CHAPTER – 15

SAMPLING THEORY

LEARNING OBJECTIVES

In this chapter the student will learn-

- Different procedure of sampling which will be the best representative of the population;
- The concept of sampling distribution;
- The techniques of construction and interpretation of confidence interval estimates as well as sample size with defined degree of precision.

15.1 INTRODUCTION

There are situations when we would like to know about a vast, infinite universe or population. But some important factors like time, cost, efficiency, vastness of the population make it almost impossible to go for a complete enumeration of all the units constituting the population. Instead, we take recourse to selecting a representative part of the population and infer about the unknown universe on the basis of our knowledge from the known sample. A somewhat clear picture would emerge out if we consider the following cases.

In the first example let us share the problem faced by Mr. Basu. Mr. Basu would like to put a big order for electrical lamps produced by Mr. Ahuja's company "General Electricals". But before putting the order, he must know whether the claim made by Mr. Ahuja that the lamps of General Electricals last for at least 1500 hours is justified.

Miss Manju Bedi is a well-known social activist. Of late, she has noticed that the incidence of a particular disease in her area is on the rise. She claims that twenty per cent of the people in her town have been suffering from the disease.

In both the situations, we are faced with three different types of problems. The first problem is how to draw a representative sample from the population of electrical lamps in the first case and from the population of human beings in her town in the second case. The second problem is to estimate the population parameters i.e., the average life of all the bulbs produced by General Electricals and the proportion of people suffering form the disease in the first and second examples respectively on the basis of sample observations. The third problem relates to decision making i.e., is there enough evidence, once again on the basis of sample observations, to suggest that the claims made by Mr. Ahuja or Miss Bedi are justifiable so that Mr. Basu can take a decision about buying the lamps from General Electricals in the first case and some effective steps can be taken in the second example with a view to reducing the outbreak of the disease. We consider tests of significance or tests of hypothesis before decision making.

15.2 BASIC PRINCIPLES OF SAMPLE SURVEY

Sample Survey is the study of the unknown population on the basis of a proper representative sample drawn from it. How can a part of the universe reveal the characteristics of the unknown universe? The answer to this question lies in the basic principles of sample survey comprising the following components:

(a) Law of Statistical regularity

(b) Principle of Inertia

(c) Principle of Optimization

(d) Principle of Validity

(a) According to the law of statistical regularity, if a sample of fairly large size is drawn from the population under discussion at random, then on an average the sample would posses the characteristics of that population.

Thus the sample, to be taken from the population, should be moderately large. In fact larger the sample size, the better in revealing the identity of the population. The reliability of a statistic in estimating a population characteristics varies as the square root of the sample size. However, it is not always possible to increase the sample size as it would put an extra burden on the available resource. We make a compromise on the sample size in accordance with some factors like cost, time, efficiency etc.

Apart from the sample size, the sample should be drawn at random from the population which means that each and every unit of the population should have a pre-assigned probability to belong to the sample.

(b) The results derived from a sample, according to the principle of inertia of large numbers, are likely to be more reliable, accurate and precise as the sample size increases, provided other factors are kept constant. This is a direct consequence of the first principle.

(c) The principle of optimization ensures that an optimum level of efficiency at a minimum cost or the maximum efficiency at a given level of cost can be achieved with the selection of an appropriate sampling design.

(d) The principle of validity states that a sampling design is valid only if it is possible to obtain valid estimates and valid tests about population parameters. Only a probability sampling ensures this validity.

15.3 COMPARISON BETWEEN SAMPLE SURVEY AND COMPLETE ENUMERATION

When complete information is collected for all the units belonging to a population, it is defined as complete enumeration or census. In most cases, we prefer sample survey to complete enumeration due to the following factors:

(a) Speed: As compared to census, a sample survey could be conducted, usually, much more quickly simply because in sample survey, only a part of the vast population is enumerated.

(b) Cost: The cost of collection of data on each unit in case of sample survey is likely to be more as compared to census because better trained personnel are employed for conducting a sample survey. But when it comes to total cost, sample survey is likely to be less expensive as only some selected units are considered in a sample survey.

(c) Reliability: The data collected in a sample survey are likely to be more reliable than that in a complete enumeration because of trained enumerators better supervision and application of modern technique.

(d) Accuracy: Every sampling is subjected to what is known as sampling fluctuation which is termed as sampling error. It is obvious that complete enumeration is totally free from this sampling error. However, errors due to recording observations, biases on the part of the enumerators, wrong and faulty interpretation of data etc. are prevalent in both sampling and census and this type of error is termed as non-sampling errors. It may be noted that in sample survey, the sampling error can be reduced to a great extent by taking several steps like increasing the sample size, adhering to a probability sampling design strictly and so on. The non-sampling errors also can be contained to a desirable degree by a proper planning which is not possible or feasible in case of complete enumeration.

(e) Necessity: Sometimes, sampling becomes necessity. When it comes to destructive sampling where the items get exhausted like testing the length of life of electrical bulbs or sampling from a hypothetical population like coin tossing, there is no alternative to sample survey.

However, when it is necessary to get detailed information about each and every item constituting the population, we go for complete enumeration. If the population size is not large, there is hardly any merit to take recourse to sampling. If the occurrence of just one defect may lead to a complete destruction of the process as in an aircraft, we must go for complete enumeration.

15.4 ERRORS IN SAMPLE SURVEY

Errors or biases in a survey may be defined as the deviation between the value of population parameter as obtained from a sample and its observed value. Errors are of two types.

I. Sampling Errors

II. Non-Sampling Errors

Sampling Errors : Since only a part of the population is investigated in a sampling, every sampling design is subjected to this type of errors. The factors contributing to sampling errors are listed below:

(a) Errors arising out due to defective sampling design:

Selection of a proper sampling design plays a crucial role in sampling. If a non- probabilistic sampling design is followed, the bias or prejudice of the sampler affects the sampling technique thereby resulting some kind of error.

(b) Errors arising out due to substitution:

A very common practice among the enumerators is to replace a sampling unit by a suitable unit in accordance with their convenience when difficulty arises in getting information from the originally selected unit. Since the sampling design is not strictly adhered to, this results in some type of bias.

(c) Errors owing to faulty demarcation of units:

It has its origin in faulty demarcation of sampling units. In case of an agricultural survey, the sampler has, usually, a tendency to underestimate or overestimate the character under consideration.

(d) Errors owing to wrong choice of statistic:

One must be careful in selecting the proper statistic while estimating a population characteristic.

(e) Variability in the population:

Errors may occur due to variability among population units beyond a degree. This could be reduced by following somewhat complicated sampling design like stratified sampling, Multistage sampling etc.

Non-sampling Errors

As discussed earlier, this type of errors happen both in sampling and complete enumeration. Some factors responsible for this particular kind of biases are lapse of memory, preference for certain digits, ignorance, psychological factors like vanity, non- responses on the part of the interviewees wrong measurements of the sampling units, communication gap between the interviewers and the interviewees, incomplete coverage etc. on the part of the enumerators also lead to non-sampling errors.

15.5 SOME IMPORTANT TERMS ASSOCIATED WITH SAMPLING

Population or Universe

It may be defined as the aggregate of all the units under consideration. All the lamps produced by "General Electricals" in our first example in the past, present and future constitute the population. In the second example, all the people living in the town of Miss Manju form the population. The number of units belonging to a population is known as population size. If there are one lakh people in her town then the population size, to be denoted by N, is 1 lakh.

A population may be finite or infinite. If a population comprises only a finite number of units, then it is known as a finite population. The population in the second example is obviously, finite. If the population contains an infinite or uncountable number of units, then it is known as an infinite population. The population of electrical lamps of General Electricals is infinite. Similarly, the population of stars, the population of mosquitoes in Kolkata, the population of flowers in Mumbai, the population of insects in Delhi etc. are infinite population.

Population may also be regarded as existent or hypothetical. A population consisting of real objects is known as an existent population. The population of the lamps produced by General Electricals and the population of Miss Manju's town are example of existent populations. A population that exists just hypothetically like the population of heads when a coin is tossed infinitely is known as a hypothetical or an imaginary population.

Sample

A sample may be defined as a part of a population so selected with a view to representing the population in all its characteristics selection of a proper representative sample is pretty important because statistical inferences about the population are drawn only on the basis of the sample observations. If a sample contains n units, then n is known as sample size. If a sample of 500 electrical lamps is taken from the production process of General Electricals, then n = 500. The units forming the sample are known as "Sampling Units". In the first example, the sampling unit is electrical lamp and in the second example, it is a human. A detailed and complete list of all the sampling units is known as a "Sampling Frame". Before drawing sample, it is a must to have a updated sampling frame complete in all respects before the samples are actually drawn.

