MODULE-3

SAMPLING TECHNIQUES

Sampling may be defined as the selection of some part of an aggregate or totality on the basis of which a judgement or inference about the aggregate or totality is made. In other words, it is the process of obtaining information about an entire population by examining only a part of it.

Sampling is used in practice for a variety of reasons such as:

1. Sampling can save time and money. A sample study is usually less expensive than a census study and produces results at a relatively faster speed.

2. Sampling may enable more accurate measurements for a sample study is generally conducted by trained and experienced investigators.

3. Sampling remains the only way when the population contains infinitely many members.

4. Sampling remains the only choice when a test involves the destruction of the item under study.

5. Sampling usually enables us to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristics of the population.

Universe/Population: From a statistical point of view, the term ‘Universe ‘refers to the total of the items or units in any field of inquiry, whereas the term ‘population’ refers to the total of items about which information is desired.

The attributes that are the object of study are referred to as **characteristics** and the units possessing them are called **elementary units**. The aggregate of such units is generally described as population. Thus, all units in any field of inquiry constitute the universe and all elementary units (based on one characteristic or more) constitute population.

The population or universe can be finite or infinite. The population is said to be **finite** if it consists of a fixed number of elements so that it is possible to enumerate it in its totality. An **infinite** population is that population in which it is theoretically impossible to observe all the elements. Thus, in an infinite population the number of items is infinite i.e., we cannot have any idea about the total number of items.

Sampling frame: The elementary units or the group or cluster of such units may form the basis of sampling process in which case they are called sampling units. A list containing all such sampling units is known as sampling frame. Thus, the sampling

frame consists of a list of items from which the sample is to be drawn. For instance, one can use telephone directory as a frame for conducting opinion survey in a city. Whatever the frame may be, it should be a good representative of the population.

Sampling design: A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data is collected.

***Statistic(s) and parameter(s):** A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population.

Thus, when we work out certain measures such as mean, median, mode or the like ones from samples, then they are called statistic(s) for they describe the characteristics of a sample. But when such measures describe the characteristics of a population, they are known as parameter(s). For instance, the population mean μ is a parameter, whereas the sample mean is a statistic.

***Precision:** Precision is the range within which the population average (or other parameter) will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate ± or as a numerical quantity. For instance, if the estimate is Rs 4000 and the precision desired is ± 4%, then the true value will be no less than Rs 3840 and no more than Rs 4160. This is the range (Rs 3840 to Rs 4160) within which the true answer should lie. But if we desire that the estimate should not deviate from the actual value by more than Rs 200 in either direction, in that case the range would be Rs 3800 to Rs 4200.

***Confidence level and significance level:** The confidence level or reliability is the expected percentage of times that the actual value will fall within the stated precision limits. Thus, if we take a confidence level of 95%, then we mean that there are 95 chances in 100 (or .95 in 1) that the sample results represent the true condition of the population within a specified precision range against 5 chances in 100 (or .05 in 1) that it does not. Precision is the range within which the answer may vary and still be acceptable; confidence level indicates the likelihood that the answer will fall within that range, and the significance level indicates the likelihood that the answer will fall outside that range. We can always remember that if the confidence level is 95%, then the significance level will be (100 – 95) i.e., 5%; if the confidence level is 99%, the significance level is (100 – 99) i.e., 1%, and so on.

Sampling error: Sample surveys do imply the study of a small portion of the population and as such there would naturally be a certain amount of inaccuracy in the information collected. This inaccuracy may be termed as sampling error or error variance. In other words, sampling errors are those errors which arise on account of sampling, and they generally happen to be random variations (in case of random sampling) in the sample estimates around the true population values. The magnitude of the sampling error depends upon the nature of the universe, the more homogeneous the universe, the

smaller the sampling error. Sampling error is inversely related to the size of the sample i.e., sampling error decreases as the sample size increases and vice-versa. A measure of the random sampling error can be calculated for a given sample design and size and this measure is often called the precision of the sampling plan.

Sampling error = Frame error + Chance error + Response error

***Standard Error:** The standard deviation of sampling distribution of a statistic is called S.E.

The utility of the concept of standard error in statistical induction arises on account of the following reasons:

1)The standard error helps in testing whether the difference between observed and expected frequencies could arise due to chance.

