Unit VI: Sampling methods

Sampling

Sampling is the process by which inference is made to the whole by examining only a part. In other words it is the process of getting information about the totality or universe or population by performing examination of only some parts of the population under investigation. For example, a doctor tests a drop of blood of a patient to know about the characteristic of the whole blood of that patient.

Objective of sampling

The major objectives of sampling techniques are as follows:

- To obtain maximum information about the population under consideration by examining number picked up in the sample
- To ascertain confidence interval to the estimate of the population parameter.
- To test the significance of the population parameter at given level of significance.

Population

In statistics, the term population or universe refers to the all totality of cases (or items) under investigation.

There are three types of population: finite, infinite and hypothetical population.

Finite population If the number of units constituting the population is fixed then it is known as finite population. For example, the number of leprosy patients in a community, all HIV positive persons, all diabetic patients etc.

Infinite population If the population contains an infinite number of members it is called infinite population. For example, all possible hemoglobin (Hb) values in a given interval, all possible height within the range 150 cm to 165 cm etc.

Hypothetical population A hypothetical population is one which is assumed for theoretical purpose. Suppose a few guinea pigs are given vitamin A-deficient diet and they are watched for vitamin A deficiency symptoms within a predetermined period, the result observed is generalized to all such groups of guinea pigs fed on a vitamin A-deficient diet. All such groups constitute a population. This is only a imaginary population which exist hypothetically.

Sampling frame

Sampling frame is a list of sampling elements with identification particulars or a map showing the positions of the sampling elements. A Sampling frame represents the population under investigation. This sampling frame is the base for drawing a sample and therefore, it should be made up to date, i.e. free from omission and duplications.

Sample

Some units selected from the population is known as sample and the process of selecting some units from the population in order to draw conclusion about the population is known as sampling.

Census survey
The complete enumeration of all units of population is known as census survey. It is the process of complete enumeration in which every member of the defined population is included.

Parameter and statistic
Parameter is the value of a variable or attribute calculated from the population under study. The parameter is generally unknown and is fixed. Statistic refers to that value of a variable or attribute calculated from a sample taken out of a population. Statistic may vary depending upon the sample. But the average of statistic is always equals to the parameter.

Characteristic	Parameter	Statistic
No. of units	N	n
Mean	μ	$\bar{x}$
Standard deviation	σ	S
Proportion	P	p
Correlation coefficient	ρ	r
Regression coefficient	β	b

Sampling and non-sampling errors
The errors occurred in the process of collecting, processing and analyzing data may be classified into the following types: sampling errors and non-sampling errors.

Sampling errors
Different samples selected from the population will give different results as the elements included in the sample will be different. This will give rise to the sampling errors. Because of these errors there may be difference between sample mean and population mean. These errors or biases are because of number of reasons. Some of these reasons will be as follows:

- Incomplete or faulty selection of samples
- By substituting certain units whose characteristics are not homogeneous with the characteristics of the original sampling units
- Applying improper statistics for estimating the population parameter.
- Not using proper sampling design.

Non-sampling errors
Non-sampling errors can occur at the different stages of observation, ascertainment and processing of data in both complete enumeration survey and the sample survey. Hence it should be noted that data collected through the complete census may possess non-sampling error whereas data collected through the sample survey may have possessed both sampling and non-sampling errors.
Non-sampling errors can occur at different stages of planning and implementing of census or sample survey. It is quite difficult to make complete list of all the source of non-sampling errors. However, following are some important factors involved in non-sampling errors.

- Inappropriate planning or definitions
- Response and non-response errors
- Compiling and publication errors etc.

Method of sampling

The method of selecting a sample is of fundamental importance in sampling theory and usually depends upon the nature of investigation. The sampling procedures, which are commonly used, may be broadly classified under the following heads.

1. Probability sampling (random sampling)
2. Non probability sampling (non random sampling)

Probability sampling

Probability sampling is the scientific method of selecting samples according to some laws of chance in which each unit in the population has some definite pre assigned probability of being selected in the sample. Some of the probability sampling methods are:

1. Simple random sampling
2. Stratified sampling
3. Systematic sampling
4. Cluster sampling
5. Multistage sampling

Simple random sampling

The simplest and common most method of sampling is the simple random sampling (SRS) in which the sample drawn is unit by unit with equal probability of selection for each unit at each draw. Therefore, SRS is method of selecting 'n' units out of a population of size 'N' units by giving equal probability to all units. Sampling in this method can be done either with replacement or without replacement.