Parameter

A parameter may be defined as a characteristic of a population based on all the units of the population. Statistical inferences are drawn about population parameters based on the sample observations drawn from that population. In the first example, we are interested about the parameter "Population Mean". If x_α denotes the α^{th} member of the population, then population mean μ, which represents the average length of life of all the lamps produced by General Electricals is given by

$$\mu = \frac{\sum_{\alpha=1}^{n} x_\alpha}{N} \qquad \text{.....(15.1)}$$

Where N denotes the population size i.e. the total number of lamps produced by the company. In the second example, we are concerned about the population proportion P, representing the ratio of the people suffering from the disease to the total number of people in the town. Thus if there are X people possessing this attribute i.e. suffering from the disease, then we have

$$P = \frac{X}{N} \qquad \text{.....(15.2)}$$

Another important parameter namely the population variance, to be denoted by σ^2 is given by

$$\sigma^2 = \frac{\Sigma(X_\alpha - \mu)^2}{N} \qquad \text{.....(15.3)}$$

Also we have SD = $\sigma = \sqrt{\frac{\Sigma(X_\alpha - \mu)^2}{N}}$ .....(15.4)

Statistics

A statistic may be defined as a statistical measure of sample observation and as such it is a function of sample observations. If the sample observations are denoted by x_1, x_2, x_3, …….. x_n, then a statistic T may be expressed as T = f(x_1, x_2, x_3, …….. x_n)

A statistic is used to estimate a particular population parameter. The estimates of population mean, variance and population proportion are given by

$$\bar{x} = \hat{\mu} = \frac{\sum x_i}{n} \quad \text{.....(15.5)}$$

$$S^2 = \hat{\sigma}^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad \text{.....(15.6)}$$

$$\text{and } p = \hat{P} = \frac{x}{n} \quad \text{.....(15.7)}$$

Where x, in the last case, denotes the number of units in the sample in possession of the attribute under discussion.

Sampling Distribution and Standard Error of a Statistic

Starting with a population of N units, we can draw many a sample of a fixed size n. In case of sampling with replacement, the total number of samples that can be drawn is $(N)^n$ and when it comes to sampling without replacement of the sampling units, the total number of samples that can be drawn is ${}^{N}C_{n}$.

If we compute the value of a statistic, say mean, it is quite natural that the value of the sample mean may vary from sample to sample as the sampling units of one sample may be different from that of another sample. The variation in the values of a statistic is termed as "Sampling Fluctuations".

If it is possible to obtain the values of a statistic (T) from all the possible samples of a fixed sample size along with the corresponding probabilities, then we can arrange the values of the statistic, which is to be treated as a random variable, in the form of a probability distribution. Such a probability distribution is known as the sampling distribution of the statistic. The sampling distribution, just like a theoretical probability distribution possesses different characteristics. The mean of the statistic, as obtained from its sampling distribution, is known as "Expectation" and the standard deviation of the statistic T is known as the "Standard Error (SE)" of T. SE can be regarded as a measure of precision achieved by sampling. SE is inversely proportional to the square root of sample size. It can be shown that

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \text{ for SRS WR}$$

$$= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}} \quad \text{for SRS WOR} \quad \text{.....(15.8)}$$

Standard Error for Proporation

$$SE(p) = \sqrt{\frac{Pq}{n}} \quad \text{for SRS WR}$$

$$= \sqrt{\frac{Pq}{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{for SRS WOR} \quad \text{..... (15.9)}$$

SRSWR and SRSWOR stand for simple random sampling with replacement and simple random sampling without replacement.

The factor $\sqrt{\frac{N - n}{N - 1}}$ is known as finite population correction (fpc) or finite population multiplier and may be ignored as it tends to 1 if the sample size (n) is very large or the population under consideration is infinite when the parameters are unknown, they may be replaced by the corresponding statistic.

Illustrations

Example 15.1: A population comprises the following units: a, b, c, d, e. Draw all possible samples of size three without replacement.

Solution: Since in this case, sample size (n) = 3 and population size (N) = 5. the total number of possible samples without replacement = ${}^{5}c_{3} = 10$

These are abc, abd, abe, acd, ace, ade, bcd, bce, bde, cde.

Example 15.2: A population comprises 3 member 1, 5, 3. Draw all possible samples of size two
(i) with replacement

(ii) without replacement

Find the sampling distribution of sample mean in both cases.

Solution: (i) With replacement :- Since n = 2 and N = 3, the total number of possible samples of size 2 with replacement = $3^2 = 9$.

These are exhibited along with the corresponding sample mean in table 15.1. Table 15.2 shows the sampling distribution of sample mean i.e., the probability distribution of $\bar{X}$.

Table 15.1

All possible samples of size 2 from a population comprising 3 units under WR scheme

Serial No.	Sample of size 2 with replacement	Sample mean ($\bar{x}$)
1	1, 1	1
2	1, 5	3
3	1, 3	2
4	5, 1	3
5	5, 5	5
6	5, 3	4
7	3, 1	2
8	3, 5	4
9	3, 3	3

Table 15.2

Sampling distribution of sample mean

$\bar{X}$	1	2	3	4	5	Total
P	1 / 9	2 / 9	3 / 9	2 / 9	1 / 9	1

(ii) without replacement: As N = 3 and n = 2, the total number of possible samples without replacement = ${}^{N}C_{2} = {}^{3}C_{2} = 3$.

Table 15.3

Possible samples of size 2 from a population of 3 units under WOR scheme

Serial No.	Sample of size 2 without replacement	Sample mean ($\bar{x}$)
1	1 , 3	2
2	1 , 5	3
3	3 , 5	4

Table 15.4

Sampling distribution of mean

$\bar{X}$:	2	3	4	Total
P:	1 / 3	1/3	1/3	1

Example 15.3: Compute the standard deviation of sample mean for the last problem. Obtain the SE of sample mean applying 15.8 and show that they are equal.

Solution: We consider the following cases:

(i) with replacement :

Let U = $\bar{x}$ The sampling distribution of U is given by

U:	1	2	3	4	5
P:	1/9	2/9	3/9	2/9	1/9

$\therefore$ E (U) $= \Sigma P_i U_i$

$= 1/9\times1 + 2/9\times2 + 3/9\times3 + 2/9\times4 + 1/9\times5$

$= 3$

$E(U^2) = \Sigma P_i U_i^2$

$= 1/9\times1^2 + 2/9\times2^2 + 3/9\times3^2 + 2/9\times4^2 + 1/9\times5^2$

$= 31/3$

$\therefore v(\bar{x}) = v(u') = E(U^2) - [E(U)]^2$

$= 31/3 - 3^2$

$= 4/3$

Hence $SE_{\bar{x}} = \frac{2}{\sqrt{3}}$ .....(1)

Since the population comprises 3 units, namely 1, 5, and 3 we may take $X_1 = 1$, $X_2 = 5$, $X_3 = 3$

The population mean (μ) is given by

$$\mu = \frac{\sum X_\alpha}{N}$$

$$= \frac{1+5+3}{3} = 3$$

and the population variance $\sigma^2 = \frac{\sum (X_\alpha - \mu)^2}{N}$

$$= \frac{(1-3)^2 + (5-3)^2 + (3-3)^2}{3} = 8/3$$

Applying 15.8 we have , $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{8}{3}} \times \frac{1}{\sqrt{2}} = \frac{2}{\sqrt{3}}$ ...(2)

Thus comparing (1) and (2), we are able to verify the validity of the formula.

(ii) without replacement :

In this case, the sampling distribution of $V = \bar{x}$ is given by

V:	2	3	4
P:	1/3	1/3	1/3

$\therefore E(\bar{x}) = E(V) = 1/3\times2 + 1/3\times3 + 1/3\times4$

$= 3$

$V(\bar{x}) = Var(V) = E(v^2) - [E(v)]^2$

$= 1/3\times2^2 + 1/3\times3^2 + 1/3\times4^2 - 3^2$

$= 29/3 - 9$

$= 2/3$

$\therefore SE_{\bar{x}} = \sqrt{\frac{2}{3}}$

Applying 15.8, we have

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{8}{3}} \times \frac{1}{\sqrt{2}} \times \sqrt{\frac{3-2}{3-1}}$$

$$= \sqrt{\frac{2}{3}}$$

and thereby , we make the same conclusion as in the previous case.

15.6 TYPES OF SAMPLING

There are three different types of sampling which are

I. Probability Sampling

II. Non–Probability Sampling

III. Mixed Sampling

In the first type of sampling there is always a fixed, pre assigned probability for each member of the population to be a part of the sample taken from that population . When each member of the population has an equal chance to belong to the sample, the sampling scheme is known as Simple Random Sampling. Some important probability sampling other than simple random

sampling are stratified sampling, Multi Stage sampling, Multi Phase Sampling, Cluster Sampling and so on. In non- probability sampling , no probability attached to the member of the population and as such it is based entirely on the judgement of the sampler. Non-probability sampling is also known as Purposive or Judgement Sampling. Mixed sampling is based partly on some probabilistic law and partly on some pre decided rule. Systematic sampling belongs to this category. Some important and commonly used sampling process are described now.

