8

Sampling Fundamentals

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. In most of the research work and surveys, the usual approach happens to be to make generalisations or to draw inferences based on samples about the parameters of population from which the samples are taken. The researcher quite often selects only a few items from the universe for his study purposes. All this is done on the assumption that the sample data will enable him to estimate the population parameters. The items so selected constitute what is technically called a sample, their selection process or technique is called sample design and the survey conducted on the basis of sample is described as sample survey. Sample should be truly representative of population characteristics without any bias so that it may result in valid and reliable conclusions.

NEED FOR SAMPLING

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 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 to estimate the sampling errors and, thus, assists in obtaining information concerning some characteristic of the population.

SOME FUNDAMENTAL DEFINITIONS

Before we talk about details and uses of sampling, it seems appropriate that we should be familiar with some fundamental definitions concerning sampling concepts and principles.

1. *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 as elementary units. The aggregate of such units is generally described as population. Thus, all units in any field of inquiry constitute universe and all elementary units (on the basis of one characteristic or more) constitute population. Quit often, we do not find any difference between population and universe, and as such the two terms are taken as interchangeable. However, a researcher must necessarily define these terms precisely.

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. For instance, the population of a city, the number of workers in a factory are examples of finite populations. The symbol '*N*' is generally used to indicate how many elements (or items) are there in case of a finite population. 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. The number of stars in a sky, possible rolls of a pair of dice are examples of infinite population. One should remember that no truly infinite population of physical objects does actually exist in spite of the fact that many such populations appear to be very very large. From a practical consideration, we then use the term infinite population for a population that cannot be enumerated in a reasonable period of time. This way we use the theoretical concept of infinite population as an approximation of a very large finite population.

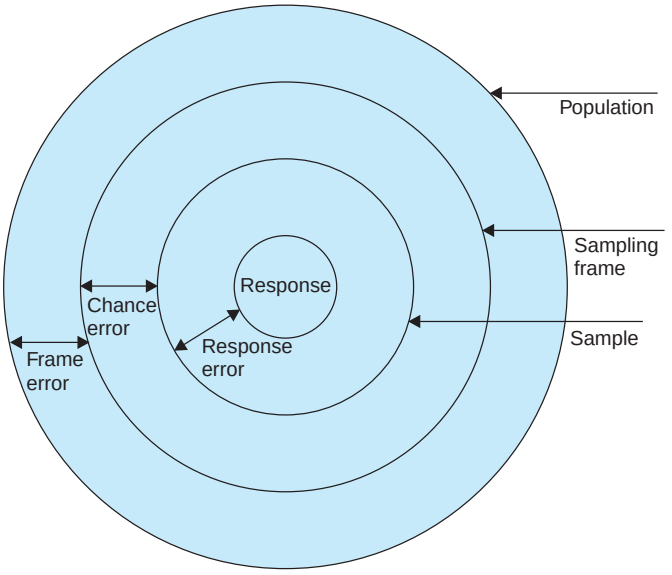
2. *Sampling frame:* The elementary units or the group or cluster of such units may form the basis of sampling process in which case they are called as sampling units. A list containing all such sampling units is known as sampling frame. Thus sampling frame consists of a list of items from which the sample is to be drawn. If the population is finite and the time frame is in the present or past, then it is possibe for the frame to be identical with the population. In most cases they are not identical because it is often impossible to draw a sample directly from population. As such this frame is either constructed by a researcher for the purpose of his study or may consist of some existing list of the population. 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.

3. *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 are collected. Various sampling designs have already been explained earlier in the book.

4. *Statisitc(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 ($\overline{X}$) is a statistic. To obtain the estimate of a parameter from a statistic constitutes the prime objective of sampling analysis.

5. *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 meaning of sampling error can be easily understood from the following diagram:



Sampling error = Frame error
+ chance error + response error.
(If we add measurement error or the non-sampling error to sampling error, we get total error)

Fig. 8.1

Sampling error = Frame error + Chance error + Response error

(If we add measurement error or the non-sampling error to sampling error, we get total error).

Sampling errors occur randomly and are equally likely to be in either direction. 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 is usually worked out as the product of the critical value at a certain level of significance and the standard error.

As opposed to sampling errors, we may have non-sampling errors which may creep in during the process of collecting actual information and such errors occur in all surveys whether census or sample. We have no way to measure non-sampling errors.

6. *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 $\pm$ or as a numerical quantity. For instance, if the estimate is Rs 4000 and the precision desired is $\pm$ 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.

7. *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. We should also remember that the area of normal curve within precision limits for the specified confidence level constitute the acceptance region and the area of the curve outside these limits in either direction constitutes the rejection regions.*

8. *Sampling distribution:* We are often concerned with sampling distribution in sampling analysis. If we take 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 particular 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; and (5) Chi-square distribution. A brief mention of each one of these sampling distribution will be helpful.

1. *Sampling distribution of mean:* Sampling distribution of mean refers to the probability distribution of all the possible means of random samples of a given size that we take from a population. If samples are taken from a normal population, $N\left(\mu, \sigma_p\right)$, the sampling distribution of mean would also be normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation $= \sigma_p \sqrt{n}$, where μ is the mean of the population, σ_p is the standard deviation of the population and *n* means the number of items in a sample. But when sampling is from a population which is not normal (may be positively or negatively skewed), even then, as per the central limit theorem, the sampling distribution of mean tends quite closer to the normal distribution, provided the number of sample items is large i.e., more than 30. In case we want to reduce the sampling distribution of mean to unit normal distribution i.e., *N* (0,1), we can write the

*See Chapter 9 Testing of Hypotheses I for details.

normal variate $z = \frac{\bar{x} - \mu}{\sigma_p / \sqrt{n}}$ for the sampling distribution of mean. This characteristic of the sampling distribution of mean is very useful in several decision situations for accepting or rejection of hypotheses.

2. *Sampling distribution of proportion:* Like sampling distribution of mean, we can as well have a sampling distribution of proportion. This happens in case of statistics of attributes. Assume that we have worked out the proportion of defective parts in large number of samples, each with say 100 items, that have been taken from an infinite population and plot a probability distribution of the said proportions, we obtain what is known as the sampling distribution of the said proportions, we obtain what is known as the sampling distribution of proportion. Usually the statistics of attributes correspond to the conditions of a binomial distribution that tends to become normal distribution as *n* becomes larger and larger. If *p* represents the proportion of defectives i.e., of successes and *q* the proportion of non-defectives i.e., of failures (or $q = 1 - p$) and if *p* is treated as a random variable, then the sampling distribution of proportion of successes has a mean = *p* with standard deviation $= \sqrt{\frac{p \cdot q}{n}}$, where *n* is the sample size. Presuming the binomial distribution approximating the normal distribution for large *n*, the normal variate of the sampling distribution of proportion $z = \frac{\hat{p} - p}{\sqrt{(p \cdot q)/n}}$, where $\hat{p}$ (pronounced as *p*-hat) is the sample proportion of successes, can be used for testing of hypotheses.