[FOR FURTHER READING ONLY :The criterion usually adopted is that if a difference is less than 3 times the S.E., the difference is supposed to exist as a matter of chance and if the difference is equal to or more than 3 times the S.E., chance fails to account for it, and we conclude the difference as significant difference. This criterion is based on the fact that at X ± 3 (S.E.) the normal curve covers an area of 99.73 per cent. Sometimes the criterion of 2 S.E. is also used in place of 3 S.E. Thus, the standard error is an important measure in significance tests or in examining hypotheses. If the estimated parameter differs from the calculated statistic by more than 1.96 times the S.E., the difference is taken as significant at 5 per cent level of significance. This, in other words, means that the difference is outside the limits i.e., it lies in the 5 per cent area (2.5 per cent on both sides) outside the 95 per cent area of the sampling distribution. Hence, we can say with 95 per cent confidence that the said difference is not due to fluctuations of sampling. In such a situation our hypothesis that there is no difference is rejected at 5 per cent level of significance. But if the difference is less than 1.96 times the S.E., then it is considered not significant at 5 per cent level and we can say with 95 per cent confidence that it is because of the fluctuations of sampling. In such a situation our null hypothesis stands true. 1.96 is the critical value at 5 per cent level. The product of the critical value at a certain level of significance and the S.E. is often described as ‘Sampling Error’ at that particular level of significance. We can test the difference at certain other levels of significance as well depending upon our requirement.]

2) The standard error gives an idea about the reliability and precision of a sample. The smaller the S.E., the greater the uniformity of sampling distribution and hence, greater is the reliability of sample. Conversely, the greater the S.E., the greater the difference between observed and expected frequencies. In such a situation the unreliability of the sample is greater. The size of S.E. depends upon the sample size to a great extent, and it varies inversely with the size of the sample. If double reliability is required i.e., reducing S.E. to 1/2 of its existing magnitude, the sample size should be increased four-fold.

3) The standard error enables us to specify the limits within which the parameters of the population are expected to lie with a specified degree of confidence. Such an interval is usually known as confidence interval.

SAMPLING THEORY

Sampling theory is a study of relationships existing between a population and samples drawn from the population. Sampling theory is applicable only to random samples. The theory of sampling studies the relationships that exist between the universe and the sample or samples drawn from it. The main problem of sampling theory is the problem of relationship between a parameter and a statistic. The theory of sampling is concerned with estimating the properties of the population from those of the sample and also with gauging the precision of the estimate. This sort of movement from particular (sample) towards general (universe) is what is known as statistical induction or statistical inference.

Sampling theory is designed to attain one or more of the following objectives:

(i) Statistical estimation: Sampling theory helps in estimating unknown population parameters from a knowledge of statistical measures based on sample studies. In other words, to obtain an estimate of parameter from statistic is the main objective of the sampling theory. The estimate can either be a point estimate or it may be an interval estimate. Point estimate is a single estimate expressed in the form of a single figure, but interval estimate has two limits viz., the upper limit and the lower limit within which the parameter value may lie. Interval estimates are often used in statistical induction.

(ii) Testing of hypotheses: The second objective of sampling theory is to enable us to decide whether to accept or reject hypothesis; the sampling theory helps in determining whether observed differences are due to chance or whether they are really significant.

(iii) Statistical inference: Sampling theory helps in making generalization about the population/ universe from the studies based on samples drawn from it. It also helps in determining the accuracy of such generalizations.

The theory of sampling can be studied under two heads viz., the sampling of attributes and the sampling of variables and that too in the context of large and small samples (By small sample is commonly understood any sample that includes 30 or fewer items, whereas a large sample is one in which the number of items is more than 30). When we study some qualitative characteristic of the items in a population, we obtain statistics of attributes in the form of two classes; one class consisting of items wherein the attribute is present and the other class consisting of items wherein the attribute is absent. The presence of an attribute may be termed as a ‘success’ and its absence a ‘failure’.

We generally consider the following three types of problems in case of sampling of attributes:

(i) The parameter value may be given, and it is only to be tested if an observed 'statistic' is its estimate.

(ii) The parameter value is not known, and we have to estimate it from the sample.

iii) Examination of the reliability of the estimate i.e., the problem of finding out how far the estimate is expected to deviate from the true value for the population.

The theory of sampling can be applied in the context of statistics of variables (i.e., data relating to some characteristic concerning population which can be measured or enumerated with the help of some well-defined statistical unit) in which case the objective happens to be :

(i) To compare the observed and expected values and to find if the difference can be ascribed to the fluctuations of sampling.
(ii) To estimate population parameters from the sample, and
(iii) To find out the degree of reliability of the estimate.

Sampling distribution: If we take a certain number of samples and for each sample compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, say mean, together with their relative frequencies will constitute the sampling distribution of the statistic, say mean. Accordingly, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. The sampling distribution tends quite closer to the normal distribution if the number of samples is large. The significance of sampling distribution follows from the fact that the mean of a sampling distribution is the same as the mean of the universe. Thus, the mean of the sampling distribution can be taken as the mean of the universe.

IMPORTANT SAMPLING DISTRIBUTIONS

Some important sampling distributions, which are commonly used, are:

(1) sampling distribution of mean

(2) sampling distribution of proportion

(3) student's 't' distribution

(4) F distribution

(5) Chi-square distribution.

