CHAPTER SEVEN

7 Sampling and Sampling Distribution of Sample Statistic

7.1 Sampling and Sampling Distribution

- **Introduction**

Sampling is simply the process of learning about the population on the basis of a sample drawn from it. Thus, in the sampling technique instead of every unit of the universe only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe.

- **Motivating Example**

Although much of the development in the theory of sampling has taken place only in recent years, the idea of sampling is pretty old. Since times immemorial people have examined a handful of grains to ascertain the quality of the entire lot. A house wife examines only two or three grains of boiling rice to know whether the pot of rice is ready or not. A doctor examines a few drops of blood and draws conclusion about the blood constitution of the whole body. A businessman places orders for material examining only a small sample of the same. A teacher may put questions to one or two students and find out whether the class as a whole is following the lesson. In fact, there is hardly any field where the technique of sampling is not used either consciously or unconsciously.

It should be noted that a sample is not studied for its own sake. The basic objective of its study is to draw inference about the population. In other words, sampling is a tool which helps to know the characteristics of the universe or population by examining only a small part of it. The values obtained from the study of a sample, such as the average and dispersion, are known as 'statistic'. On the other hand, such values for population are called 'parameters.

When secondary data are not available for the problem under study, a decision may be taken to collect primary data by using any of the methods discussed in chapter two. The required information may be obtained by following either the census method or the sample method.

What is the Difference between Census and Sample Method?

- **Census and Sample Method**

Under the census or complete enumeration survey method, data are collected for each and every unit (person, household, field, shop, factory etc.), as the case may be of the population or universe, which is the complete set of items, which are of interest in any particular situation.

The merits of the census method are

1. Data are obtained from each and every unit of the population.
2. The results obtained are likely to be more representative, accurate and reliable.
3. It is an appropriate method of obtaining information on rare events such as areas under some crops and yield thereof, the number of persons of certain age groups, their distribution by sex, educational level of people, etc. This is the reason why throughout the world the population data are obtained by conducting a census generally every 10 years by the census method.
4. Data of complete enumeration census can be widely used as a basis for various surveys.

Demerits

However, despite these advantages the census method is not very popularly used in practice.

1. The effort, money and time required for carrying out complete enumeration will generally be very large and in many cases cost may be so prohibitive that the very idea of collecting information may have to be dropped. This is truer of underdeveloped countries where resources constitute a big constraint.

2. Also, if the population is infinite or the evaluation process destroys the population unit, the method cannot be adopted.

What is sampling?

Sampling is a method used in statistical analysis in which a decided number of considerations are taken from a comprehensive population or a sample survey. Thus, in the sampling technique instead of every unit of the universe only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe. A sample is a subset of population units.

- The process of sampling involves three elements:

  a. Selecting the sample.
  b. Collecting the information, and
  c. Making an inference about the population.