3. *Student's t-distribution:* When population standard deviation (σ_p) is not known and the sample is of a small size $(\text{i.e., } n \leq 30)$, we use *t* distribution for the sampling distribution of mean and workout *t* variable as:

$$t = \left(\bar{X} - \mu\right) \Big/ \left(\sigma_s / \sqrt{n}\right)$$

where $\sigma_s = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{n} - 1}$

i.e., the sample standard deviation . *t*-distribution is also symmetrical and is very close to the distribution of standard normal variate, *z*, except for small values of *n*. The variable *t* differs from *z* in the sense that we use sample standard deviation (σ_s) in the calculation of *t*, whereas we use standard deviation of population (σ_p) in the calculation of *z*. There is a different *t* distribution for every possible sample size i.e., for different degrees of freedom. The degrees of freedom for a sample of size *n* is $n - 1$. As the sample size gets larger, the shape of the *t* distribution becomes apporximately equal to the normal distribution. In fact for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we can use the normal to approximate the *t*-distribution. But when *n* is small, the *t*-distribution is far from normal but when $n \to \alpha$, *t*-distribution is identical with normal distribution. The *t*-distribution tables are available which give the critical values of *t* for different degrees of freedom at various levels of significance. The table value of *t* for given degrees of freedom at a

certain level of significance is compared with the calculated value of *t* from the sample data, and if the latter is either equal to or exceeds, we infer that the null hypothesis cannot be accepted.*

4. *F distribution:* If $(\sigma_{s1})^2$ and $(\sigma_{s2})^2$ 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, $(\sigma_{p1})^2 = (\sigma_{p2})^2$, the ratio $F = (\sigma_{s1})^2 / (\sigma_{s2})^2$, where $(\sigma_{s1})^2 = \Sigma \left(\bar{X}_{1i} - \bar{X}_1\right)^2 / n_1 - 1$ and $(\sigma_{s2})^2 = \Sigma \left(\bar{X}_{2i} - \bar{X}_2\right)^2 / n_2 - 1$ 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. Tables have been prepared for *F* distribution that give critical values of *F* for various values of degrees of freedom for larger as well as smaller variances. The calculated value of *F* from the sample data is compared with the corresponding table value of *F* and if the former is equal to or exceeds the latter, then we infer that the null hypothesis of the variances being equal cannot be accepted. We shall make use of the *F* ratio in the context of hypothesis testing and also in the context of ANOVA technique.

5. *Chi-square* $\left(\chi^2\right)$ *distribution:* Chi-square distribution is encountered when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and thus have distributions that are related to chi-square distribution. If we take each one of a collection of sample variances, divide them by the known population variance and multiply these quotients by $(n-1)$, where *n* means the number of items in the sample, we shall obtain a chi-square distribution. Thus, $\left(\sigma_s^2 / \sigma_p^2\right)(n-1)$ would have the same distribution as chi-square distribution with $(n-1)$ degrees of freedom. Chi-square distribution is not symmetrical and all the values are positive. One must know the degrees of freedom for using chi-square distribution. This distribution may also be used for judging the significance of difference between observed and expected frequencies and also as a test of goodness of fit. The generalised shape of χ^2 distribution depends upon the d.f. and the χ^2 value is worked out as under:

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

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 relevant d.f. at a desired level of significance for testing hypotheses. We will take it up in detail in the chapter 'Chi-square Test'.

CENTRAL LIMIT THEOREM

When sampling is from a normal population, the means of samples drawn from such a population are themselves normally distributed. But when sampling is not from a normal population, the size of the

* This aspect has been dealt with in details in the context of testing of hypotheses later in this book.

sample plays a critical role. When *n* is small, the shape of the distribution will depend largely on the shape of the parent population, but as *n* gets large ($n > 30$), the thape of the sampling distribution will become more and more like a normal distribution, irrespective of the shape of the parent population. The theorem which explains this sort of relationship between the shape of the population distribution and the sampling distribution of the mean is known as the central limit theorem. This theorem is by far the most important theorem in statistical inference. It assures that the sampling distribution of the mean approaches normal distribtion as the sample size increases. In formal terms, we may say that the central limit theorem states that "the distribution of means of random samples taken from a population having mean μ and finite variance σ^2 approaches the normal distribution with mean μ and variance σ^2/n as *n* goes to infinity."[1]

"The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample."[2]

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. For this purpose the population or a universe may be defined as an aggregate of items possessing a common trait or traits. In other words, a universe is the complete group of items about which knowledge is sought. The universe may be finite or infinite. finite universe is one which has a definite and certain number of items, but when the number of items is uncertain and infinite, the universe is said to be an infinite universe. Similarly, the universe may be hypothetical or existent. In the former case the universe in fact does not exist and we can only imagin the items constituting it. Tossing of a coin or throwing a dice are examples of hypothetical universe. Existent universe is a universe of concrete objects i.e., the universe where the items constituting it really exist. On the other hand, the term sample refers to that part of the universe which is selected for the purpose of investigation. 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. In more clear terms "from the sample we attempt to draw inference concerning the universe. In order to be able to follow this inductive method, we first follow a deductive argument which is that we imagine a population or universe (finite or infinite) and investigate the behaviour of the samples drawn from this universe applying the laws of probability."[3] The methodology dealing with all this is known as sampling theory.

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

[1] Donald L. Harnett and James L. Murphy, *Introductory Statistical Analysis,* p.223.

[2] Richard I. Levin, *Statistics for Management,* p. 199.

[3] J.C. Chaturvedi: *Mathematical Statistics,* p. 136.

(*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 actually due to chance or whether they are really significant.

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

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 alarge 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'. Thus, if out of 600 people selected randomly for the sample, 120 are found to possess a certain attribute and 480 are such people where the attribute is absent. In such a situation we would say that sample consists of 600 items (i.e., $n = 600$) out of which 120 are successes and 480 failures. The probability of success would be taken as 120/600 = 0.2 (i.e., $p = 0.2$) and the probability of failure or $q = 480/600 = 0.8$. With such data the sampling distribution generally takes the form of binomial probability distribution whose mean (μ) would be equal to $n \cdot p$ and standard deviation (σ_p) would be equal to $\sqrt{n \cdot p \cdot q}$. If n is large, the binomial distribution tends to become normal distribution which may be used for sampling analysis. 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.

All the above stated problems are studied using the appropriate standard errors and the tests of significance which have been explained and illustrated in the pages that follow.

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.

The tests of significance used for dealing with problems relating to large samples are different from those used for small samples. This is so because the assumptions we make in case of large samples do not hold good for small samples. In case of large samples, we assume that the sampling distribution tends to be normal and the sample values are approximately close to the population values. As such we use the characteristics of normal distribution and apply what is known as *z*-test*. When *n* is large, the probability of a sample value of the statistic deviating from the parameter by more than 3 times its standard error is very small (it is 0.0027 as per the table giving area under normal curve) and as such the *z*-test is applied to find out the degree of reliability of a statistic in case of large samples. Appropriate standard errors have to be worked out which will enable us to give the limits within which the parameter values would lie or would enable us to judge whether the difference happens to be significant or not at certain confidence levels. For instance, $\overline{X} \pm 3\sigma_{\overline{X}}$ would give us the range within which the parameter mean value is expected to vary with 99.73% confidence. Important standard errors generally used in case of large samples have been stated and applied in the context of real life problems in the pages that follow.