Simple Random Sampling (SRS)

When the units are selected independent of each other in such a way that each unit belonging to the population has an equal chance of being a part of the sample, the sampling is known as Simple random sampling or just random sampling. If the units are drawn one by one and each unit after selection is returned to the population before the next unit is being drawn so that the composition of the original population remains unchanged at any stage of the sampling then the sampling procedure is known as Simple Random Sampling with replacement. If, however, once the units selected from the population one by one are never returned to the population before the next drawing is made, then the sampling is known as sampling without replacement. The two sampling methods become almost identical if the population is infinite i.e. vary large or a very large sample is taken from the population. The best method of drawing simple random sample is to use random sampling numbers.

Simple random sampling is a very simple and effective method of drawing samples provided (i) the population is not very large (ii) the sample size is not very small and (iii) the population under consideration is not heterogeneous i.e. there is not much variability among the members forming the population. Simple random sampling is completely free from Sampler's biases. All the tests of significance are based on the concept of simple random sampling.

Stratified Sampling

If the population is large and heterogeneous, then we consider a somewhat, complicated sampling design known as stratified sampling which comprises dividing the population into a number of strata or sub-populations in such a way that there should be very little variations among the units comprising a stratum and maximum variation should occur among the different strata. The stratified sample consists of a number of sub samples, one from each stratum. Different sampling scheme may be applied to different strata and , in particular, if simple random sampling is applied for drawing units from all the strata, the sampling procedure is known as stratified random sampling. The purpose of stratified sampling are (i) to make representation of all the sub populations (ii) to provide an estimate of parameter not only for all the strata but also and overall estimate (iii) reduction of variability and thereby an increase in precision.

There are two types of allocation of sample size. When there is prior information that there is not much variation between the strata variances. We consider "Proportional allocation" or "Bowely's allocation where the sample sizes for different strata are taken as proportional to the population sizes. When the strata-variances differ significantly among themselves, we take recourse to "Neyman's allocation" where sample size vary jointly with population size and population standard deviation i.e. $n_i \propto N_i S_i$. Here n_i denotes the sample size for the i^{th} stratum, N_i and S_i being the corresponding population size and population standard deviation. In case of Bowley's allocation, we have $n_i \propto N_i$.

Stratified sampling is not advisable if (i) the population is not large (ii) some prior information is not available and (iii) there is not much heterogeneity among the units of population.

Multi Stage Sampling

In this type of complicated sampling, the population is supposed to compose of first stage sampling units, each of which in its turn is supposed to compose of second stage sampling units, each of which again in its turn is supposed to compose of third stage sampling units and so on till we reach the ultimate sampling unit.

Sampling also, in this type of sampling design, is carried out through stages. Firstly, only a number of first stage units is selected. For each of the selected first stage sampling units, a number of second stage sampling units is selected. The process is carried out until we select the ultimate sampling units. As an example of multi stage sampling, in order to find the extent of unemployment in India, we may take state, district, police station and household as the first stage, second stage, third stage and ultimate sampling units respectively.

The coverage in case of multistage sampling is quite large. It also saves computational labour and is cost-effective. It adds flexibility into the sampling process which is lacking in other sampling schemes. However, compared to stratified sampling, multistage sampling is likely to be less accurate.

Systematic Sampling

It refers to a sampling scheme where the units constituting the sample are selected at regular interval after selecting the very first unit at random i.e., with equal probability. Systematic sampling is partly probability sampling in the sense that the first unit of the systematic sample is selected probabilistically and partly non- probability sampling in the sense that the remaining units of the sample are selected according to a fixed rule which is non-probabilistic in nature.

If the population size N is a multiple of the sample size n i.e. N = nk, for a positive integer k which must be less than n, then the systematic sampling comprises selecting one of the first k units at random, usually by using random sampling number and thereby selecting every k^{th} unit till the complete, adequate and updated sampling frame comprising all the members of the population is exhausted. This type of systematic sampling is known as "linear systematic sampling". K is known as "sample interval".

However, if N is not a multiple of n, then we may write N = nk + p, p < k and as before, we select the first unit from 1 to k by using random sampling number and thereafter selecting every kth unit in a cyclic order till we get the sample of the required size n. This type of systematic sampling is known as "circular systematic sampling."

Systematic sampling is a very convenient method of sampling when a complete and updated sampling frame is available. It is less time consuming, less expensive and simple as compared to the other methods of sampling. However, systematic sampling has a severe drawback. If there is an unknown and undetected periodicity in the sampling frame and the sampling interval is a multiple of that period, then we are going to get a most biased sample, which, by no stretch of imagination, can represent the population under investigation. Furthermore, since it is not a probability sampling, no statistical inference can be drawn about population parameter.

Purposive or Judgement sampling

This type of sampling is dependent solely on the discretion of the sampler and he applies his own judgement based on his belief, prejudice, whims and interest to select the sample. Since this type of sampling is non-probabilistic, it is purely subjective and, as such, varies from person to person. No statistical hypothesis can be tested on the basis of a purposive sampling.

15.7 THEORY OF ESTIMATION

While inferring statistically about a population parameter on the basis of a random sample drawn from the population, we face two different types of problems. In the first situation, the population under discussion is completely unknown to us and we would like to guess about the population parameter (s) from our knowledge about the sample observations. Thus, we may like to guess about the mean length of life of all the lamps produced by General Electricals once a random sample of lamps is drawn from the production process. This aspect is known as Estimation of population parameters.

In the second situation, some information about the population is already available and we would like to verify how far that information is valid on the basis of the random sample drawn from that population. This second aspect is known as tests of significance. As for example, we may be interested to verify whether the producer's claim in the first example that the lamps produced by General Electricals last at least 1500 hours is valid on the basis of a random sample of lamps produced by the company.

Point Estimation

Let us consider a population characterised by an unknown population parameter θ where θ could be population mean or population variance of a normal population. In order to estimate the parameter, we draw a random sample of size n from the population and let us denote the sample observations by, $x_1, x_2, x_3, \ldots\ldots\ldots x_n$. We are in search of a statistic T, which is a function of the sample observations $x_1, x_2, x_3, \ldots\ldots\ldots x_n$, that can estimate the parameter. T is known to be an estimator of the parameter θ if it estimates θ and this is denoted by

$$\hat{T} = \theta \quad \ldots\ldots\ldots\ldots (15.10)$$

T is described as, to be more precise, a point estimator of θ as T represents θ by a single value or point and the value of T, as obtained from the sample, is known as point estimate. The point estimator of population mean, population variance and population proportion are the corresponding sample statistics. Hence

$$\hat{\mu} = \bar{x}$$

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

and $\hat{P} = p$

which we have already discussed.

The criterion for an ideal estimator are

(a) Unbiased ness and minimum variance

(b) Consistency and Efficiency

(c) Sufficiency

(a) Unbiased ness and minimum variance:

A statistic T is known to be an unbiased estimator of the parameter θ if the expectation of T is θ. Thus T is unbiased of θ if

$E(T) = \theta$ ……(15.11)

If (15.11) does not hold then T is known to be a biased estimator of θ. The bias is known to be positive if $E(T) - \theta > 0$ and negative if $E(T) - \theta < 0$.

A statistic T is known to be a minimum variance unbiased estimator (MVUE) of θ if (i) T is unbiased for θ and (ii) T has the minimum variance among all the unbiased estimators of θ.

For a parameter θ, there exists a good number of unbiased statistics and that is why unbiased ness is considered along with minimum variance. The sample mean is an MVUE for population mean. The sample standard deviation

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

is a biased estimator of the population standard deviation σ. However, a slight adjustment can produce an unbiased estimator of σ. Instead of S if we consider

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

i.e. the sample standard deviation with divisor as (n – 1), then we get an unbiased estimator of σ. The sample proportion p is an MVUE for the population proportion P.

(b) Consistency and Efficiency

A statistic T is known to be consistent estimator of the parameter θ if the difference between T and θ can be made smaller and smaller by taking the sample size n larger and larger. Mathematically, T is consistent for θ if

$E(T) \to \theta$

and $V(T) \to 0$ as $n \to \infty$ (15.12)

the sample mean, sample SD and sample proportion are all consistent estimators for the corresponding population parameters.

A statistic T is known to be an efficient estimator of θ if T has the minimum standard error among all the estimators of θ when the sample size is kept fixed. Like unibiased estimators, more than one consistent estimator exists for θ. To choose the best among them, we consider that estimator which is both consistent and efficient. The sample mean is both consistent and efficient estimator for the population mean.