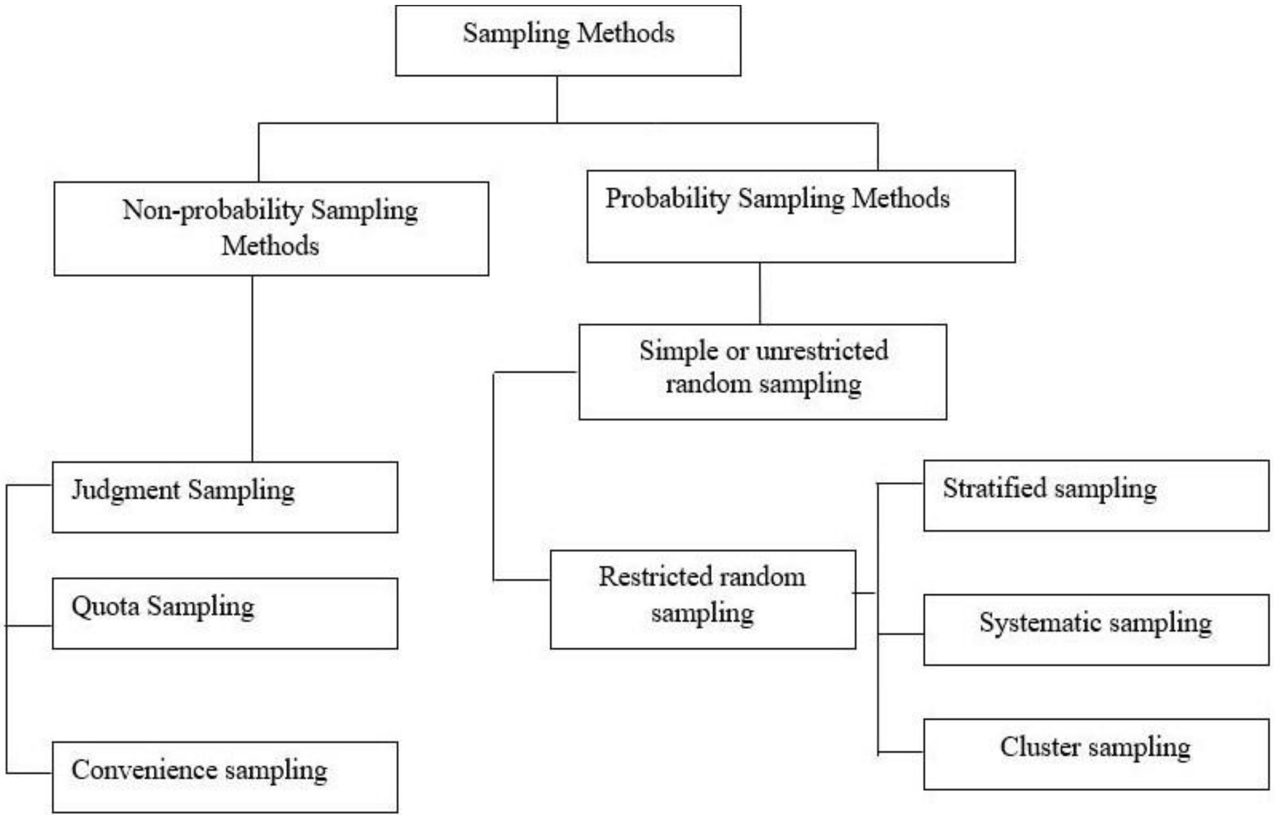
The three elements cannot generally be considered in isolation from one another. Sample selection, data collection and estimation are all interwoven and each has an impact on the others. Sampling is not haphazard selection-it embodies definite rules for selecting the sample.

7.1.1 Methods of Sampling

- The various methods of sampling can be grouped under two broad heads these are:

1. **Probability sampling** (also known as random sampling) and

2. **non-probability** (or non-random) sampling.



1. Probability (Random) Sampling

Probability sampling methods are those in which every item in the universe (population) has a known chance, or probability, of being chosen for sample. This implies that the selection of sample items is independent of the person making the study-that is, the sampling operation is controlled so objectively that the items will be chosen strictly at random. It may be noted that the term random sample is not used to describe the data in the sample but the process employed to select the sample. Randomness is thus a property of the sampling procedure instead of an individual sample.

❖ **Advantages of Probability Sampling**

The following are the basic advantages of probability sampling methods:

1. Probability sampling does not depend upon the existence of detailed information about the universe for its effectiveness.

2. Probability sampling provides estimates which are essentially unbiased and have measurable precision
3. It is possible to evaluate the relative efficiency of various sample designs only when probability sampling is used.

- **Limitations of Probability Sampling**

The limitations are;

1. Probability sampling requires a very high level of skill and experience for its use
2. It requires a lot of time to plan and execute a probability sample.
3. The costs involved in probability sampling are generally large as compared to non-probability sampling.

a) Simple or Unrestricted Random Sampling

Simple random sampling refers to that sampling technique in which each and every unit of the population has an equal opportunity of being selected in the sample. In simple random sampling which items get selected in the sample is just a matter of chance and personal bias of the investigator does not influence the selection. It should be noted that the word random does not mean 'haphazard' or 'hit-or-miss'-it rather means that the selection process is such that chance only determines which items shall be included in the sample.

- All n items of the sample are selected independently of one another and all N items in the population have the same chance of being included in the sample. By independence of selection, we mean that the selection of a particular item in one draw has no influence on the probabilities of selection in any other draw.
- All the possible samples of a given size n are equally likely to be selected.

What Methods to Ensure Randomness?