Sampling distributions: Any statistic, being a random variable, has a probability distribution. The probability distribution of a statistic is called Sampling distribution.

Types of Sampling distribution

1. Sampling distribution of Mean
2. Sampling distribution of proportion
3. t - distribution
4. χ^2 - distribution (Chi-square)
5. F - distribution.

Standard error: The standard deviation of the sampling distribution of a statistic is called its standard error.

1. Sampling distribution of mean ($\bar{X}$)

It is the probability distribution of all the possible means of random samples of a given size that we take from the population.

Let $X_1, X_2 \ldots X_n$ represent n independent observations from a population with mean μ and S.D σ.

let $\bar{X}$ denote the mean of these observations

ie, $\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n}$

$\bar{X}$ is a random variable known as sampling distribution of sample mean.

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots X_n}{n}\right)$$

$$= \frac{1}{n}\{E(X_1) + \cdots E(X_n)\}$$

$$= \frac{1}{n}\{\underbrace{\mu + \cdots \mu}_{n \text{ times}}\}$$

$$= \mu$$

$$V(\bar{X}) = V\left(\frac{X_1 + \cdots X_n}{n}\right)$$

$$= \frac{1}{n^2}\{V(X_1) + \cdots V(X_n)\}$$

$$= \frac{1}{n^2}\{\underbrace{\sigma^2 + \cdots \sigma^2}_{n \text{ times}}\}$$

$$= \frac{\sigma^2}{n}$$

$$\therefore S.D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Note: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ is called standard error of the sample mean.

- If $X_1, X_2 \cdots X_n$ be a random sample from a normal population with mean μ and S.D σ,
  ie, $X_i \sim N(\mu, \sigma^2)$ then by CLT

  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

To reduce the sampling distribution of mean to unit normal distribution ie, $N(0,1)$ we can write

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

- If the population is not normal by CLT, $\bar{X}$ is normally distributed provided the no. of sample items are large $\{n \geq 30\}$

S.E

- Standard error of mean when population S.D is known

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

σ_p – S.D of population
n – sample size

- S.E of mean when population S.D is unknown

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

σ_s – S.D of sample

$$\sigma_s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

- S.E of difference between means of two samples

(a) When two samples are drawn from the same population

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

{ If σ_p is not known, Sample S.D for Combined samples denoted by $\sigma_{s_{1.2}}$ may be substituted }

(b) When two samples are drawn from different population

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{\sigma_{p_1}^2}{n_1} + \frac{\sigma_{p_2}^2}{n_2}}$$

{ If σ_{p_1}, σ_{p_2} are not known, then σ_{s_1} and σ_{s_2} respectively may be substituted. }

- S.E of mean when population is finite

$$\sigma_{\overline{x}} = \frac{\sigma_p}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$\sqrt{\frac{N-n}{N-1}}$ is called Correction factor.

(2) Sampling distribution of proportion ($\hat{p}$)

The distribution of the values of the sample proportion ($\hat{p}$) in repeated samples of same size is called the sampling distribution of $\hat{p}$.

but X = No. of individuals in the sample with the same characteristic

n = No. of individuals in the sample

$\therefore \hat{p} = X/n$.

- If X is a binomial random variable i.e, $X \sim B(n, p)$ then $\hat{p}$ takes one of the possible values $0, 1/n, 2/n \cdots n/n$

Mean of the sampling distribution of sample proportion

$$E(\hat{p}) = E(X/n) = 1/n\, E(X) = 1/n \cdot np = p$$

$(\mu_{\hat{p}})$

$$V(\hat{p}) = \sigma^2_{\hat{p}} = V(X/n) = 1/n^2 [npq(1-p)]$$

$(\sigma^2_{\hat{p}})$

$$= \frac{pq(p+1)}{n}$$

$$S.D(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$(\sigma_{\hat{p}})$

- If the sample size is large by CLT binomial distribution can be approximating by Normal distribution. Normal variable

$$Z = \frac{\hat{p} - p}{\sqrt{pq(1-p)/n}}$$

S.E

- S.E of proportion ($\hat{P}$)

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \text{ (with replacement) or } \sqrt{\frac{pq}{n}}\sqrt{\frac{N-n}{N-1}} \text{ (without replacement)}$$

(when population size is given)

n = samplesize
p = population proportion
$q = 1-p$
N = population size

- S.E of difference blw proportions of two samples

a) $$\sigma_{p_1-p_2} = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}$$

where $$p = \frac{n_1p_1 + n_2p_2}{n_1+n_2}$$ { It is the best estimate of proportion in the population.

$$q = 1-p$$

n_1 = No. of events in sample 1

n_2 = " " 2

b) $$\sigma_{p_1-p_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

when samples are drawn from two heterogeneous population.