(c) A statistic T is known to be a sufficient estimator of θ if T contains all the information about θ. However, the sufficient statistics do not exists for all the parameters. The sample mean is a sufficient estimator for the corresponding population mean.

Illustrations

Example 15.4: A random sample of size 5 is taken from a population containing 100 units. If the sample observations are 10, 12, 13, 7, 18, find

(i) an estimate of the population mean

(ii) an estimate of the standard error of sample mean

Solution: The estimate of the population mean (μ) is given by

$$\hat{\mu} = \bar{x}$$

The estimate of the standard error of sample mean is given by

$$\hat{SE}_{\bar{x}} = \frac{\sqrt{n}}{\sqrt{n-1}} \frac{S}{\sqrt{n}} \text{ for SRSWR} = \frac{\sqrt{n}}{\sqrt{n-1}} \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \quad \text{For SRSWOR}$$

$$\text{i.e. } \hat{SE}_{\bar{x}} = S/\sqrt{n-1} \text{ for SRSWR} = \frac{S}{\sqrt{n-1}} \cdot \sqrt{\frac{(N-n)}{(N-1)}} \text{ for SRSWOR}$$

Table 15.5

Computation of sample mean and sample SD

x_i	x_i^2
10	100
12	144
13	169
7	49
18	324
60	786

$$\bar{x} = \frac{\sum x_i}{n} = 60/5 = 12$$

$$S^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$= 786/5 - 12$$

$$= 157.20 - 144$$

$$= 13.20 = (3.633)^2$$

Hence we have $\hat{\mu} = 12$

$$SE_{\bar{x}} = \frac{3.633}{\sqrt{5-1}} \text{ for SRSWR}$$

$$= \frac{3.633}{\sqrt{5-1}} \cdot \sqrt{\frac{100-5}{100-1}} \text{ for SRSWOR}$$

i.e. $\hat{SE}_{\bar{x}} = 1.82$ for SRSWR

$= 1.78$ for SRSWOR

Example 15.5: A random sample of 200 articles taken from a large batch of articles contains 15 defective articles.

(i) What is the estimate of the proportion of defective articles in the entire batch?

(ii) What is the estimate of the sample proportion of defective articles?

Solution: Since it is a very large batch, the fpc is ignored and we have

$$\hat{P} = p = \frac{15}{200} = 0.075$$

$$\hat{SE}_p = \sqrt{\frac{p(1-p)}{n}}$$

$$= \sqrt{\frac{0.075 \times (1-0.075)}{200}}$$

$= 0.02$

Interval Estimation

Instead of estimating a parameter θ by a single value, we may consider an interval of values which is supposed to contain the parameter θ. An interval estimate is always expressed by a pair of unequal real values and the unknown parameter θ lies between these two values. Hence, an interval estimation may be defined as specifying two values that contains the unknown parameter θ on the basis of a random sample drawn from the population in all probability.

On the basis of a random sample drawn from the population characterised by an unknown parameter θ, let us find two statistics T_1 and T_2 such that

$$P(T_1 < \theta) = \alpha_1$$

$$P(T_2 > \theta) = \alpha_2,$$

for any two small positive quantities α_1 and α_2.

Combining these two conditions, we may write

$P(T_1 \le \theta \le T_2) = 1 - \alpha$ where $\alpha = \alpha_1 + \alpha_2$ .....(15.13)

(15.13) implies that the probability that the unknown parameter q lies between the two statistic T_1 and T_2 is $(1 - \alpha)$. The interval $[T_1, T_2]$, $T_1 < T_2$, is known as 100 $(1 - \alpha)$ % confidence limits to θ. T_1 is known as the lower confidence limit (LCL) and T_2 is known as upper confidence limit (UCL) to θ.

$(1 - \alpha)$ is termed as confidence coefficient corresponding to the confidence interval $[T_1 , T_2]$. The term "confidence interval" has its origin in the fact that if we select $\alpha = 0.05$, then we feel confident that the interval $[T_1 , T_2]$, would contain the parameter θ in$(1 - \alpha)$ % or $(1 - 0.05)$ % or 95 per cent of cases and the amount of confidence is 95 percent. This further means that if repeated samples of a fixed size are taken from the population with the unknown parameter θ, then in 95 per cent of the cases, the interval $[T_1 , T_2]$ would contain θ and in the remaining 5 percent of the cases, it would fail to contain θ.

Confidence Interval for population mean

To begin with, let us assume that we have taken a random sample of size n from a normal population with mean μ and standard deviations σ. We assume further that the population standard deviation σ, is known i.e. its value is specified. From our discussion in the last chapter, we know that the sample mean $\bar{x}$ is normally distributed with mean μ and standard deviation = SE of $\bar{x} = \frac{\sigma}{\sqrt{n}}$

If the assumption of normality is not tenable, then also the sample mean follows normal distribution approximately, statistically known as asymptotically, with population mean μ and standard deviation as $\frac{\sigma}{\sqrt{n}}$, provided the sample size n is sufficiently large. If the sample size exceeds 30, then the asymptotic normality assumption holds. In order to select the appropriate confidence interval to the population mean, we need determine a quantity p, say, such that

$$P [\bar{x} - p \times SE (\bar{x}) \le \mu \le \bar{x} + p \times SE (\bar{x})] = 1 - \alpha \quad \ldots\ldots(15.14)$$

(15.14) finally leads to

$$\phi (p) = 1 - \alpha / 2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(15.15)$$

choosing α as 0.05, (15.15) becomes

$$\phi (p) = 0.975 = \phi (1.96)$$

$$=> p = 1.96$$

Hence 95% confidence interval to μ is given by

$$[\bar{x} - 1.96 \times SE (\bar{x}) , \bar{x} + 1.96 \times SE (\bar{x})] \ldots..(15.16)$$

In a similar manner, 99% confidence interval to μ is given by

$$[\bar{x} - 2.58 \times SE (\bar{x}) , \bar{x} + 2.58 \times SE (\bar{x})] \ldots\ldots.. (15.17)$$

In case the Population standard deviation σ is unknown, we replace σ by the corresponding sample standard deviation. With divisor as (n–1) instead of n and obtain 95% confidence interval to μ as

$$[\bar{x} - 1.96 \times \frac{S'}{\sqrt{n}} , \bar{x} + 1.96 \times \frac{S'}{\sqrt{n}}] \ldots. (15.18)$$

Also 99% confidence interval to μ is

$$[\bar{x} - 2.58 \times \frac{S'}{\sqrt{n}} , \bar{x} + 2.58 \times \frac{S'}{\sqrt{n}}] \ldots. (15.19)$$

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

These are shown in figure (15.1) and (15.2) respectively.

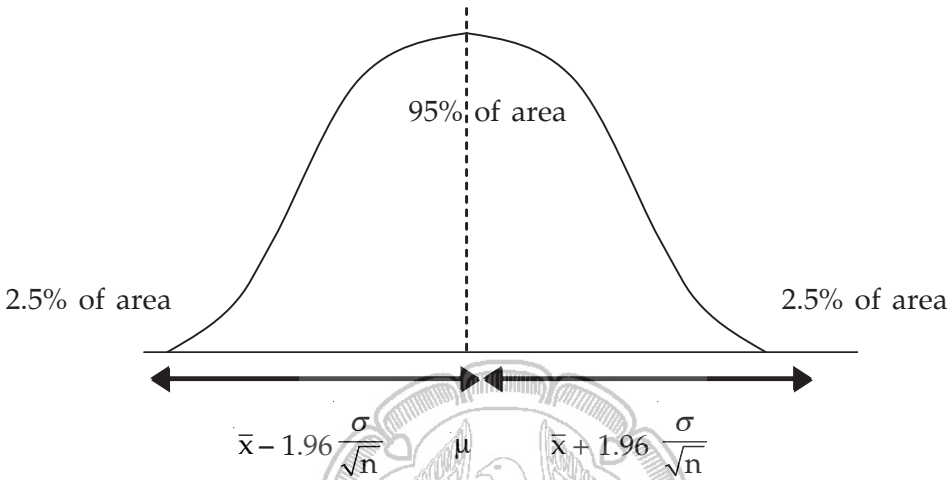

Figure 15.1
Showing 95 per cent confidence interval for population mean

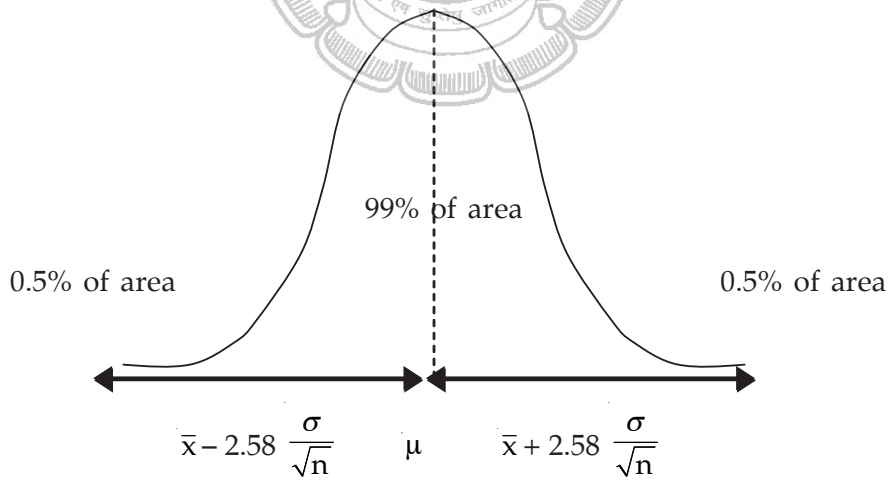

Figure 15.2
Showing 99 per cent confidence interval for population mean

After simplifying (15.18) and (15.19), we have

95% confidence interval to $\mu = \bar{x} \pm 1.96\, S/\sqrt{n-1}$

and 99% confidence interval $\bar{x} \pm 2.58\frac{S}{\sqrt{n-1}}$ .....(15.20)

