Sampling

Sampling:

Sampling is a process of selecting a suitable sample or a representative part of a population for purpose of determining parameters or characteristics of the whole population.

The word ‘population’ or ‘universe’ means an aggregate of all objects, animate or inanimate, about which an information is required.

Population may be a finite (e.g., patients in a hospital, faculty members in a medical college) or an infinite (e.g., microbial organisms in a river, fishes in a pond) depending on whether individual elements are countable or uncountable.

It is not possible for any scientific study to cover the whole population because of the cost, time and practicability. So a representative portion of the population is taken for the study. It is called a sample.

A sample is a finite subset or a representative part of a population. The number of individuals (or units) in a sample is called sample size.

The observation made out of a representative sample is applied to the universe at large, but generalization is valid, only if the sample is sufficiently large and representative.

Population units selected as sample are called sampling units. The list of population units from which the sample units are to be selected is called sampling frame.

The ratio of sample size to the size of the target population is called sampling ratio. If the population size is 20,000 people and a sample of 200 people is selected, the sampling ratio is 200/2000 = 0.01 or 1%.

The population parameter refers to a value calculated from a defined population, such as mean (μ) and standard deviation (σ). It is a constant value because it covers all the members of the population. A value calculated from a sample is called statistic such as mean ($\bar{x}$), standard deviation (s)and proportion (p).

Objectives of Sampling:

- **To obtain information about population characteristics based on samples**
- **To test the hypothesis about the population from which the sample or samples are drawn.**
- **To estimate the population parameters (mean, proportion, etc.) from sample statistics**

Characteristics of a Representative Sample:

- **It is selected by sampling technique from the population it represents**
- **Usually it does not differ from the population in composition. If it differs, it is solely by chance.**
- **Each unit of the population has equal chance of being selected in the sample.**
- **Bias has been ruled out and the sample will give a estimate of the attribute understudy, almost equal to the population value called ‘true value’**

- **Sample should be sufficiently large in size to represent the population from which it is drawn.**

- **It is chosen according to rule that is independent of the observations to be made in the sample.**

- **The two main characteristics of representative sample are precision and unbiased character. Precision depends on the sample size. Unbiased statistics of sample like mean will be close to population mean.**

Census Vs Sampling:

	Census	Sample Survey
1	Investigation is carried out on the entire population	Investigation is carried out only for a sample
2	Useful when detailed information is required	Useful when overall information is required
3	Organization is costly and consumes more time	Organization is cheap and consumes less time
4	Greater attention can not be paid for each unit because of the vastness of the population	Greater attention can be paid for each unit because of relative smallness of the sample
5	Difficult to achieve a complete and accurate study	Easy to achieve a complete and accurate study
6	Require more personnel (trained and experienced) for the study	Require less personnel(trained and experienced) for the study

Types of Sampling Techniques:

1. Probability Sampling (Random Sampling)

These are employed when the sampling frame is available. The methods of probability sampling are:

- **Simple Random Sampling**
- **Stratified Sampling**
- **Systematic Sampling**
- **Cluster Sampling**
- **Multistage Sampling**
- **Multiphase Sampling**

2. Non-probability Sampling (Non-random Sampling)

These are employed when the sampling frame is not available. Some of the methods of non-probability sampling are:

- **Convenience Sampling**
- **Purposive / Judgment Sampling**
- **Quota Sampling**
- **Snowball sampling**

Probability Sampling (Random Sampling):

1. Simple Random Sampling:

This method is applicable when the population is small, homogeneous and readily available such as patients coming to hospital or lying in the wards. It is used in experimental medicine or clinical trials like testing the efficacy of particular drug. The principle here is that every unit of the population has an equal chance of being selected.

The sample may be drawn unit by unit, either by numbering the units such as persons, families or households of a particular population on the cards or from the published table of random numbers. To ensure randomness of selections, one may adopt either lottery method or refer to table of random numbers.

Merits:

i) Free from personal bias

ii) Sampling error can be calculated.

Demerits:

i) Not suitable for large and heterogeneous population

ii) Time consuming and costly method

2. Stratified Random Sampling:

This method is followed when the population is not homogeneous. The population under study is first divided into homogenous groups or classes called strata and the sample is drawn from each stratum at random in proportion to its size.

It is a method of sampling for giving representation to all strata or society or population such as selecting sample from defined areas, classes, ages, sexes, etc. This technique gives more representative sample than simple random sampling in a given large population.

Merits :

i) Proportionate representative sample from each strata is secured

ii) It gives greater accuracy.