Sampling with replacement

The unit drawn in the first draw is replaced before the second draw. The possible samples of size 'n' out of 'N' units of the population is N^n and the probability of selecting each sample is $\frac{1}{N^n}$

Sampling without replacement

The unit drawn in the first draw is not replaced before the second draw. The possible samples of size 'n' out of 'N' units of the population is ${}^{N}C_{n}$ and the probability of selecting each sample is $\frac{1}{{}^{N}C_{n}}$.

Procedures of selecting random sample

i. Lottery method ii. Use of random number tables

Stratified sampling

When the population characteristics are heterogeneous, the SRS does not serve as a good design so as to represent the sample units from each characteristic. Then the entire population is divided into different groups called strata in such a way that within strata they are homogeneous in nature and between strata heterogeneous. Then a simple random sampling procedure is used to draw samples from each stratum.

Systematic sampling

Systematic sampling is a commonly used technique if a complete and up to date sampling frame is available. In this sampling the first unit is selected randomly and the remaining units are selected automatically with some predetermined pattern. Suppose N units of the population are numbered from 1 to N in some order. Let N = nk where n is the sample size and k is an integer known as sampling interval and a random number called random start is selected between 1 to k. Then every k^{th} unit will be selected automatically.

For example, let us consider the population of size 100 starting from 1 to 100. Let us take a sample of size 10. To select 10 samples out of 100, a number is selected at random and other 9 numbers are selected at equal sampling interval.

Sampling interval (k) = N/n = 100/10 = 10

Select a number (which is less than or equal to k), say a number 3 is taken first, then other numbers will be 13, 23, 33, 43, 53, 63, 73, 83 and 93.

Cluster sampling

This sampling method is also appropriate if the population characteristics are not homogeneous. In this method, the population is classified into sub population, called cluster in such a way that the characteristics within the cluster are heterogeneous and between cluster homogeneous. A cluster is then selected for the desired sample.

Multistage sampling

As the name suggests, multistage sampling refers to the sampling technique which is carried out in various stages. Multistage sampling consists in sampling first stage units by suitable methods of sampling. From among the selected first stage units, a sub sample of secondary stage is drawn by some suitable method of sampling which may be same or different from the method used in the first stage. Further stages may be added to arrive at a sample of desired sampling units.

Non probability sampling

This is the method of selecting samples, in which the choice of sampling units depends entirely on the judgment of the sampler. This method is mainly used for opinion surveys, but can not be recommended for general use at it is subject to the drawbacks of prejudice and bias of the investigator. The types of non probability sampling are:

1. Judgment or purposive sampling
2. Convenience sampling
3. Quota sampling

Judgment or purposive sampling

A sampling method, in which the researcher selects the sample according to personal judgement, is called purposive sampling. This method is suitable only when a universe is small and a quick decision is needed. For example, medical representative contacts to the popular and busy doctor purposively. This method gives valid results when used properly i.e. if the researcher is skilled and apply the method unbiasedly otherwise because of personal judgment, biases and prejudices lead improper conclusions.

Convenience sampling
The investigator selects the samples on the basis of the convenience of the investigator. This is also known as chunk sampling.

Quota sampling
Some quotas are set up according to some criteria and selection of quota is made accordingly to the personal judgment of the investigator. In this method, the investigator is told in advance the number of the sample units he is to enumerate.

Snowball sampling
This technique is used by the researchers to identify the potential subjects in studies where the subjects are hard to locate. This type of sampling technique works like chain referral. So it is also known as chain sampling or chain- referral sampling or referral sampling. After observing the initial subject, the researcher asks for the assistance from the subject to help identify people with a similar trait of interest.

Distribution of sample mean and proportion

Sampling distribution of statistic
If we draw a sample of size n with simple random sampling without replacement (SRSWOR) from a given finite population of size N, then the total number of possible samples will be

$${}^{N}C_{n} = \frac{N!}{n!(N-n)!} = k \text{ (say)}$$

For each of these k values we can compute statistic $t = t(x_1, x_2, \ldots, x_n)$. Say the statistic is mean, standard deviation etc. which are shown below.