When the population standard deviation is unknown and the sample size does not exceed 30, we consider

$$\sqrt{n-1}\,\frac{(\bar{x}-\mu)}{S}$$

which, as we have discussed in the last chapter follows t – distribution with (n–1) degrees of freedom (df). The 100 (1 – α) % confidence interval to μ is given by

$$\bar{x} - \frac{s}{\sqrt{n-1}} t_{\frac{\alpha}{2},(n-1)}, \bar{x} + \frac{s}{\sqrt{n-1}} t_{\frac{\alpha}{2},(n-1)} \quad \text{.....(15.21)}$$

Where S denotes the sample standard deviation and t_{p} ; (n–1) denotes upper p per cent point of the t - distribution with (n–1) df. The values of t_{p} ; (n–1) for different values of p and n are provided in the Biometrika Table. In particular, if we take α = 0.05 then the 95% lower confidence limit to μ is

$$\bar{x} - \frac{s}{\sqrt{n-1}} . t_{0.025,(n-1)}$$

and the corresponding upper confidence limit to μ is

$$\bar{x} + \frac{s}{\sqrt{n-1}} t_{0.025,(n-1)} \quad \text{..... (15.22)}$$

Similarly, 99% LCL to μ is $\bar{x} - \frac{s}{\sqrt{n-1}} \cdot t_{0.005\ (n-1)}$

and 99% UCL to μ is $\bar{x} + \frac{s}{\sqrt{n-1}} \cdot t_{0.005,\ (n-1)}$ .....(15.23)

Interval estimation of population proportion

When the sample size is large and both p and q = 1 – p, p being sample proportion, are not very small, the sample proportion follows asymptotic normal distribution with mean P and

SD = SE (p) $\sqrt{\frac{PQ}{n}}$

The estimate of SE (p) is given by

$\sqrt{\frac{PQ}{n}}$, ignoring the fpc.

Hence 100 $(1 - \alpha)$% confidence interval to p is

$$p - z_\alpha\sqrt{\frac{pq}{n}}, p + z_\alpha\sqrt{\frac{pq}{n}} \quad \text{.....(15.24)}$$

We take z_α = 1.96 for α = 0.05

= 2.58 for α = 0.01

Illustrations:

Example 15.5: A factory produces 60000 pairs of shoes on a daily basis. From a sample of 600 pairs, 3 per cent were found to be of inferior quality. Estimate the number of pairs that can be reasonably expected to be spoiled in the daily production process at 95% level of confidence.

Solution : Here we are given p = 0.03 , n = 600

and N = 60000

$$\therefore \hat{SE}(p) \sqrt{\frac{Pq}{n}}\sqrt{\frac{N-n}{N-1}} \quad \text{(including fpc)}$$

$$= \sqrt{\frac{0.03\times(1-0.03)}{600}} \times \sqrt{\frac{60000-600}{60000-1}}$$

= 0.0069.

Hence, 95% confidence limit to P

= [p – 1.96×SE (p) , p + 1.96 SE (p)] (from 15.24)

= [0.03 – 1.96×0.00692, 0.03 + 1.96×0.006]

= [0.01636, 0.04364]

Thus the number of pairs that can be reasonably expected to be spoiled in the entire production process on a daily basis at 95% level of confidence

= [0.01636×60000, 0.04364×60000]

[982, 2618]

Example 15.6: The marks obtained by a group of 15 students in statistic in an examination have a mean 55 and variance 49. What are the 99% confidence limits for the mean of the population of marks, assuming it to be normal. Given that the upper 0.5 per cent value of t distribution with 14 df is 2.98.

Solution: Let X denote the marks of the students in the population. Since (i) X is normally distributed as per the assumption (ii) the population standard deviation unknown (iii) the sample size (n) is less than 30, we consider t- distribution for finding confidence limits to the population mean μ of marks.

Here $\bar{x}$ = 55, S = 7, n = 15

From (15.23), 99% LCL to μ

$$= \bar{x} - \frac{s}{\sqrt{n-1}} \times t_{0.005,(n-1)}$$

$$= 55 - \frac{7}{\sqrt{15-1}} \times t_{0.005,(15-1)}$$

$$= 55 - \frac{7}{\sqrt{14}} \times t_{0.005,14}$$

= 55 – 1.8708 × 2.98 (as given $t_{0.005, 14} = 2.98$)

= 55 – 5.5750

= 49.43

The 99% UCL to μ

$$= \bar{x} + \frac{s}{\sqrt{n-1}} t_{0.005,\ (n-1)}$$

= 55+5.5750

= 60.58

Example 15.7: A pharmaceutical company wants to estimate the mean life of a particular drug under typical weather conditions. A simple random sample of 81 bottles yields the following information:

Sample mean = 23 months

population variance = 6.25 (months) ^{2}

Find an interval estimate with a confidence level of (i) 90% (ii) 98%

Solution: Since the sample size n = 81 is large, the mean life of the drug under consideration ($\bar{x}$) is asymptotically normal with population mean μ and SE = standard deviation

$$= \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{6.25}}{\sqrt{81}}$$

$$= \frac{2.50}{9} = 0.2778$$

(i) Consulting Biometrika table, we find that $\phi(p) = 1 - \alpha / 2$

$$\Rightarrow \phi(p) = 1 - \frac{0.10}{2}$$

$$= 0.95 = \phi(1.645)$$

$$\Rightarrow p = 1.6450$$

From (15.14), 90% confidence interval for μ is

$$[\bar{x} - p \times SE(\bar{x}),\ \bar{x} + p \times SE(\bar{x})]$$

$$= [23 - 1.6450 \times 0.2778,\ 23 + 1.645 \times 0.27778]$$

$$= [22.5430,\ 23.4570]$$

(ii) In this case, $\phi(p) = 1 - 0.02 / 2 = 0.99 = \phi(2.325)$

$$\Rightarrow p = 2.3250$$

thus, 98% confidence interval to μ

$$= (23 - 2.3250 \times 0.27778,\ 23 + 2.325 \times 0.27778)$$

$$= [22.3542,\ 23.6458]$$

Example 15.8: A random sample of 100 days shows an average daily sale of Rs. 1000 with a standard deviation of Rs. 250 in a particular shop. Assuming a normal distribution, find the limits which have a 95% chance of including the expected sales per day.

Solution: As given, n= 100,

$\bar{x}$ = average sales of the shop as obtained from the sample = Rs. 1000

S = standard deviation of sales as obtained from sample = Rs 250

From (15.20), we find that the 95% confidence interval to the expected sales per day (μ) is given by

$$\text{Rs. } [\bar{x} \pm 1.96 \frac{s}{\sqrt{n-1}}]$$

$$= \text{Rs. } [1000 \pm 1.96 \times \frac{250}{\sqrt{99}}]$$

$$= \text{Rs. } [1000 \pm 49.25]$$

$$= [\text{Rs } 950.75 , \text{Rs. } 1049.25]$$

15.8 DETERMINATION OF SAMPLE SIZE FOR A SPECIFIC PRECISION

In case of variable, we know that the sample mean $\bar{x}$ follows normal distribution with population mean μ and

SD = SE ($\bar{x}$) = $\frac{\sigma}{\sqrt{n}}$,

n denoting the size of the random sample drawn from the population . Letting E stands for the admissible error while estimating μ, the approximate sample size is given by

$$n = \left[\frac{\sigma p_\alpha}{E}\right]^2 \quad \text{.....(15.25)}$$

p_α denotes upper α per cent points of the standard normal distribution and assumes the values 1.96 and 2.58 respectively for 5% and 1% level of significance.

For an attribute, we have

$$n = \frac{Pqp^2\alpha}{E^2} \quad \text{.....(15.26)}$$

Where P= population proportion

$q = 1 - P$

where P is unknown, we replace it by the corresponding sample estimate p.