The sampling theory for large samples is not applicable in small samples because when samples are small, we cannot assume that the sampling distribution is approximately normal. As such we require a new technique for handlng small samples, particularly when population parameters are unknown. Sir William S. Gosset (pen name Student) developed a significance test, known as Student's *t*-test, based on *t* distribution and through it made significant contribution in the theory of sampling applicable in case of small samples. Student's *t*-test is used when two conditions are fulfilled viz., the sample size is 30 or less and the population variance is not known. While using *t*-test we assume that the population from which sample has been taken is normal or approximately normal, sample is a random sample, observations are independent, there is no measurement error and that in the case of two samples when equality of the two population means is to be tested, we assume that the population variances are equal. For applying *t*-test, we work out the value of test statistic (i.e., '*t*') and then compare with the table value of *t* (based on '*t*' distribution) at certain level of significance for given degrees of freedom. If the calculated value of '*t*' is either equal to or exceeds the table value, we infer that the difference is significant, but if calculated value of *t* is less than the concerning table value of *t*, the difference is not treated as significant. The following formulae are commonly used to calculate the *t* value:

(i) To test the significance of the mean of a random sample

$$t = \frac{\left(\overline{X} - \mu\right)}{\sigma_{\overline{X}}}$$

where $\overline{X}$ = Mean of the sample

μ = Mean of the universe/population

$\sigma_{\overline{X}}$ = Standard error of mean worked out as under

$$\sigma_{\overline{X}} = \frac{\sigma_s}{\sqrt{n}} = \sqrt{\frac{\sum \left(Xi - \overline{X}\right)^2}{n-1}} \Big/ \sqrt{n}$$

and the degrees of freedom = $(n - 1)$.

*The *z*-test may as well be applied in case of small sample provided we are given the variance of the population.

(ii) To test the difference between the means of two samples

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

where $\bar{X}_1$ = Mean of sample one

$\bar{X}_2$ = Mean of sample two

$\sigma_{\bar{X}_1 - \bar{X}_2}$ = Standard error of difference between two sample means worked out as

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the d.f. = $(n_1 + n_2 - 2)$.

(iii) To test the significance of the coefficient of simple correlation

$$t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{n - 2} \text{ or } t = r\sqrt{\frac{n - 2}{1 - r^2}}$$

where

r = the coefficient of simple correlation

and the d.f. = $(n - 2)$.

(iv) To test the significance of the coefficient of partial correlation

$$t = \frac{r_p}{\sqrt{1 - r_p^2}} \times \sqrt{n - k} \quad \text{or} \quad t = r_p\sqrt{\frac{(n - k)}{1 - r_p^2}}$$

where r_p is any partial coeficient of correlation

and the d.f. = $(n - k)$, n being the number of pairs of observations and k being the number of variables involved.

(v) To test the difference in case of paired or correlated samples data (in which case t test is ofter described as difference test)

$$t = \frac{\bar{D} - \mu_D}{\sigma_D}\sqrt{n} \quad \text{i.e.,} \quad t = \frac{\bar{D} - 0}{\sigma_D}\sqrt{n}$$

where

Hypothesised mean difference (μ_D) is taken as zero (0),

$\bar{D}$ = Mean of the differences of correlated sample items

σ_D = Standard deviation of differences worked out as under

$$\sigma_D = \sqrt{\frac{\Sigma\, D_i^2 - \bar{D}/n}{n - 1}}$$

D_i = Differences {i.e., $D_i = (X_i - Y_i)$}

n = number of pairs in two samples and the d.f. = $(n - 1)$.

SANDLER'S *A*-TEST

Joseph Sandler has developed an alternate approach based on a simplification of *t*-test. His approach is described as Sandler's *A*-test that serves the same purpose as is accomplished by *t*-test relating to paired data. Researchers can as well use *A*-test when correlated samples are employed and hypothesised mean difference is taken as zero i.e., $H_0 : \mu_D = 0$. Psychologists generally use this test in case of two groups that are matched with respect to some extraneous variable(s). While using *A*-test, we work out *A*-statistic that yields exactly the same results as Student's *t*-test*. *A*-statistic is found as follows:

$$A = \frac{\text{the sum of squares of the differences}}{\text{the squares of the sum of the differences}} = \frac{\Sigma D_i^2}{(\Sigma D_i)^2}$$

The number of degrees of freedom (d.f.) in *A*-test is the same as with Student's *t*-test i.e., d.f. = $n - 1$, n being equal to the number of pairs. The critical value of *A*, at a given level of significance for given d.f., can be obtained from the table of *A*-statistic (given in appendix at the end of the book). One has to compare the computed value of *A* with its corresponding table value for drawing inference concerning acceptance or rejection of null hypothesis.** If the calculated value of *A* is equal to or less than the table value, in that case *A*-statistic is considered significant where upon we reject H_0 and accept H_a. But if the calculated value of *A* is more than its table value, then *A*-statistic is taken as insignificant and accordingly we accept H_0. This is so because the two test statistics viz., *t* and *A* are inversely related. We can write these two statistics in terms of one another in this way:

(i) '*A*' in terms of '*t*' can be expressed as

$$A = \frac{n-1}{n \cdot t^2} + \frac{1}{n}$$

(ii) '*t*' in terms of '*A*' can be expressed as

$$t = \sqrt{\frac{n-1}{A \cdot n - 1}}$$

Computational work concerning *A*-statistic is relatively simple. As such the use of *A*-statistic result in considerable saving of time and labour, specially when matched groups are to be compared with respect to a large number of variables. Accordingly researchers may replace Student's *t*-test by Sandler's *A*-test whenever correlated sets of scores are employed.

Sandler's *A*-statistic can as well be used "in the one sample case as a direct substitute for the Student *t*-ratio."[4] This is so because Sandler's *A* is an algebraically equivalent to the Student's *t*. When we use *A*-test in one sample case, the following steps are involved:

(*i*) Subtract the hypothesised mean of the population (μ_H) from each individual score (X_i) to obtain D_i and then work out ΣD_i.

* For proof, see the article, "*A* test of the significance of the difference between the means of correlated measures based on a simplification of Student's" by Joseph Sandler, published in the *Brit. J Psych.*, 1955, pp. 225–226.

** See illustrations 11 and 12 of Chapter 9 of this book for the purpose.

[4] Richard P. Runyon, *Inferential Statistics: A Contemporary Approach*, p.28

(*ii*) Square each D_i and then obtain the sum of such squares i.e., ΣD_i^2.

(*iii*) Find *A*-statistic as under:

$$A = \Sigma D_i^2 \Big/ (\Sigma D_i)^2$$

(*iv*) Read the table of *A*-statistic for $(n-1)$ degrees of freedom at a given level of significance (using one-tailed or two-tailed values depending upon H_a) to find the critical value of *A*.

(*v*) Finally, draw the inference as under:

When calculated value of *A* is equal to or less than the table value, then reject H_0 (or accept H_a) but when computed *A* is greater than its table value, then accept H_0.

The practical application/use of *A*-statistic in one sample case can be seen from Illustration No. 5 of Chapter IX of this book itself.

CONCEPT OF STANDARD ERROR

The standard deviation of sampling distribution of a statistic is known as its standard error (S.E) and is considered the key to sampling theory. 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. 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 $\overline{X} \pm 3$ (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. The following table gives some idea about the criteria at various levels for judging the significance of the difference between observed and expected values:

Table 8.1: Criteria for Judging Significance at Various Important Levels

Significance level	*Confidence level*	*Critical value*	*Sampling error*	*Confidence limits*	*Difference Significant if*	*Difference Insignificant if*
5.0%	95.0%	1.96	1.96σ	$\pm 1.96\sigma$	$\geq 1.96\sigma$	$\leq 1.96\sigma$
1.0%	99.0%	2.5758	$2.5758\ \sigma$	$\pm\ 2.5758\ \sigma$	$\geq 2.5758\ \sigma$	$\leq 2.5758\ \sigma$
2.7%	99.73%	3	$3\ \sigma$	$\pm\ 3\ \sigma$	$\geq 3\ \sigma$	$\leq 3\ \sigma$
4.55%	95.45%	2	$2\ \sigma$	$\pm\ 2\ \sigma$	$\geq 2\ \sigma$	$\leq 2\ \sigma$

σ = Standard Error.