Distribution of the sample proportion is

- Normal if $np \geq 5$ and $n(1-p) \geq 5$
- Binomial if $np < 5$ and $n(1-p) < 5$

(3) Students t- distribution

If we draw a random sample of n observations from a normally distributed population with mean μ and variance σ^2 then $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ (1) follows Std normal distribution. In some cases σ is not known (population S.D is denoted by σ) replace σ in (1) by sample standard deviation σ_s. So (1) does not follow normal distribution. This new type of distribution is called t distribution

ie, $t = \frac{\bar{X} - \mu}{\frac{\sigma_s}{\sqrt{n}}}$ where $\sigma_s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$

has $(n-1)$ degrees of freedom.

{ Degrees of freedom of a statistic is a +ve integer denoted by ν is the no. of values in the final calculation of a statistic that are free to vary.

In general, $\nu = n - k$

n = sample size

k = no. of independent Constraints imposed on the Observations in the sample.

Note: When population S.D σ is not known and the sample is of a small size (i.e, $n \leq 30$), we use t-distribution for the sampling distribution of mean.

(4) F-distribution

This test is usually used to test equality of variances. If $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$ are the variances of two independent samples of size n_1 and n_2 respectively taken from two independent normal populations having the same variance, $\sigma_{p_1}^2 = \sigma_{p_2}^2$ the ratio

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

where

$$\sigma_{s_1}^2 = \frac{\sum (X_{1i} - \bar{X}_1)^2}{n_1 - 1}$$

$$\sigma_{s_2}^2 = \frac{\sum (X_{2i} - \bar{X}_2)^2}{n_2 - 1}$$

has an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. F ratio is computed in a way that the larger variance is always in the numerator.

(5) Chi-square distribution (χ^2)

χ^2-distribution with degrees of freedom, k is the distribution of a sum of the squares of k independent standard normal random variables

Chi-square distribution is used to compute the following.

1) To check the relationships between categorical variables.
2) To check independence of two criteria of classification of multiple qualitative variables
3) To conduct a The Chi-square test.

The formula for the chisquare statistic used in a chi-square test is

$$\chi^2_\nu = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$\nu \rightarrow$ Degrees of freedom
$O_i \rightarrow$ Observed value
$E_i \rightarrow$ Expected value.

Tables are there that give the value of χ^2 for given d.f which may be used with calculated value of χ^2 for relevent d.f at a desired level of significance for testing hypotheses.

Definitions

Precision: It is the range within which the population parameter will lie in accordance with the reliability specified in the confidence level as a percentage of the estimate or as a numerical quantity.

Eg: If the estimate is 10000 and precision desired is ±4% then true value is in blw 9600 and 10400 & $10000 - \frac{4}{100} \times 10000$, $10000 + \frac{4}{100} \times 10000$}

Confidence level: It is the expected % of times that the actual value will fall within the stated precision limits.

Eg: If we take a confidence level of 95% then it means that there are 95 chances in 100 that the sample results represent the true condition of the population with in a specific precision range against 5 chances in 100 that it does not.

Significance level: It is the probability that the answer will fall outside precision.

If Confidence level is 95%. then significance level is 5%.

Confidence interval: A probability that a parameter will fall between a set of values.

or If t is the statistic used to estimate parameter θ, then $1-\alpha$ Confidence limits for θ is

$$t \pm S.E(t)\, t_{\alpha/2}$$

where $t_{\alpha/2}$ is the significant or critical value of t at level of significance α for a two tailed test.

ie, $P(t - S.E(t)\, t_{\alpha/2}, t + S.E(t)\, t_{\alpha/2}) = 1-\alpha$

$1-\alpha$ is called Confidence Coefficient and α is called level of significance.

Result: If the sample size is large then the underlaying distribution of the Standardised variate corresponding to the Sampling distribution of the statistic t follows normal distribution

ie, $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

So $1-\alpha$ Confidence limits for θ is

$$t \pm Z_{\alpha/2} \cdot S.E(t)$$

- To estimate population mean μ
Sample mean $\bar{x}$ is the best estimate of population mean (μ)
$1-\alpha$ Confidence limits for μ is

$\bar{x} \pm z_{\alpha/2} \frac{\sigma_s}{\sqrt{n}}$, population size is not given (N)

$\bar{x} \pm z_{\alpha/2} \frac{\sigma_s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, population size($N$) is given

$\bar{x} \pm t_{\alpha/2} \frac{\sigma_s}{\sqrt{n}}$, size of sample is small

- To estimate population proportion
Sample proportion is the best estimator for population proportion

$1-\alpha$ Confidence limits for population proportion is

$$p \pm z_{\alpha/2} \sqrt{\frac{pq}{n}}$$

A sample of 36 New