Example 15.9: In measuring reaction time, a psychologist estimated that the standard deviation is 1.08 seconds. What should be the size of the sample in order to be 99% confident that the error of her estimates of mean would not exceed 0.18 seconds ?

Solution: Let n be the size of the random sample.

As given, σ = 1.08, p_α = 2.58, E = 0.18

Applying (15.25) , we have $n = \left[\frac{1.08\times 2.58}{0.18}\right]^2$

$\cong$ 240

Example 15.10: The incidence of a particular disease in an area is such that 20 per cent people of that area suffers from it. What size of sample should be taken so as to ensure that the error of estimation of the proportion should not be more than 5 per cent with 95 per cent confidence?

Solution: Let n denote the required sample size.

As given P = 0.2, q = 1 – P = 0.8 p_α= 1.96 and E = 0.05

Applying (15.26), we have $n = \frac{Pqp^2_\alpha}{E^2}$

$$= \frac{0.2\times 0.8\times (1.96)^2}{(0.05)^2}$$

$\cong$ 246

EXERCISE

Set A

Answer the following questions. Each question carries one mark.

1. Sampling can be described as a statistical procedure
   (a) To infer about the unknown universe from a knowledge of any sample
   (b) To infer about the known universe from a knowledge of a sample drawn from it
   (c) To infer about the unknown universe from a knowledge of a random sample drawn from it
   (d) Both (a) and (b).
2. The Law of Statistical Regularity says that
   (a) Sample drawn from the population under discussion possesses the characteristics of the population
   (b) A large sample drawn at random from the population would posses the characteristics of the population
   (c) A large sample drawn at random from the population would possess the characteristics of the population on an average
   (d) An optimum level of efficiency can be attained at a minimum cost.
3. A sample survey is prone to
   (a) Sampling errors (b) Non-sampling errors
   (c) Either (a) or (b) (d) Both (a) and (b)
4. The population of roses in Salt Lake City is an example of
   (a) A finite population (b) An infinite population
   (c) A hypothetical population (d) An imaginary population.
5. Statistical decision about an unknown universe is taken on the basis of
   (a) Sample observations (b) A sampling frame
   (c) Sample survey (d) Complete enumeration
6. Random sampling implies
   (a) Haphazard sampling (b) Probability sampling
   (c) Systematic sampling (d) Sampling with the same probability for each unit.
7. A parameter is a characteristic of
   (a) Population (b) Sample (c) Both (a) and (b) (d) (a) or (b)

8. A statistic is
   (a) A function of sample observations
   (b) A function of population units
   (c) A characteristic of a population
   (d) A part of a population.
9. Sampling Fluctuations may be described as
   (a) The variation in the values of a statistic
   (b) The variation in the values of a sample
   (c) The differences in the values of a parameter
   (d) The variation in the values of observations.
10. The sampling distribution is
   (a) The distribution of sample observations
   (b) The distribution of random samples
   (c) The distribution of a parameter
   (d) The probability distribution of a statistic.
11. Standard error can be described as
   (a) The error committed in sampling
   (b) The error committed in sample survey
   (c) The error committed in estimating a parameter
   (d) Standard deviation of a statistic.
12. A measure of precision obtained by sampling is given by
   (a) Standard error
   (b) Sampling fluctuation
   (c) Sampling distribution
   (d) Expectation.
13. As the sample size increases, standard error
   (a) Increases
   (b) Decreases
   (c) Remains constant
   (d) Decreases proportionately.
14. If from a population with 25 members, a random sample without replacement of 2 members is taken, the number of all such samples is
   (a) 300 (b) 625 (c) 50 (d) 600
15. A population comprises 5 members. The number of all possible samples of size 2 that can be drawn from it with replacement is
   (a) 100 (b) 15 (c) 125 (d) 25

16. Simple random sampling is very effective if
    (a) The population is not very large
    (b) The population is not much heterogeneous
    (c) The population is partitioned into several sections.
    (d) Both (a) and (b)

17. Simple random sampling is
    (a) A probabilistic sampling
    (b) A non- probabilistic sampling
    (c) A mixed sampling
    (d) Both (b) and (c).

18. According to Neyman's allocation, in stratified sampling
    (a) Sample size is proportional to the population size
    (b) Sample size is proportional to the sample SD
    (c) Sample size is proportional to the sample variance
    (d) Population size is proportional to the sample variance.

19. Which sampling provides separate estimates for population means for different segments and also an over all estimate?
    (a) Multistage sampling
    (b) Stratified sampling
    (c) Simple random sampling
    (d) Systematic sampling

20. Which sampling adds flexibility to the sampling process?
    (a) Simple random sampling
    (b) Multistage sampling
    (c) Stratified sampling
    (d) Systematic sampling

21. Which sampling is affected most if the sampling frame contains an undetected periodicity?
    (a) Simple random sampling
    (b) Stratified sampling
    (c) Multistage sampling
    (d) Systematic sampling

22. Which sampling is subjected to the discretion of the sampler?
    (a) Systematic sampling
    (b) Simple random sampling
    (c) Purposive sampling
    (d) Quota sampling.

23. The criteria for an ideal estimator are
    (a) Unbiasedness, consistency, efficiency and sufficiency
    (b) Unbiasedness, expectation, sampling and estimation
    (c) Estimation, consistency, sufficiency and efficiency
    (d) Estimation, expectation, unbiasedness and sufficiency.

24. The sample standard deviation is
    (a) A biased estimator (b) An unbiased estimator.
    (c) A biased estimator for population SD
    (d) A biased estimator for population variance.
25. The sample mean is
    (a) An MVUE for population mean
    (b) A consistent and efficient estimator for population mean
    (c) A sufficient estimator for population mean
    (d) All of these.
26. For an unknown parameter, how many interval estimates exist?
    (a) Only one (b) Two (c) Three (d) Many
27. The most commonly used confidence interval is
    (a) 95 percent (b) 90 percent (c) 94 percent (d) 98 percent.

Set B

Answer the following question. Each question carries 2 marks.

1. If a random sample of size 2 with replacement is taken from the population containing the units 3,6 and 1, then the samples would be
    (a) (3,6), (3,1), (6,1) (b) (3,3), (6,6), (1,1)
    (c) (3,3), (3,6), (3,1), (6,6), (6,3), (6,1), (1,1), (1,3), (1,6)
    (d) (1,1), (1,3), (1,6), (6,1), (6,2), (6,3), (6,6), (1,6), (1,1)
2. If a random sample of size two is taken without replacement from a population containing the units a,b,c and d then the possible samples are
    (a) (a, b), (a, c), (a, d) (b) (a, b),(b, c), (c, d)
    (c) (a, b), (b, a), (a, c), (c,a), (a, d), (d, a) (d) (a, b), (a, c), (a, d), (b, c), (b, d), (c,d)
3. If a random sample of 500 oranges produces 25 rotten arranges, then the estimate of SE of the proportion of rotten arranges in the sample is
    (a) 0.01 (b) 0.05 (c) 0.028 (d) 0.0593
4. If the population SD is known to be 5 for a population containing 80 units, then the standard error of sample mean for a sample of size 25 without replacement is
    (a) 5 (b) 0.20 (c) 1 (d) 0.83
5. A simple random sample of size 16 is drawn from a population with 50 members. What is the SE of sample mean if the population variance is known to be 25 given that the sampling is done with replacement?
    (a) 1.25 (b) 6.25 (c) 1.04 (d) 1.56

6. A simple random sample of size 10 is drawn without replacement from a universe containing 85 units. If the mean and SD, as obtained from the sample, are 90 and 4 respectively, what is the estimate of the standard error of sample mean?

(a) 0.58 (b) 0.63 (c) 0.67 (d) 0.72

7. A sample of size 3 is taken from a population of 10 members with replacement. If the sample observations are 1, 3 and 5, what is the estimate of the standard error of sample mean?

(a) 1.96 (b) 2.00 (c) 2.25 (d) 2.28

8. If n numbers are drawn at random without replacement from the set { 1, 2,..,m}, then var. $(\bar{x})$ would be

(a) (m+1) (m–n)/12n (b) (m–1) (m+ n)/12

(c) (m–1) (m +n)/12n (d) (m–1) (m+n) / 12m

9. A random sample of the heights of 100 students from a large population of students having SD as 0.35m show an average height of 1.75m. What are the 95% confidence limits for the average height of all the students forming the population?

