population parameter.

$$E(\hat{\theta}) = \theta$$

Example:
The sample mean ($\bar{X}$) is an unbiased estimator of population mean (μ).

2. Consistency

An estimator is **consistent** if it approaches the true value of the parameter as the sample size increases.

$$\hat{\theta}_n \xrightarrow{p} \theta$$

Example:
As n → ∞, the sample mean converges to the population mean.

3. Efficiency

An estimator is **efficient** if it has the **smallest variance** among all unbiased estimators of a parameter.
Efficient estimators make use of all available information and provide more reliable estimates.

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2) \Rightarrow \hat{\theta}_1 \text{ is more efficient}$$

4. Sufficiency

An estimator is **sufficient** if it uses all relevant information about the parameter contained in the sample data.
Example:
The sample mean is a sufficient estimator for μ in a normal population.

5. Minimum Variance Unbiased Estimator (MVUE)

An estimator that is both **unbiased** and has **minimum variance** among all unbiased estimators.