Demerits:

i) This method is not effective if each stratum does not contain homogenous units

3. Systematic Sampling:

This method is popularly used in those cases when a complete list of population from which sample is to be drawn, is available. It is more often applied to field studies when the population is large, scattered and not homogenous. Systematic procedure is followed to choose a sample by taking every K^{th} house or patient where K refers to sample interval, which is calculated by the formula:

$$K = \frac{\text{Total population}}{\text{Sample size desired}}$$

For example: If 10% sample is to be taken out of one thousand patients, the sample interval K will be:

$$K = \frac{1000}{10\% \text{ of } 1000} = 10$$

One number chosen from 01 to 10 using lottery method or two digits number from table of random numbers is 6 (say) then sample will consists of units with sample number 6, 16, 26, 36, 46 and so on. Select every 10^{th} house after the 6^{th} house. Such as this may apply in assessing the incidence of influenza in an epidemic in a large city.

Merits:

i) Simple and convenient method

ii) Less time and labor required

iii) Give accurate results in large and homogenous sample

Demerits:

i) It may give biased results when there are periodic features.

4. Cluster Sampling:

A cluster is a randomly selected group. This method is used when units of population are natural groups or clusters such as villages, wards. Blocks, slums of a town, children of a school etc From the chosen cluster entire population is surveyed. WHO module has stressed its usefulness in evaluation of vaccination coverage in expanded immunization program.

Merits: i) Data collection in this method is simpler and involves less time and cost.

Demerits: i) It gives higher standard error.

5. Multistage Sampling:

As the name implies, this method refers to the sampling procedures carried out several stages using random sampling technique.

This is employed in large country survey. As for example, the districts are chosen randomly in the first stage, the VDCs and municipalities from within the chosen district are selected randomly in second stage, and wards within selected VDCs and municipalities are chosen in third stage and number of households in the fourth stage.

Merits: i) It introduces flexibility in sampling

ii) It enables use of existing division and subdivision which saves time

Demerits: i) It is very expensive and involves time and labor in collection and processing of data due to nature of large scale survey.

6. Multiphase Sampling:

In this method, part of the information is collected from the whole sample and part from the subsample. There will be several phases for information collection. The number of subsamples in each phase will become successively smaller and smaller.

As for example take a case of TB survey:

1st phase: Physical exam. or Mantoux test to all cases

2nd phase: Chest X-ray to Montoux positive patients

3rd phase: Sputum test to chest x-ray positive patients

Non-Probability (Non-Random) Sampling:

1. Convenience Sampling:

Convenience sampling is a method in which for convenience sake the study units (or the subjects) that happen to be available at the time of data collection are selected in the sample. This method is also referred to as accidental sampling.

The convenience sampling is done when population is not clearly obtained or when sampling frame is not available. It is relatively a quick and low cost process but main disadvantage is that it is not representative.

As for example, a researcher wants to study the attitudes of villages toward family planning services provided by the health centre. He may decide to interview all the adult patients who visit the out-patient clinic during one particular day. This is more convenient than taking a random sample of people in the village.

2. Purposive Sampling:

This is also known as judgmental sampling. In this method, the attitude is quite different. Recognizing that the population may well contain different types of individuals or items, with differing measures and ease of access, the researcher exercises deliberate subjective choice in drawing what he / she regards as a 'representative' sample.

This is a sampling procedure in which an experienced researcher selects the sample based on some appropriate characteristic of sample members to serve a purpose of the study.

3. Quota Sampling:

The population is first divided into different categories with specific characteristics. There are different subgroups of population from which samples are to be drawn. The number of units to be selected from each subgroup are fixed as quota with specific criteria. The categories for which quotas are to be used and the quotas to be allocated are determined based on the issue to be addressed. Common criteria are age, gender, occupation and whether a people live in project or non-project areas, etc.

4. Snowball Sampling:

Some medical condition carry stigma. Cases of these conditions are difficult to locate. For example, people exposed to sex worker or those injecting drugs in the context of HIV are difficult to identify. But if one person is detected, he can help identify others of his clan. They generally know each other but are otherwise not known to be exposed to the risk factor. Those identified by him are asked to identify others and so on. This is called snowball sampling.

Errors in sampling:

1. Sampling Error:

Sampling errors can be attributed to selection of sampling procedure, the study / research design and the sample design.

Sampling errors may arise due to various reasons that include:

- **faulty selection of sampling**
- **faulty demarcation of sampling units**
- **substitution**
- **improper selection of sampling statistic**

2. Non-Sampling Error:

It includes errors attributed to acquiring, recording, and tabulating the data. Non-sampling errors are due to following reasons:

- **faulty definition / planning**
- **responses error**
- **non-responses**
- **errors in coverage**
- **compiling errors**
- **publication errors**

Determination of Sample Size:

An optimum size of the sample is usually a compromise of what is desirable and what is feasible?