(a) [1.68 m, 1.82 m] (b) [1.58 m , 1.90 m] (c) [1.58m, 1.92m] (d) [1.5m, 2.0m]

10. A random sample of size 17 has 52 as mean. The sum of squares of deviation from mean is 160. The 99% confidence limits for the mean are

[Given $t_{0.01,15} = 2.60$, $t_{0.01,16} = 2.58$ $t_{0.01,17} = 2.57$ $t_{0.005,15} = 2.95$ $t_{0.005,16} = 2.92$ $t_{0.05,17} = 2.90$]

(a) [43, 6] (b) [45, 59] (c) [42.77, 61.23] (d) [48, 56]

11. A random sample of size 82 was taken to estimate the mean annual income of 500 families and the mean and SD were found to be Rs.7500 and Rs.80 respectively. What is upper confidence limit to the average income of all the families when the confidence level is 90 percent?

[Given ϕ (2.58) = 0.95]

(a) Rs.7600 (b) Rs.7582 (c) Rs.7520.98 (d) Rs.7522.93

12. 8 Life Insurance Policies in a sample of 100 taken out of 20,000 policies were found to be insured for less than Rs.10,000. How many policies in the whole lot can be expected to be insured for less than Rs. 10,000 at 95% confidence level?

(a) 1050 and 2150 (b) 1058 and 2142 (c) 1040 and 2160 (d) 1023 and 2057

13. A random sample of a group of people is taken and 120 were found to be in favor of liberalizing licensing regulations. If the proportion of people in the population found in favor of liberalization with 95% confidence lies between 0.683 and 0.817, then the number of people in the group is

(a) 140 (b) 150 (c) 160 (d) 175

14. A Life Insurance Company has 1500 policies averaging Rs.2000 on lives at age 30. From experience, it is found that out of 100,000 alive at age 30, 99,000 survive at age 31. What is the lower value of the amount that the company will have to pay in insurance during the year?

    (a) Rs.6000 (b) Rs.8000 (c) Rs.8200 (d) Rs.8500

15. If it is known that the 95% LCL and UCL to population mean are 48.04 and 51.96 respectively, what is the value of the population variance when the sample size is 100?

    (a) 8 (b) 10 (c) 12 (d) 12.50

ANSWERS

Set A											
1.	(c)	2.	(c)	3.	(d)	4.	(b)	5.	(a)	6.	(d)
7.	(a)	8.	(a)	9.	(a)	10.	(d)	11.	(d)	12.	(a)
13.	(b)	14.	(a)	15.	(c)	16.	(d)	17.	(a)	18.	(a)
19.	(b)	20.	(d)	21.	(d)	22.	(c)	23.	(a)	24.	(c)
25.	(d)	26.	(d)	27.	(a)						
Set B											
1.	(c)	2.	(d)	3.	(a)	4.	(d)	5.	(a)	6.	(b)
7.	(b)	8.	(a)	9.	(c)	10.	(c)	11.	(c)	12.	(b)
13.	(c)	14.	(a)	15.	(b)						

ADDITIONAL QUESTION BANK

1. Statistical data may be collected by complete enumeration called

   (a) Census inquiry (b) Sample inquiry (c) both (d) none

2. Statistical data may be collected by partial enumeration called

   (a) Census inquiry (b) Sample inquiry (c) both (d) none

3. The primary object of sampling is to obtain ———— information about population with ———— effort.

   (a) maximum, minimum (b) minimum, maximum
   (c) some, less (d) none

4. A ———— is a complete or whole set of possible measurements/data corresponding to the entire collection of units.

   (a) Sample (b) Population (c) both (d) none

5. A ———— is the set of measurement/data that are actually selected in the course of an investigation/enquiry.

   (a) Sample (b) Population (c) both (d) none

6. Sampling error is ———— proportional to the square root of the number of items in the sample.

   (a) inversely (b) directly (c) equally (d) none

7. Two basic Statistical laws concerning a population are

   (a) the law of statistical irregularity and the law of inertia of large numbers.

   (b) the law of statistical regularity and the law of inertia of large number Rs.

   (c) The law of statistical regularity and the law of inertia of small number Rs.

   (d) The law of statistical regularity and the law of inertia of small number Rs.

8. The ———— the size of the sample more reliable is the result.

   (a) medium (b) smaller (c) larger (d) none

9. Sampling is the process of obtaining a

   (a) population (b) sample (c) frequency (d) none

10. By using sampling methods we have

    (a) the error estimation & less quality data

    (b) less quality data & lower costs.

    (c) The error estimation & higher quality data.

    (d) higher quality data & higher costs.

11. Under ———————— method selection is often based on certain predetermined criteria.
    (a) Block or Cluster sampling
    (b) Area sampling
    (c) Quota sampling
    (d) Deliberate, purposive or judgment sampling.
12. ———————— sampling is similar to cluster sampling.
    (a) Judgment (b) Quota (c) Area (d) none
13. Value of a ———————— is different for different samples.
    (a) statistic (b) skill (c) both (d) none
14. A statistic is a ———————— variable.
    (a) simple (b) compound (c) random (d) none
15. The distribution of a ———————— is called sampling distribution of that ————————.
    (a) statistic, statistic (b) probability, probability (c) both (d) none
16. A ———————— distribution is a theoretical distribution that expresses the functional relation between each of the distinct values of the sample statistic and the corresponding probability.
    (a) normal (b) Binomial (c) Poisson (d) sampling.
17. Sampling distribution is a frequency distribution.
    (a) true (b) false (c) both (d) none
18. Sampling distribution approaches ———————— distribution when the population distribution is not normal provided the sample size is sufficiently large.
    (a) Binomial (b) Normal (c) Poisson (d) none
19. The Standard deviation of the ———————— distribution is called standard error.
    (a) normal (b) Poisson (c) Binomial (d) sampling
20. The difference of the actual value and the expected value using a model is
    (a) Error in statistics (b) Absolute error (c) Percentage error (d) Relative error.
21. The measure of divergence is ———————— as the size of the sample approaches that of the population.
    (a) more (b) less (c) same (d) none
22. The distribution of sample ———————— being normally or approximately normally distributed about the population.
    (a) median (b) mode (c) mean (d) none

23. The standard error of the ———————— is the standard deviation of sample means.

(a) median (b) mode (c) mean (d) none

24. There are ———————— types of estimates about a population parameter.

(a) five (b) Two (c) three (d) four.

25. To estimate an unknown population parameter

(a) interval estimate (b) Error estimate (c) Point estimate (d) none is used.

26. When we have an idea of the error that might be involved, we use

(a) Point estimate (b) interval estimate (c) both (d) none

27. The estimate which is used in making estimation of a population parameter is

(a) point (b) interval (c) both (d) none

28. A ———————— estimate is a single number.

(a) point (b) interval (c) both (d) none

29. A range of values is

(a) a point estimate (b) an interval estimate (c) both (d) none

30. If we do not have any knowledge of population variance, then we have to estimate it from the

(a) frequency (b) sample data (c) distribution (d) none

31. The sample standard deviation may be a good estimate for population standard deviation in case of ———————— samples.

(a) small (b) moderately sized
(c) large (d) none

32. The sample standard deviation is a biased estimator of population standard deviation in case of ———————— samples.

(a) small (b) moderately sized
(c) large (d) none

33. If the expected value of the estimator is the value of the parameter of estimation then a good estimator shall be

(a) biased (b) unbiased (c) both (d) none

34. The difference between sample S.D and the estimate of population S.D is negligible if the sample size is

(a) small (b) moderate (c) sufficiently large (d) none

35. Finite population multiplier is

(a) square root of (N –1)/ (N –n) (b) square root of (N –n)/ (N –1)
(c) square of (N –1)/ (N –n) (d) square of (N –n)/ (N –1)

36. Sampling fraction is

(a) n/N (b) N/n (c) (n + 1)/N (d) (N + 1)/n

37. The standard error of the mean for finite population is very close to the standard error of the mean for infinite population when the sampling fraction is

(a) small (b) large (c) moderate (d) none

38. The finite population multiplier is ignored when the sampling fraction is

(a) greater than 0.05 (b) less than 0.5 (c) less than 0.05 (d) greater than 0.5

39. The ———— that we associate with an interval estimate is called the confidence level.

(a) probability (b) statistics (c) both (d) none

40. The higher the probability the ———————— is the confidence.

(a) moderate (b) less (c) more (d) none

41. The most commonly used confidence levels are

(a) greater than and equal to 90% (b) less than 90%
(c) greater than 90% (d) less than and equal to 90%

42. The confidence limits are the upper & lower limits of the

(a) point estimate (b) interval estimate
(c) confidence interval (d) none

43. We use t- distributions when the sample size is

(a) big (b) small (c) moderate (d) none

44. We use t- distributions when samples are drawn from the ———————— population.

(a) normal (b) Binomial (c) Poisson (d) none

45. For 2 sample values, we have ———— degree of freedom.

(a) 2 (b) 1 (c) 3 (d) 4

46. For 5 sample values, we have ———— degree of freedom.

(a) 5 (b) 3 (c) 4 (d) none

47. The ratio of the no. of elements possessing a characteristic to the total no. of elements in the population is known as

