Theory of Sampling

Unit-1

Basic Terminology

The five words population, sample, parameter, statistic, and variable are the basic terminology of sampling, are described as follows.

1. Population:

In general it is a number of People residing in a given area. But in statistics it is totality of objects having certain characteristics.

In Statistics, population has a much broader meaning. It does not only refer to people but also the group of elements or units under consideration by the analyst. Thus, population is the collection or group of individuals /items /units/observations under study.

The total number of elements / items / units / observations in a population is known as population size and denoted by N. The characteristic under study may be denoted by X or Y.

Example:

- Population of Mumbai people who are currently registered to vote,
- Annual income of people residing in a given area.
- Heights of students doing graduation from a given college
- IQ level of students selected of engineering course

We may classify the population into two types as:

Finite Population: If a population contains finite number of units or observations, it is called a finite population. For example, population of students in a class, the population of bolts produced in a factory in a day, the population of electric bulbs in a lot, the population of books in a library, etc. are the finite populations.

Infinite Population: If a population contains an infinite (uncountable) number of units or observations, it is called an infinite population. For example : Population of real numbers between 1 and 2.

Real population: A population comprising the items or units which are all physically present is known as real population.

Hypothetical Population: If a population consists the items or units which are not physically present but the existence of them can only be imagined or conceptualized then it is known as hypothetical population.

2. Sample:

A part of the population selected according to some rule for drawing conclusions about the target population.

i.e It is a finite subset of population and expected to be a representative of the population.

"A sample is a part / fraction / subset of the population."

Example:

- if we take handful of wheat or rice from 100 kg bag ,we expect the same quality of wheat in hand as inside the bag.
- It is expected that a drop of blood will give same information as well as the blood in the body
- The patients selected to fill out a patient-satisfaction questionnaire,
- 10 bulbs selected from an electric bulb factory's production line.

3. Variable:

A Characteristic of an individual that will be analyzed using statistical tools.

Example:

Gender, Household income Individuals height, weight, Blood pressure, blood sugar ..etc

4. Parameter:

The characteristics of a population can be described with some measures such as total numbers of elements in the population, Population mean, Population standard deviation, population variance, Population proportion, population correlation…. etc. These measures are known as parameters of the population.

We know that the population can be described with the help of distribution and the distribution is fully determined with the help of its constants such as, in case of normal distribution, we need to know μ and σ^2 to determine the normal distribution, in case of Poisson distribution, we need to know λ, etc. These unknown constants are known as **parameter.**

4. Statistic:

Any function obtained by using sample values or observations only and does not contain any unknown parameter is known as statistic. It is denoted by t = t (x_1, x_2.....x_n) .For example, if x_1, x_2.....x_n is a random sample of size n taken from a population with mean μ and variance σ^2 (both are unknown) then sample mean $\bar{x} = \frac{\sum x_i}{n}$ is a statistic. But $\bar{x} - \mu$ and $\frac{\bar{x}}{\sigma}$are not statistics because both are function of unknown parameters.

5. Estimator and Estimate (Guess)

Generally, population parameters are unknown and the whole population is too large to find out the parameters. Since the sample drawn from a population always contains some or more information about the population, therefore in such situations, we guess or estimate the value of the parameter under study based on a random sample drawn from that population.

So, if a statistic, t = t (x_1, x_2.....x_n) (function of sample observations) is used to estimate a population parameter then it is known as estimator and the value of the estimator is known as estimate of parameter.

Example:

- Sample means, Sample SD or SE, Sample Correlation......

  Suppose, if we want to estimate the average height (μ) of students in a college with the help of sample mean $\bar{x}$then $\bar{x}$ is the estimator of μ and its particular value, say, 165 cm is the estimate or guess of the population's average height (μ).

5. Bias estimator:

If θ be the unknown constant (Parameter) of a probability distribution f(x,θ) of population under study and t is an estimator of θ, then t is said to be biased estimator if

$$E(t) \neq \theta$$

Amount of bias = $E(t) - \theta \neq 0$

Amount of bias of an estimator gives us an idea that on average how far our guess is from true value of θ.

If $E(t) - \theta > 0$, then estimator is said to be positively biased (over estimation)

If $E(t) - \theta < 0$, then estimator is said to be negatively biased (under estimation)

6. Unbiased Estimator: A statistic t =t(x_1,x_2.....x_n) is said to be unbiased estimator of parameter θ iff

$$E(t) = \theta$$

Example : Let there is a population having four units {2,4,6,8} and if we want toselect a sample of size 2 from the population , then there would be

(1) ${}^{N}C_n$=^{4} C_2 = 6 possible samples can be drawn if ordering and repletion of units are not allowed.

(2) N^n = 4^2 = 16 possible samples can be drawn if ordering and repletion of units are allowed.

Population mean ⍰ $= \frac{2+4+6+8}{4} = 5$

If X_1 : Number selected at the first draw, X_2: Number selected at the Second draw

Then following distribution can be drawn:

S.N	(X_1, X_2)	$t = \frac{X_1 + X_2}{2}$
1	(2,4)	3
2	(2,6)	4
3	(2,8)	5
4	(4,6)	5
5	(4,8)	6
6	(6,8)	7

$$E(t) = \frac{t_1 + t_2 + \cdots \ldots \ldots . t_6}{6} = 5 = \mu$$

Here we observed that

$$E(t) = \mu$$

And say that sample mean $\bar{x}$ is unbiased estimator of µ.