The feasible sample size is determined by the availability of resources in terms of cost, manpower and time.

The following points are taken into account while in determining the desirable sample size.

- **Variance or heterogeneity of population**
- **The degree of acceptable error (confidence interval)**
- **Confidence level**
- **Judgments on all related variable characteristics**

1. Calculation of sample size for quantitative data

a) When population is infinite

$$n = \frac{Z^2 S^2}{E^2}$$

Where,

n = Size of sample

Z = The value of SNV at a given confidence level
1.96 for a 95% confidence interval.

E = Permissible error in the estimation of mean

S = Standard deviation

b) When population is finite

$$n = \frac{Z^2 N S^2}{E^2(N-1) + Z^2 S^2}$$

Where, N : Size of the population

1. Calculation of Sample Size for Qualitative Data

a) When population is infinite

$$n = \frac{Z^2 \, p \, q}{E^2}$$ Where,

n = Size of sample

Z = The value of SNV at a given confidence level 1.96 for a 95% confidence interval.

E = Permissible error in the estimate of p

p = Sample proportion or percentage of incidence or prevalence: q = 1 - p

b) When population is finite

$$n = \frac{Z^2 pqN}{E^2(N-1) + Z^2 pq}$$ Where, N : Size of the population

Exercise:

1. A clinician would like to explore the mean fasting blood glucose value of diabetic patients at BS hospital over the past ten years. We are interested in determining the number of records that the clinician should examine in order to obtain a 95% confidence interval for μ (the population mean fasting blood glucose) if the desired width of interval is 6 units and a pilot sample yields a variance of 60. Also we want to find the sample size that should be taken to examine if the total number of records is 5000.

Solution

Determination of sample size for estimating mean

a) Population is infinite

Z = 1.96 (for a 95% confidence interval)

E = Margin of error = ± 3 (given the length of the interval 6 units)

S^2 = Variance (sample / population) = 60

$$n = \frac{Z^2 S^2}{E^2} = \frac{1.96^2 x 60}{3^2} = \frac{230.50}{9} = 25.61$$

The sample size or the number of records to be examined = 26

b) Population is finite

Here, N = 5000

$$n = \frac{Z^2 N S^2}{E^2 (N-1) + Z^2 S^2} = \frac{1.96^2 x 5000\ x 60}{3^2 (5000\ -1) + 1.96^2 x 60} = \frac{1152480}{45221.5} = 25.49$$

The number of records to be examined from 5000 records is 26

2. A hospital administrator wishes to know what proportion of discharged patients is unhappy with the care received during hospitalization. He is interested to find out the answer regarding this - how large a sample should be taken if e = 0.05, the confidence coefficient is 0.95 and p (proportion of unhappy patients) informally reported as at 25 percent. Also determine the sample size if the patients discharged from the hospital in recent year were 12600 in numbers.

Solution:

Determination of sample size for estimating proportion

a) Population is infinite

p = 25% = 0.25

q = 1 - p = 1 - 0.25 = 0.75

Z = 1.96 (for a confidence coefficient 0.95)

E = 0.05

$$n = \frac{Z^2 pq}{E^2} = \frac{1.96^2 x 0.25 x 0.75}{0.05^2} = \frac{0.7203}{0.0025} = 288.12$$

The required sample size is 288

b) Population is finite

Here, N = 12600

$$n = \frac{Z^2 pqN}{E^2(N-1) + Z^2 pq} = \frac{1.96^2 x 0.25 x 0.75 x 12600}{0.05^2(12600 - 1) + 1.96^2 x 0.25 x 0.75} = \frac{9075.78}{32.22} = 281.68$$

The required sample size is 282. The sample should include 282 discharged patients.

3. Mean pulse rate of a population is believed to be 70 per minute with a standard deviation of 8 beats. Calculate the minimum size of sample to verify this, if allowable error E = ±2 beat at 5% risk. What will be the size of sample if allowable error increase to ±4 beats at 5% risk?

$$\text{Hints :} \quad n = \frac{Z^2 S^2}{E^2} = \frac{1.96^2 x8x8}{2x2} = 62$$

4. Hookworm prevalence rate was known to be 30% in a community. Calculate the size of sample required to find the prevalence rate now if allowable is 0.05 and 0.06.

$$\text{Hints :} \quad n = \frac{Z^2 \, p \, q}{E^2} = \frac{1.96^2 x0.3x0.7}{0.05x0.05} = 323$$