Sample no.	1	2	3	.	.	.	k
Statistic t	t_1	t_2	t_3	.	.	.	t_k
Mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	.	.	.	$\bar{x}_k$
Variance	S_1^2	S_2^2	S_3^2	.	.	.	S_k^2

The set of values the statistic so obtained for each sample, constitute the sampling distribution of mean or variance depending upon the statistic calculated. In general we call sampling distribution of statistic.
Then the mean and variance of statistic t would be,

$$\text{Mean } \bar{t} = \frac{1}{n}\sum t_i$$

$$\text{Variance (t)} = \frac{1}{n}\sum \left(t_i - \bar{t}\right)^2 \ \forall\ i = 1, 2, \ldots, k$$

Standard error of mean
The standard deviation of the sampling distribution of mean is called standard error of mean. Thus standard error of the sampling distribution of t is given by,

Standard error of mean, SE ($\bar{X}$)= $\frac{\sigma}{\sqrt{n}}$

Standard error of difference of means $(\bar{x}_1 - \bar{x}_2)$ = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

The above formula for SE are obtained in random sampling from an infinite population so that the sample size n is relatively very small as compared with the population size N and consequently $\frac{n}{N}$ (sampling fraction) can be neglected. Thus for a finite population of size N when a sample is drawn without replacement method we have,

Standard error of mean, SE ($\bar{X}$)= $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Where $\sqrt{\frac{N-n}{N-1}}$ is called finite population correction (fpc) is applied when $\frac{n}{N} \geq 0.05$ (i.e. 5%)

Standard error of Proportion

Observed sample proportion (p)= $\sqrt{\frac{PQ}{n}}$

Difference of two proportions $(p_1 - p_2)$ = $\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Thus for a finite population of size N when a sample is drawn without replacement method we have,

Observed sample proportion (p)= $\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$

Example: A simple random sample of size 20 is drawn without replacement from a finite population of 75 units, if the number of defective units in the population is 12, find the standard error of the sample proportion.

Solution: Given n = 20, N = 75

P = Proportion of defective units in the population = 12/75 = 0.16

Q = 1 – P = 1 – 0.16 = 0.84

$$SE(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{0.16 \times 0.84}{20}}\sqrt{\frac{75-20}{75-1}} = 0.071$$

Exercise

1. The mean pulse rate of 50 people is 79.41 and a standard deviation is 0.7. Calculate standard error of the mean. **(0.099)**
2. A group of 16 subjects, a mean serum iron level was found as 148μg% with a standard deviation of 0.34μg%. Calculate the standard error of mean.

3. In a sample of 25 observations from a normal distribution with mean 98.6 and standard deviation 17.2, find the standard error of mean.
4. The Biostatistics class has a total of 60 students. Their average score in their second term exams was 70 with a standard deviation of 8. A sample of 36 these students is taken at random. Calculate the standard error of the mean for this sample. **(0.85)**
5. In a study on growth of children, one group of 100 children had a mean height of 60 cm and SD of 2.5 cm while another group of 150 children had a mean height of 62 cm and SD of 3 cm. Calculate standard error of difference of the mean. **(0.35)**
6. Case fatality rate of liver failure is 70 percent. There are 74 cases during a year. What is the standard error of sample 74? **(0.053)**
7. In a survey of 300 children in the age group 0 – 5 years, showed 15% prevalence rate of protein calorie malnutrition. Calculate the standard error of proportion of malnutrition. **(0.021)**
8. Occurrence of pyorrhea among 100 persons who did not brush their teeth was found to be 10, while in another sample of 100 persons who brushed their teeth, the occurrence was found to be 2. Calculate the standard error of proportion.
9. A sample of 20 persons drawn from a certain village showed the number of attacks of cold per person is: 3, 4, 0, 7, 5, 1, 6, 5, 8, 8, 2, 6, 0, 2, 4, 6, 1, 5, 2, 6. Calculate the standard error of mean.
10. Blood serum cholesterol level of 10 subjects is: 260, 277, 250, 240, 255, 245, 278, 288, 263, 290. Calculate the standard error of mean.

Relationship between sample size and standard error

- As the standard error decreases, the precision (accuracy) with which the sample mean can be used to estimate population mean increases.
- In other word, as the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean.
- As S.E. of a sample mean varies inversely with the square root of sample size *n* therefore, If we increase sample size then only S.E. of a sample mean decreases and the value of sample mean will be probably be closer to the value of the population mean.

Illustration

If population standard deviation σ = 100.

Case I

Consider sample of size *n* = 10

S.E. of sample mean $= \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{10}} = 31.63$

Case II

Consider sample of size *n* = 100

S.E. of sample mean $= \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10$

From this as we have increase *n* from 10 to 100 the valve of standard error decreases.