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. The following table gives the percentage of samples having their mean values within a range of population mean $(\mu) \pm$ S.E.

Table 8.2

Range	*Per cent Values*
$\mu \pm$ 1 S.E.	68.27%
$\mu \pm$ 2 S.E.	95.45%
$\mu \pm$ 3 S.E.	99.73%
$\mu \pm$ 1.96 S.E.	95.00%
$\mu \pm$ 2.5758 S.E.	99.00%

Important formulae for computing the standard errors concerning various measures based on samples are as under:

(a) *In case of sampling of attributes*:

(i) Standard error of number of successes = $\sqrt{n \cdot p \cdot q}$

where n = number of events in each sample,
p = probability of success in each event,
q = probability of failure in each event.

(ii) Standard error of proportion of successes $\sqrt{\frac{(p \cdot q)}{n}}$

(iii) Standard error of the difference between proportions of two samples:

$$\sigma_{p_1 - p_2} = \sqrt{p \cdot q \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where p = best estimate of proportion in the population and is worked out as under:

$$p = \frac{n_1 \ p_1 + n_2 \ p_2}{n_1 + n_2}$$

$q = 1 - p$

n_1 = number of events in sample one

n_2 = number of events in sample two

Note: Instead of the above formula, we use the following formula:

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 \ q_1}{n_1} + \frac{p_2 \ q_2}{n_2}}$$

when samples are drawn from two heterogeneous populations where we cannot have the best estimate of proportion in the universe on the basis of given sample data. Such a situation often arises in study of association of attributes.

(*b*) *In case of sampling of variables* (*large samples*)*:*

(i) Standard error of mean when population standard deviation is known:

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

where

σ_p = standard deviation of population

n = number of items in the sample

Note: This formula is used even when n is 30 or less.

(ii) Standard error of mean when population standard deviation is unknown:

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

where

σ_s = standard deviation of the sample and is worked out as under

$$\sigma_s = \sqrt{\frac{\Sigma \left(X_i - \bar{X} \right)^2}{n - 1}}$$

n = number of items in the sample.

(iii) Standard error of standard deviation when population standard deviation is known:

$$\sigma_{\sigma_s} = \frac{\sigma_p}{\sqrt{2\,n}}$$

(iv) Standard error of standard deviation when population standard deviation is unknown:

$$\sigma_{\sigma_s} = \frac{\sigma_s}{\sqrt{2\,n}}$$

where $$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}$$

n = number of items in the sample.

(v) Standard error of the coeficient of simple correlation:

$$\sigma_r = \frac{1 - r^2}{\sqrt{n}}$$

where

r = coefficient of simple correlation

n = number of items in the sample.

(vi) Standard error of difference between means of two samples:

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

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

(If σ_p is not known, sample standard deviation for combined samples $\left(\sigma_{s_{1,2}}\right)^*$ may be substituted.)

(b) When two samples are drawn from different populations:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\left(\sigma_{p_1}\right)^2}{n_1} + \frac{\left(\sigma_{p_2}\right)^2}{n_2}}$$

(If σ_{p_1} and σ_{p_2} are not known, then in their places σ_{s_1} and σ_{s_2} respectively may be substituted.)

(c) *In case of sampling of variables* (*small samples*):

(i) Standard error of mean when σ_p is unknown:

$$\sigma_{s_{1\cdot 2}} = \sqrt{\frac{n_1\left(\sigma_{s_1}\right)^2 + n_2\left(\sigma_{s_2}\right)^2 + n_1\left(\bar{X}_1 - \bar{X}_{1\cdot 2}\right)^2 + n_2\left(\bar{X}_2 - \bar{X}_{1\cdot 2}\right)^2}{n_1 + n_2}}$$

where $\bar{X}_{1\cdot 2} = \dfrac{n_1\left(\bar{X}_1\right) + n_2\left(\bar{X}_2\right)}{n_1 + n_2}$

Note: (1) All these formulae apply in case of infinite population. But in case of finite population where sampling is done without replacement and the sample is more than 5% of the population, we must as well use the finite population multiplier in our standard error formulae. For instance, $\text{S.E.}_{\bar{X}}$ in case of finite population will be as under:

$$SE_{\bar{X}} = \frac{\sigma_p}{\sqrt{n}} \cdot \sqrt{\frac{(N-n)}{(N-1)}}$$

It may be remembered that in cases in which the population is very large in relation to the size of the sample, the finite population multiplier is close to one and has little effect on the calculation of S.E. As such when sampling fraction is less than 0.5, the finite population multiplier is generally not used.

(2) The use of all the above stated formulae has been explained and illustrated in context of testing of hypotheses in chapters that follow.

$$\sigma_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}} = \frac{\sqrt{\dfrac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}}{\sqrt{n}}$$

(ii) Standard error of difference between two sample means when σ_p is unknown

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

ESTIMATION

In most statistical research studies, population parameters are usually unknown and have to be estimated from a sample. As such the methods for estimating the population parameters assume an important role in statistical anlysis.

The random variables (such as $\bar{X}$ and σ_s^2) used to estimate population parameters, such as μ and σ_p^2 are conventionally called as '*estimators*', while specific values of these (such as $\bar{X} = 105$ or $\sigma_s^2 = 21.44$) are referred to as '*estimates*' of the population parameters. The estimate of a population parameter may be one single value or it could be a range of values. In the former case it is referred as *point estimate*, whereas in the latter case it is termed as *interval estimate*. The

researcher usually makes these two types of estimates through sampling analysis. While making estimates of population parameters, the researcher can give only the best point estimate or else he shall have to speak in terms of intervals and probabilities for he can never estimate with certainty the exact values of population parameters. Accordingly he must know the various properties of a good estimator so that he can select appropriate estimators for his study. He must know that a good estimator possesses the following properties:

(i) An estimator should on the average be equal to the value of the parameter being estimated. This is popularly known as the *property* of *unbiasedness*. An estimator is said to be unbiased if the expected value of the estimator is equal to the parameter being estimated. The sample mean $(\bar{X})$ is he most widely used estimator because of the fact that it provides an unbiased estimate of the population mean (μ).

(ii) An estimator should have a relatively small variance. This means that the most efficient estimator, among a group of unbiased estimators, is one which has the smallest variance. This property is technically described as the *property of efficiency*.

(iii) An estimator should use as much as possible the information available from the sample. This property is known as the *property of sufficiency*.

(iv) An estimator should approach the value of population parameter as the sample size becomes larger and larger. This property is referred to as the *property of consistency*.

Keeping in view the above stated properties, the researcher must select appropriate estimator(s) for his study. We may now explain the methods which will enable us to estimate with reasonable accuracy the population mean and the population proportion, the two widely used concepts.