(a) population proportion (b) population size
(c) both (d) none

48. The ratio of the no. of elements possessing a characteristic to the total no. of elements in a sample is known as

(a) characteristic proportion (b) sample proportion
(c) both (d) none

49. The mean of the sampling distribution of sample proportion is ———————— the population proportion.

(a) greater than (b) less than (c) equal to (d) none

50. For ———————— samples , the sample proportion is an unbiased estimate of the population proportion.

(a) large (b) small (c) moderate (d) none

51. The finite population correction factors should be used when the population is

(a) infinite (b) finite & large (c) finite & small (d) none

52. Which would you prefer for ——" The universe is large"

(a) Full enumeration (b) sampling (c) both (d) none

53. Which would you prefer for ——"The Statistical inquiry is in depth"

(a) Full enumeration (b) sampling (c) both (d) none

54. Which would you prefer for ——"Where testing destroys the quality of the product"

(a) Full enumeration (b) sampling (c) both (d) none

55. In Hypothesis Testing when H_0 is true, it is called

(a) Type I error (b) Type II error (c) Type III error (d) Type IV error

56. P (type I error) means

(a) P (accepting H_0 when H_1 is true) (b) P (rejection of H_0 when H_0 is true)
(c) P (accepting H_0 when H_0 is true) (d) P (rejection of H_0 when $H_1$is true)

57. The procedures for determining the sample size for estimating a population proportion are similar to those of estimating a population mean. In this case we must know ———— —— facto Rs.

(a) 2 (b) 5 (c) 4 (d) 3

58. In determining the sample size for estimating a population mean , the no. of factors must be known is

(a) 2 (b) 3 (c) 5 (d) 4

59. In audit test Statistical Sampling methods are used.

(a) true (b) false (c) both (d) none

60. In cost accounting operation Statistical Sampling methods are used.

(a) true (b) false (c) both (d) none

61. The difference between the estimate from the sample and the parameter to be estimated is

(a) sampling error (b) permissible sampling error
(c) confidence level (d) none

62. The estimated true proportion of success is required to determine sample size for

(a) estimating a mean (b) estimating a proportion
(c) both (d) none

63. The standard deviation is required to determine sample size for

(a) estimating a mean (b) estimating a proportion
(c) both (d) none

64. The desired confidence level is required to determine sample size for

(a) estimating a mean (b) estimating a proportion
(c) both (d) none

65. The permissible sampling error is required to determine sample size for

(a) estimating a mean (b) estimating a proportion
(c) both (d) none

66. In Control of book keeping and clerical errors Statistical sampling methods are used.

(a) true (b) false (c)both (d) none

67. The Exploratory sampling is known as

(a) Estimation sampling (b) Acceptance sampling
(c) Discovery sampling (d) none

68. Single, double, multiple and sequential are several types of

(a) Discovery sampling method (b) Acceptance sampling method
(c) both (d) none

69. Standard deviation of a sampling distribution is itself the standard error.

(a) true (b) false (c) both (d) none

70. Sampling error increases with an increase in the size of the sample.

(a) true (b) false (c) both (d) none

71. Deliberate sampling is free from bias.

(a) True (b) false (c) both (d) none

72. Which would you prefer ————A higher degree of confidence is desired.

(a) Larger Sample (b) Small sample (c) both (d) none

73. Which would you prefer ———— Previous experience reveals a low rate of error.

(a) Larger Sample (b) Small sample (c) both (d) none

74. Testing the assumption that an assumed population is located at a known level of significance is known as

(a) confidence testing (b) point estimation
(c) interval estimation (d) hypothesis testing

75. Purposive selection is resorted to in case of judgment sampling

(a) True (b) false (c) both (d) none

76. In test for means of Paired data, if the computed value is _____ than the table value the difference is considered significant.

(a) lesser (b) greater (c) moderate (d) none

77. Cluster sampling is ideal in case the data are widely scattered.

(a) True (b) false (c) both (d) none

78. Stratified random sampling is appropriate when the universe is not homogeneous

(a) True (b) false (c) both (d) none

79. Sampling error increases with an increase in the size of the sample

(a) True (b) false (c) both (d) none

80. Standard deviation of a sampling distribution is it self the standard error.

(a) True (b) false (c) both (d) none

81. The magnitude of standard error increase both by absolute and relative size of the sample.

(a) True (b) false (c) both (d) none

82. In stratified sampling, the sampling is subdivided into several parts, called

(a) strata (b) strati (c) start (d) none

83. The no. of types of random sampling is

(a) 2 (b) 1 (c) 3 (d) 4

84. Random numbers are also called Random sampling number Rs.

(a) True (b) false (c) both (d) none

85. Sample mean is an example of

(a) parameter (b) statistic (c) both (d) none

86. Population mean is an example of

(a) parameter (b) statistic (c) both (d) none

87. Large sample is that sample whose size is

(a) greater than 30 (b) greater than or equal to 30
(c) less than 30 (d) less than or equal to 30

88. Standard error of mean may be defined as the standard deviation in the sampling distribution of

(a) mean (b) median (c) mode (d) none

89. If random sampling with replacement is applied, then the mean of sample means will be _____ the population mean

(a) greater than (b) less than (c) exactly equal to (d) none

90. The sample proportion is taken as an estimate of the population proportion of defectives

(a) True (b) false (c) both (d) none

91. The main object of sampling is to state the limits of accuracy of estimates base on samples

(a) yes (b) no (c) both (d) none

92. The sample is a selected part of the

(a) estimation (b) population (c) both (d) none

93. The ways of selecting a sample are

(a) Random sampling (b) multi – stage sampling
(c) both (d) none

94. _____ sampling is the most appropriate in cases when the population is more or less homogeneous with respect to the characteristic under study

(a) Multi – stage (b) Stratified (c) Random (d) none

95. Random sampling is called lottery sampling

(a) True (b) false (c) both (d) none

96. _____ sampling is absolutely free from the influence of human bias

(a) multi – stage (b) Random (c) purposive (d) none

97. The standard deviation in the sampling deviation is called

(a) standard error (b) Absolute error
(c) relative error (d) none of the statistic

98. Standard error is used to set confidence limits for population parameter and in tests of significance

(a) True (b) false (c) both (d) none

99. In _____ estimation, the estimate is given by a single quantity

(a) Interval (b) Point (c) both (d) none

100. The estimate of the parameter is stated as on interval with a specified degree of

(a) confidence (b) interval (c) class (d) none

101. The interval bounded by upper and lower limits is known as

(a) estimate interval (b) confidence interval
(c) point interval (d) none

102. Statistical hypothesis is an

(a) error (b) assumption (c) both (d) none

103. A die was thrown 400 times and 'six' resulted 80 times then observed value of proportion is

(a) 0.4 (b) 0.2 (c) 5 (d) none

104. In a sample of 400 parts manufactured by a factory, the no. of defective parts was found to be 30. The observed value is

(a) $\frac{7}{60}$ (b) $\frac{3}{40}$ (c) $\frac{40}{3}$ (d) $\frac{60}{7}$

105. If S. D.= 20 and sample size is 100 then standard error of mean is

(a) 2 (b) 5 (c) $\frac{1}{5}$ (d) none

ANSWERS

1	(a)	2	(b)	3	(a)	4	(b)	5	(a)
6	(a)	7	(b)	8	(c)	9	(b)	10	(c)
11	(d)	12	(c)	13	(a)	14	(c)	15	(a)
16	(d)	17	(a)	18	(b)	19	(d)	20	(a)
21	(b)	22	(c)	23	(c)	24	(b)	25	(c)
26	(a)	27	(b)	28	(a)	29	(b)	30	(b)
31	(c)	32	(b)	33	(b)	34	(c)	35	(b)
36	(a)	37	(a)	38	(c)	39	(a)	40	(c)
41	(a)	42	(c)	43	(b)	44	(a)	45	(b)
46	(c)	47	(a)	48	(b)	49	(c)	50	(a)
51	(c)	52	(b)	53	(b)	54	(b)	55	(a)
56	(b)	57	(d)	58	(b)	59	(a)	60	(a)
61	(b)	62	(b)	63	(a)	64	(c)	65	(c)
66	(a)	67	(c)	68	(b)	69	(a)	70	(b)
71	(b)	72	(c)	73	(b)	74	(d)	75	(a)
76	(b)	77	(b)	78	(b)	79	(b)	80	(a)
81	(a)	82	(a)	83	(a)	84	(a)	85	(b)
86	(a)	87	(b)	88	(a)	89	(c)	90	(a)
91	(a)	92	(b)	93	(c)	94	(c)	95	(a)
96	(b)	97	(a)	98	(a)	99	(b)	100	(a)
101	(b)	102	(b)	103	(b)	104	(b)	105	(a)