ESTIMATING THE POPULATION MEAN (μ)

So far as the point estimate is concerned, the sample mean $\bar{X}$ is the best estimator of the population mean, μ, and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution. If we know the sampling distribution of $\bar{X}$, we can make statements about any estimate that we may make from the sampling information. Assume that we take a sample of 36 students and find that the sample yields an arithmetic mean of 6.2 i.e., $\bar{X} = 6.2$. Replace these student names on the population list and draw another sample of 36 randomly and let us assume that we get a mean of 7.5 this time. Similarly a third sample may yield a mean of 6.9; fourth a mean of 6.7, and so on. We go on drawing such samples till we accumulate a large number of means of samples of 36. Each such sample mean is a separate point estimate of the population mean. When such means are presented in the form of a distribution, the distribution happens to be quite close to normal. This is a characteristic of a distribution of sample means (and also of other sample statistics). Even if the population is not normal, the sample means drawn from that population are dispersed around the parameter in a distribution that is generally close to normal; the mean of the distribution of sample means is equal to the population mean.[5] This is true in case of large samples as per the dictates of the central limit theorem. This relationship between a population distribution and a distribution of sample

[5] C. William Emory, *Business Research Methods*, p.145

mean is critical for drawing inferences about parameters. The relationship between the dispersion of a population distribution and that of the sample mean can be stated as under:

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

where $\sigma_{\bar{X}}$ = standard error of mean of a given sample size

σ_p = standard deviation of the population

n = size of the sample.

How to find σ_p when we have the sample data only for our analysis? The answer is that we must use some best estimate of σ_p and the best estimate can be the standard deviation of the sample, σ_s. Thus, the standard error of mean can be worked out as under:[6]

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

where

$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}}$$

With the help of this, one may give interval estimates about the parameter in probabilistic terms (utilising the fundamental characteristics of the normal distribution). Suppose we take one sample of 36 items and work out its mean $\left(\bar{X}\right)$ to be equal to 6.20 and its standard deviation $\left(\sigma_s\right)$ to be equal to 3.8, Then the best point estimate of population mean $\left(\mu\right)$ is 6.20. The standard error of mean $\left(\sigma_{\bar{X}}\right)$ would be $3.8/\sqrt{36} = 3.8/6 = 0.663$. If we take the interval estimate of μ to be $\bar{X} \pm 1.96\left(\sigma_{\bar{X}}\right)$ or 6.20 ± 1.24 or from 4.96 to 7.44, it means that there is a 95 per cent chance that the population mean is within 4.96 to 7.44 interval. In other words, this means that if we were to take a complete census of all items in the population, the chances are 95 to 5 that we would find the population mean lies between 4.96 to 7.44*. In case we desire to have an estimate that will hold for a much smaller range, then we must either accept a smaller degree of confidence in the results or take a sample large enough to provide this smaller interval with adequate confidence levels. Usually we think of increasing the sample size till we can secure the desired interval estimate and the degree of confidence.

Illustration 1

From a random sample of 36 New Delhi civil service personnel, the mean age and the sample standard deviation were found to be 40 years and 4.5 years respectively. Construct a 95 per cent confidence interval for the mean age of civil servants in New Delhi.

Solution: The given information can be written as under:

[6] To make the sample standard deviation an unbiased estimate of the population, it is necessary to divide $\Sigma\left(X_i - \bar{X}\right)^2$ by $(n-1)$ and not by simply (n).

* In case we want to change the degree of confidence in the interval estimate, the same can be done using the table of areas under the normal curve.

$$n = 36$$
$$\bar{X} = 40 \text{ years}$$
$$\sigma_s = 4.5 \text{ years}$$

and the standard variate, z, for 95 per cent confidence is 1.96 (as per the normal curve area table).

Thus, 95 per cent confidence inteval for the mean age of population is:

$$\bar{X} \pm z \frac{\sigma_s}{\sqrt{n}}$$

or $$40 \pm 1.96 \frac{4.5}{\sqrt{36}}$$

or $$40 \pm (1.96)(0.75)$$

or $$40 \pm 1.47 \text{ years}$$

Illustration 2

In a random selection of 64 of the 2400 intersections in a small city, the mean number of scooter accidents per year was 3.2 and the sample standard deviation was 0.8.

(1) Make an estimate of the standard deviation of the population from the sample standard deviation.
(2) Work out the standard error of mean for this finite population.
(3) If the desired confidence level is .90, what will be the upper and lower limits of the confidence interval for the mean number of accidents per intersection per year?

Solution: The given information can be written as under:

$N = 2400$ (This means that population is finite)

$n = 64$

$\bar{X} = 3.2$

$\sigma_s = 0.8$

and the standard variate (z) for 90 per cent confidence is 1.645 (as per the normal curve area table).

Now we can answer the given questions thus:

(1) The best point estimate of the standard deviation of the population is the standard deviation of the sample itself.

Hence,

$$\hat{\sigma}_p = \sigma_s = 0.8$$

(2) Standard error of mean for the given finite population is as follows:

$$\sigma_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{0.8}{\sqrt{64}} \times \sqrt{\frac{2400 - 64}{2400 - 1}}$$

$$= \frac{0.8}{\sqrt{64}} \times \sqrt{\frac{2336}{2399}}$$

$$= (0.1)\ (.97)$$

$$= .097$$

(3) 90 per cent confidence interval for the mean number of accidents per intersection per year is as follows:

$$\bar{X} \pm z \left\{ \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N - n}{N - 1}} \right\}$$

$$= 3.2 \pm (1.645)\ (.097)$$

$$= 3.2 \pm .16 \text{ accidents per intersection.}$$

When the sample size happens to be a large one or when the population standard deviation is known, we use normal distribution for detemining confidence intervals for population mean as stated above. But how to handle estimation problem when population standard deviation is not known and the sample size is small (i.e., when $n < 30$)? In such a situation, normal distribution is not appropriate, but we can use *t*-distribution for our purpose. While using *t*-distribution, we assume that population is normal or approximately normal. There is a different *t*-distribution for each of the possible degrees of freedom. When we use *t*-distribution for estimating a population mean, we work out the degrees of freedom as equal to $n - 1$, where n means the size of the sample and then can look for cirtical value of '*t*' in the *t*-distribution table for appropriate degrees of freedom at a given level of significance. Let us illustrate this by taking an example.

Illustration 3

The foreman of *ABC* mining company has estimated the average quantity of iron ore extracted to be 36.8 tons per shift and the sample standard deviation to be 2.8 tons per shift, based upon a random selection of 4 shifts. Construct a 90 per cent confidence interval around this estimate.

Solution: As the standard deviation of population is not known and the size of the sample is small, we shall use *t*-distribution for finding the required confidence interval about the population mean. The given information can be written as under:

$$\bar{X} = 36.8 \text{ tons per shift}$$

$$\sigma_s = 2.8 \text{ tons per shift}$$

$$n = 4$$

degrees of freedom = $n - 1 = 4 - 1 = 3$ and the critical value of '*t*' for 90 per cent confidence interval or at 10 per cent level of significance is 2.353 for 3 *d.f.* (as per the table of *t*-distribution).

Thus, 90 per cent confidence interval for population mean is

$$\bar{X} \pm t \frac{\sigma_s}{\sqrt{n}}$$

$$= 36.8 \pm 2.353 \frac{2.8}{\sqrt{4}} = 36.8 \pm (2.353)(1.4)$$

$$= 36.8 \pm 3.294 \text{ tons per shift.}$$

ESTIMATING POPULATION PROPORTION

So far as the point estimate is concerned, the sample proportion (p) of units that have a particular characteristic is the best estimator of the population proportion $(\hat{p})$ and its sampling distribution, so long as the sample is sufficiently large, approximates the normal distribution. Thus, if we take a random sample of 50 items and find that 10 per cent of these are defective i.e., $p = .10$, we can use this sample proportion ($p = .10$) as best estimator of the population proportion $(\hat{p} = p = .10)$. In case we want to construct confidence interval to estimate a population poportion, we should use the binomial distribution with the mean of population $(\mu) = n \cdot p$, where n = number of trials, p = probability of a success in any of the trials and population standard deviation $= \sqrt{n\,p\,q}$. As the sample size increases, the binomial distribution approaches normal distribution which we can use for our purpose of estimating a population proportion. The mean of the sampling distribution of the proportion of successes (μ_p) is taken as equal to p and the standard deviation for the proportion of successes, also known as the standard error of proportion, is taken as equal to $\sqrt{pq/n}$. But when population proportion is unknown, then we can estimate the population parameters by substituting the corresponding sample statistics p and q in the formula for the standard error of proportion to obtain the estimated standard error of the proportion as shown below:

$$\sigma_p = \sqrt{\frac{p\,q}{n}}$$

Using the above estimated standard error of proportion, we can work out the confidence interval for population proportion thus:

$$p \pm z \cdot \sqrt{\frac{p\,q}{n}}$$

where

p = sample proportion of successes;
$q = 1 - p$;
n = number of trials (size of the sample);
z = standard variate for given confidence level (as per normal curve area table).

We now illustrate the use of this formula by an example.

Illustration 4

A market research survey in which 64 consumers were contacted states that 64 per cent of all consumers of a certain product were motivated by the product's advertising. Find the confidence limits for the proportion of consumers motivated by advertising in the population, given a confidence level equal to 0.95.

Solution: The given information can be written as under:

$n = 64$

$p = 64\%$ or $.64$

$q = 1 - p = 1 - .64 = .36$

and the standard variate (z) for 95 per cent confidence is 1.96 (as per the normal curve area table).

Thus, 95 per cent confidence interval for the proportion of consumers motivated by advertising in the population is:

$$p \pm z \cdot \sqrt{\frac{pq}{n}}$$

$$= .64 \pm 1.96 \sqrt{\frac{(0.64)(0.36)}{64}}$$

$$= .64 \pm (1.96)(.06)$$

$$= .64 \pm .1176$$

Thus, lower confidence limit is 52.24%

upper confidence limit is 75.76%

For the sake of convenience, we can summarise the formulae which give confidence intevals while estimating population mean (μ) and the population proportion $(\hat{p})$ as shown in the following table.

Table 8.3: Summarising Important Formulae Concerning Estimation

	In case of infinite population	*In case of finite population**
Estimating population mean (μ) when we know σ_p	$\bar{X} \pm z \cdot \frac{\sigma_p}{\sqrt{n}}$	$\bar{X} \pm z \cdot \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
Estimating population mean (μ) when we do not know σ_p	$\bar{X} \pm z \cdot \frac{\sigma_s}{\sqrt{n}}$	$\bar{X} \pm z \cdot \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$

Contd.

	In case of infinite population	*In case of finite population**
and use σ_s as the best estimate of σ_p and sample is large (i.e., $n > 30$)		
Estimating population mean (μ) when we do not know σ_p and use σ_s as the best estimate of σ_p and sample is small (i.e., $n \leq 30$)	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{n}}$	$\bar{X} \pm t \cdot \frac{\sigma_s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$
Estimating the population proportion $(\hat{p})$ when p is not known but the sample is large.	$p \pm z \cdot \sqrt{\frac{pq}{n}}$	$p \pm z \cdot \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}$

* In case of finite population, the standard error has to be multiplied by the finite population multiplier viz., $\sqrt{(N-n)/(N-1)}$.

SAMPLE SIZE AND ITS DETERMINATION

In sampling analysis the most ticklish question is: What should be the size of the sample or how large or small should be '*n*'? If the sample size ('*n*') is too small, it may not serve to achieve the objectives and if it is too large, we may incur huge cost and waste resources. As a general rule, one can say that the sample must be of an optimum size i.e., it should neither be excessively large nor too small. Technically, the sample size should be large enough to give a confidence inerval of desired width and as such the size of the sample must be chosen by some logical process before sample is taken from the universe. Size of the sample should be determined by a researcher keeping in view the following points:

(i) *Nature of universe:* Universe may be either homogenous or heterogenous in nature. If the items of the universe are homogenous, a small sample can serve the purpose. But if the items are heteogenous, a large sample would be required. Technically, this can be termed as the dispersion factor.

(ii) *Number of classes proposed:* If many class-groups (groups and sub-groups) are to be formed, a large sample would be required because a small sample might not be able to give a reasonable number of items in each class-group.

(iii) *Nature of study:* If items are to be intensively and continuously studied, the sample should be small. For a general survey the size of the sample should be large, but a small sample is considered appropriate in technical surveys.

(iv) *Type of sampling:* Sampling technique plays an important part in determining the size of the sample. A small random sample is apt to be much superior to a larger but badly selected sample.

(v) *Standard of accuracy and acceptable confidence level:* If the standard of acuracy or the level of precision is to be kept high, we shall require relatively larger sample. For doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold.

(vi) *Availability of finance:* In prctice, size of the sample depends upon the amount of money available for the study purposes. This factor should be kept in view while determining the size of sample for large samples result in increasing the cost of sampling estimates.

(vii) *Other considerations:* Nature of units, size of the population, size of questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study are a few other considerations to which a researcher must pay attention while selecting the size of the sample.

There are two alternative approaches for determining the size of the sample. The first approach is "to specify the precision of estimation desired and then to determine the sample size necessary to insure it" and the second approach "uses Bayesian statistics to weigh the cost of additional information against the expected value of the additional information."[7] The first approach is capable of giving a mathematical solution, and as such is a frequently used technique of determining '*n*'. The limitation of this technique is that it does not analyse the cost of gathering information *vis-a-vis* the expected value of information. The second approach is theoretically optimal, but it is seldom used because of the difficulty involved in measuring the value of information. Hence, we shall mainly concentrate here on the first approach.

DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON PRECISION RATE AND CONFIDENCE LEVEL

To begin with, it can be stated that whenever a sample study is made, there arises some sampling error which can be controlled by selecting a sample of adequate size. Researcher will have to specify the precision that he wants in respect of his estimates concerning the population parameters. For instance, a researcher may like to estimate the mean of the universe within ± 3 of the true mean with 95 per cent confidence. In this case we will say that the desired precision is ± 3, i.e., if the sample mean is Rs 100, the true value of the mean will be no less than Rs 97 and no more than Rs 103. In other words, all this means that the acceptable error, *e*, is equal to 3. Keeping this in view, we can now explain the determination of sample size so that specified precision is ensured.

(a) *Sample size when estimating a mean:* The confidence interval for the universe mean, μ, is given by

$$\bar{X} \pm z \frac{\sigma_p}{\sqrt{n}}$$

where $\bar{X}$ = sample mean;

z = the value of the standard variate at a given confidence level (to be read from the table giving the areas under normal curve as shown in appendix) and it is 1.96 for a 95% confidence level;

n = size of the sample;

[7] Rodney D. Johnson and Bernard R. Siskih, *Quantitative Techniques for Business Decisions*, p. 374–375.

σ_p = standard deviation of the popultion (to be estimated from past experience or on the basis of a trial sample). Suppose, we have $\sigma_p = 4.8$ for our purpose.

If the difference between μ and $\overline{X}$ or the acceptable error is to be kept with in ± 3 of the sample mean with 95% confidence, then we can express the acceptable error, '*e*' as equal to

$$e = z \cdot \frac{\sigma_p}{\sqrt{n}} \text{ or } 3 = 1.96 \frac{4.8}{\sqrt{n}}$$

Hence, $n = \dfrac{(1.96)^2 (4.8)^2}{(3)^2} = 9.834 \cong 10$.

In a general way, if we want to estimate μ in a population with standard deviation σ_p with an error no greater than '*e*' by calculating a confidence interval with confidence corresponding to *z*, the necessary sample size, *n*, equals as under:

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

All this is applicable whe the population happens to be infinite. Bu in case of finite population, the above stated formula for determining sample size will become

$$n = \frac{z^2 \cdot N \cdot \sigma_p^{2*}}{(N-1)e^2 + z^2 \sigma_p^2}$$

* In case of finite population the confidence interval for μ is given by

$$\overline{X} \pm z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$

where $\sqrt{(N-n)/(N-1)}$ is the finite population multiplier and all other terms mean the same thing as stated above.

If the precision is taken as equal to '*e*' then we have

$$e = z \frac{\sigma_p}{\sqrt{n}} \times \sqrt{\frac{(N-n)}{(N-1)}}$$

or $$e^2 = z^2 \frac{\sigma_p^2}{n} \times \frac{N-n}{N-1}$$

or $$e^2 (N-1) = \frac{z^2 \sigma_p^2 N}{n} - \frac{z^2 \sigma_p^2 n}{n}$$

or $$e^2 (N-1) + z^2 \sigma_p^2 = \frac{z^2 \sigma_p^2 N}{n}$$

or $$n = \frac{z^2 \cdot \sigma_p^2 \cdot N}{e^2 (N-1) + z^2 \sigma_p^2}$$

or $$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)e^2 + z^2 \sigma_p^2}$$

This is how we obtain the above stated formula for determining '*n*' in the case of infinite population given the precision and confidence level.

where

N = size of population

n = size of sample

e = acceptable error (the precision)

σ_p = standard deviation of population

z = standard variate at a given confidence level.

Illustration 5

Determine the size of the sample for estimating the true weight of the cereal containers for the universe with $N = 5000$ on the basis of the following information:

(1) the variance of weight = 4 ounces on the basis of past records.

(2) estimate should be within 0.8 ounces of the true average weight with 99% probability.

Will there be a change in the size of the sample if we assume infinite population in the given case? If so, explain by how much?

Solution: In the given problem we have the following:

$N = 5000$;

σ_p = 2 ounces (since the variance of weight = 4 ounces);

$e = 0.8$ ounces (since the estimate should be within 0.8 ounces of the true average weight);

$z = 2.57$ (as per the table of area under normal curve for the given confidence level of 99%).

Hence, the confidence interval for μ is given by

$$\bar{X} \pm z \cdot \frac{\sigma_p}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

and accordingly the sample size can be worked out as under:

$$n = \frac{z^2 \cdot N \cdot \sigma_p^2}{(N-1)\, e^2 + z^2\, \sigma_p^2}$$

$$= \frac{(2.57)^2 \cdot (5000) \cdot (2)^2}{(5000-1)\,(0.8)^2 + (2.57)^2\,(2)^2}$$

$$= \frac{132098}{3199.36 + 26.4196} = \frac{132098}{3225.7796} = 40.95 \cong 41$$

Hence, the sample size (or n) = 41 for the given precision and confidence level in the above question with finite population. But if we take population to be infinite, the sample size will be worked out as under:

$$n = \frac{z^2 \, \sigma_p^2}{e^2}$$

$$= \frac{(2.57)^2 \, (2)^2}{(0.8)^2} = \frac{26.4196}{0.64} = 41.28 \simeq 41$$

Thus, in the given case the sample size remains the same even if we assume infinite population.

In the above illustration, the standard deviation of the population was given, but in many cases the standard deviation of the population is not available. Since we have not yet taken the sample and are in the stage of deciding how large to make it (sample), we cannot estimate the populaion standard deviation. In such a situation, if we have an idea about the range (i.e., the difference between the highest and lowest values) of the population, we can use that to get a crude estimate of the standard deviation of the population for geting a working idea of the required sample size. We can get the said estimate of standard deviation as follows:

Since 99.7 per cent of the area under normal curve lies within the range of ± 3 standard deviations, we may say that these limits include almost all of the distribution. Accordingly, we can say that the given range equals 6 standard deviations because of ± 3. Thus, a rough estimate of the population standard deviation would be:

$$6\hat{\sigma} = \text{the given range}$$

or

$$\hat{\sigma} = \frac{\text{the given range}}{6}$$

If the range happens to be, say Rs 12, then

$$\hat{\sigma} = \frac{12}{6} = \text{Rs } 2.$$

and this estimate of standard deviation, $\hat{\sigma}$, can be used to determine the sample size in the formulae stated above.

(b) *Sample size when estimating a percentage or proportion:* If we are to find the sample size for estimating a proportion, our reasoning remains similar to what we have said in the context of estimating the mean. First of all, we shall have to specify the precision and the confidence level and then we will work out the sample size as under:

Since the confidence interval for universe proportion, $\hat{p}$ is given by

$$p \pm z \cdot \sqrt{\frac{p \cdot q}{n}}$$

where p = sample proportion, $q = 1 - p$;

z = the value of the standard variate at a given confidence level and to be worked out from table showing area under Normal Curve;

n = size of sample.

Since $\hat{p}$ is actually what we are trying to estimate, then what value we should assign to it ? One method may be to take the value of $p = 0.5$ in which case 'n' will be the maximum and the sample will yield at least the desired precision. This will be the most conservative sample size. The other method may be to take an initial estimate of p which may either be based on personal judgement or may be the result of a pilot study. In this context it has been suggested that a pilot study of something like 225 or more items may result in a reasonable approximation of p value.

Then with the given precision rate, the acceptable error, 'e', can be expressed as under:

$$e = z \cdot \sqrt{\frac{pq}{n}}$$

or

$$e^2 = z^2 \frac{pq}{n}$$

or

$$n = \frac{z^2 \cdot p \cdot q}{e^2}$$

The formula gives the size of the sample in case of infinite population when we are to estimate the proportion in the universe. But in case of finite population the above stated formula will be changed as under:

$$n = \frac{z^2 \cdot p \cdot q \cdot N}{e^2 (N-1) + z^2 \cdot p \cdot q}$$

Illustration 6

What should be the size of the sample if a simple random sample from a population of 4000 items is to be drawn to estimate the per cent defective within 2 per cent of the true value with 95.5 per cent probability? What would be the size of the sample if the population is assumed to be infinite in the given case?

Solution: In the given question we have the following:

$N = 4000$;

$e = .02$ (since the estimate should be within 2% of true value);

$z = 2.005$ (as per table of area under normal curve for the given confidence level of 95.5%).

As we have not been given the p value being the proportion of defectives in the universe, let us assume it to be $p = .02$ (This may be on the basis of our experience or on the basis of past data or may be the result of a pilot study).

Now we can determine the size of the sample using all this information for the given question as follows:

$$n = \frac{z^2 \cdot p \cdot q \cdot N}{e^2 (N-1) + z^2 \cdot p \cdot q}$$

$$= \frac{(2.005)^2\ (.02)\ (1 - .02)\ (4000)}{(.02)^2\ (4000 - 1) + (2.005)^2\ (.02)\ (1 - .02)}$$

$$= \frac{315.1699}{1.5996 + .0788} = \frac{315.1699}{1.6784} = 187.78 \simeq 188$$

But if the population happens to be infinite, then our sample size will be as under:

$$n = \frac{z^2 \cdot p \cdot q}{e^2}$$

$$= \frac{(2.005)^2 \cdot (.02)\ (1 - .02)}{(.02)^2}$$

$$= \frac{.0788}{.0004} = 196.98 \simeq 197$$

Illustration 7

Suppose a certain hotel management is interested in determining the percentage of the hotel's guests who stay for more than 3 days. The reservation manager wants to be 95 per cent confident that the percentage has been estimated to be within $\pm 3\%$ of the true value. What is the most conservative sample size needed for this problem?

Solution: We have been given the following:

Population is infinite;

e = .03 (since the estimate should be within 3% of the true value);

z = 1.96 (as per table of area under normal curve for the given confidence level of 95%).

As we want the most conservative sample size we shall take the value of p = .5 and q = .5. Using all this information, we can determine the sample size for the given problem as under:

$$n = \frac{z^2\ p\ q}{e^2}$$

$$= \frac{(1.96)^2 \cdot (.5)\ (1 - .5)}{(.03)^2} = \frac{.9604}{.0009} = 1067.11 \simeq 1067$$

Thus, the most conservative sample size needed for the problem is = 1067.

DETERMINATION OF SAMPLE SIZE THROUGH THE APPROACH BASED ON BAYESIAN STATISTICS

This approach of determining 'n'utilises Bayesian statistics and as such is known as Bayesian approach. The procedure for finding the optimal value of 'n' or the size of sample under this approach is as under:

(i) Find the expected value of the sample information (EVSI)* for every possible *n*;

(ii) Also workout reasonably approximated cost of taking a sample of every possible *n*;

(iii) Compare the EVSI and the cost of the sample for every possible *n*. In other words, workout the expected net gain (ENG) for every possible *n* as stated below:

For a given sample size (*n*):

(EVSI) – (Cost of sample) = (ENG)

(iv) Form (iii) above the optimal sample size, that value of *n* which maximises the difference between the EVSI and the cost of the sample, can be determined.

The computation of EVSI for every possible *n* and then comparing the same with the respective cost is often a very cumbersome task and is generally feasible with mechanised or computer help. Hence, this approach although being theoretically optimal is rarely used in practice.

Questions

1. Explain the meaning and significance of the concept of "Standard Error' in sampling analysis.
2. Describe briefly the commonly used sampling distributions.
3. State the reasons why sampling is used in the context of research studies.
4. Explain the meaning of the following sampling fundamentals:
   (a) Sampling frame;
   (b) Sampling error;
   (c) Central limit theorem;
   (d) Student's *t* distribution;
   (e) Finite population multiplier.
5. Distinguish between the following:
   (a) Statistic and parameter;
   (b) Confidence level and significance level;
   (c) Random sampling and non-random sampling;
   (d) Sampling of attributes and sampling of variables;
   (e) Point estimate and interval estimation.
6. Write a brief essay on statistical estimation.
7. 500 articles were selected at random out of a batch containing 10000 articles and 30 were found defective. How many defective articles would you reasonably expect to find in the whole batch?
8. In a sample of 400 people, 172 were males. Estimate the population proportion at 95% confidence level.
9. A smaple of 16 measurements of the diameter of a sphere gave a mean $\bar{X} = 4.58$ inches and a standard deviation $\sigma_s = 0.08$ inches. Find (a) 95%, and (b) 99% confidence limits for the actual diameter.
10. A random sample of 500 pineapples was taken from a large consignment and 65 were found to be bad. Show that the standard error of the population of bad ones in a sample of this size is 0.015 and also show that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5.

* EVSI happens to be the difference between the expected value with sampling and the expected value without sampling. For finding EVSI we have to use Bayesian statistics for which one should have a thorough knowledge of Bayesian probability analysis which can be looked into any standard text book on statistics.

11. From a packet containing iron nails, 1000 iron nails were taken at random and out of them 100 were found defective. Estimate the percentage of defective iron nails in the packet and assign limits within which the percentage probably lies.

12. A random sample of 200 measurements from an infinite population gave a mean value of 50 and a standard deviation of 9. Determine the 95% confidence interval for the mean value of the population.

13. In a random sample of 64 mangoes taken from a large consignment, some were found to be bad. Deduce that the percentage of bad mangoes in the consignment almost certainly lies between 31.25 and 68.75 given that the standard error of the proportion of bad mangoes in the sample 1/16.

14. A random sample of 900 members is found to have a mean of 4.45 cms. Can it be reasonably regarded as a sample from a large population whose mean is 5 cms and variance is 4 cms?

15. It is claimed that Americans are 16 pounds overweight on average. To test this claim, 9 randomly selected individuals were examined and the average excess weight was found to be 18 pounds. At the 5% level of significance, is there reason to believe the claim of 16 pounds to be in error?

16. The foreman of a certain mining company has estimated the average quantity of ore extracted to be 34.6 tons per shift and the sample standard deviation to be 2.8 tons per shift, based upon a random selection of 6 shifts. Construct 95% as well as 98% confidence interval for the average quantity of ore extracted per shift.

17. A sample of 16 bottles has a mean of 122 ml. (Is the sample representative of a large consignment with a mean of 130 ml.) and a standard deviation of 10 ml.? Mention the level of significance you use.

18. A sample of 900 days is taken from meteorological records of a certain district and 100 of them are found to be foggy. What are the probable limits to the percentage of foggy days in the district?

19. Suppose the following ten values represent random observations from a normal parent population:

2, 6, 7, 9, 5, 1, 0, 3, 5, 4.

Construct a 99 per cent confidence interval for the mean of the parent population.

20. A survey result of 1600 Playboy readers indicates that 44% finished at least three years of college. Set 98% confidence limits on the true proportion of all Playboy readers with this background.

21. (a) What are the alternative approaches of determining a sample size? Explain.

(b) If we want to draw a simple random sample from a population of 4000 items, how large a sample do we need to draw if we desire to estimate the per cent defective within 2 % of the true value with 95.45% probability. [*M. Phil. Exam. (EAFM) RAJ. Uni. 1979*]

22. (a) Given is the following information:

(i) Universe with N =10,000.

(ii) Variance of weight of the cereal containers on the basis of past records = 8 kg. Determine the size of the sample for estimating the true weight of the containers if the estimate should be within 0.4 kg. of the true average weight with 95% probability.

(b) What would be the size of the sample if infinite universe is assumed in question number 22 (a) above?

23. Annual incomes of 900 salesmen employed by Hi-Fi Corporation is known to be approximately normally distributed. If the Corporation wants to be 95% confident that the true mean of this year's salesmen's income does not differ by more than 2% of the last year's mean income of Rs 12,000, what sample size would be required assuming the population standard deviation to be Rs 1500?

[*M. Phil. (EAFM) Special Exam. RAJ. Uni. 1979*]

24. Mr. Alok is a purchasing agent of electronic calculators. He is interested in determining at a confidence level of 95% what proportion (within plus or minus 4%), is defective. Conservatively, how many calculators should be tested to find the proportion defective?

(*Hint:* If he tests conservatively, then p = .5 and q = .5).

25. A team of medico research experts feels confident that a new drug they have developed will cure about 80% of the patients. How large should the sample size be for the team to be 98% certain that the sample proportion of cure is within plus and minus 2% of the proportion of all cases that the drug will cure?

26. Mr. Kishore wants to determine the average time required to complete a job with which he is concerned. As per the last studies, the population standard deviation is 8 days. How large should the sample be so that Mr. Kishore may be 99% confident that the sample average may remain within ± 2 days of the average?