To ensure randomness of selection one may adopt either

1. The Lottery Method or
2. Consult table of random numbers.

Lottery Method: This is a very popular method of taking a random sample. Under this method, all items of the universe are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container or drum. A blindfold selection then made of the number of slips required to constitute the desired sample size. The selection of items thus depends entirely on chance.

Table of Random Numbers: The lottery method discussed above becomes quite cumbersome as the size of population increases. An alternative method of random selection is that of using the table of random numbers.

The random numbers are generally obtained by some mechanism which, when repeated a large number of times, ensures approximately equal frequencies for the numbers from 0 to 9 and also proper frequencies for various combinations of numbers (such as 00, 01, 999, etc.) that could be expected in a random sequence of the digits 0 to 9.

❖ **Merits of Simple Random Sampling**

- Simple random sampling method has the following advantages:

- Since the selection of items in the sample depends entirely on change there is no possibility of personal bias affecting the results.
- As compared to judgment sampling a random sample represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population.

- The analyst can easily assess the accuracy of this estimate because sampling errors follow the principles of chance. The theory of random sampling is further developed than that of any other type of sampling, which enables the analyst to provide the most reliable information at the least cost.

Demerits: This method is however associated with following limitations:

- The use of simple random sampling necessitates a completely catalogued universe from which to draw the sample. But it is often difficult for the investigator to have up-to-date lists of all the items of the population to be sampled. This restricts the use of this method in economic and business data where very often we have to employ restricted random sampling designs.
- The size of the sample required to ensure statistical reliability is usually larger under simple random sampling than stratified sampling.
- From the point of view of field survey, it has been claimed that cases selected by simple random sampling tend to be too widely dispersed geographically and that the time and cost of collecting data become too large.

Note: *let* $N = \text{population size}$, $n = \text{sample size}$.

Suppose simple random sampling is used

- We have N^n possible samples if sampling is **with replacement**.
- We have $\binom{N}{n}$ possible samples if sampling is **without replacement**.

b) Restricted random sampling:

i. Systematic Sampling

A systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals until the sample has been formed. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available. The list may be prepared in alphabetical, geographical, numerical or some other order. The items are serially numbered. The first item is selected at random generally by following the Lottery method. Subsequent items are selected by taking every k^{th} item from the list where '*k* refers to the sampling interval or sampling ratio, *i.e.,* the ratio of population size to the size of the sample. Symbolically:

$$k = \frac{N}{n}$$

Where *k* = Sampling interval, N = Universe size, and *n* = Sample size.

This method of sampling is also known as quasi-random sampling method. This is because once the initial starting point is determined; the remainder of the items selected for the sample is pre-determined by the sampling interval.

While calculating *k,* it is possible that we get a fractional value. In such a case we should use approximation procedure, *i.e.,* if the fraction is less than 0.5 it should be omitted and if it is more than 0.5 it should be taken as 1. If it is exactly 0.5 it should be omitted, if the number is even and should be taken as 1, if the number is odd. This is based on the principle that the number after approximation should preferably be even. For example, if the number of students is respectively 1,020, 1,150 and 1,100 and we want to take a sample of 200, *k* shall be:

(i) $k = \frac{1020}{200}$ =5.1 or 5

(ii) $k = \frac{1150}{200}$ =5.75 or 6

(iii) $k = \frac{1100}{200}$ =5.5 or 6

Illustration

In a class there are 96 students with Roll No. from 1 to 96. It is desired to take sample of 10 students. Use the systematic sampling method to determine the sample size.

Solution

$k = \frac{N}{n} = \frac{96}{10}$ =9.6 or 10

From 1 to 96 Roll Nos. the first students between 1 and *k, i.e.,* 1 and 10, will be selected at random and then we will go on taking every k^{th} student. Suppose the first student comes out to be 4^{th}. The sample would then consist of the following Roll No's. 4, 14, 24, 34, 44, 54, 64, 74, 84, 94

❖ **Merits of Systematic Sampling**

1. The systematic sampling design is simple and convenient to adopt.
2. The time and work involved in sampling by this method are relatively less.
3. If populations are sufficiently large, systematic sampling can often be expected to yield results similar to those obtained by proportional stratified sampling.

❖ **Demerits of Systematic Sampling**

- The main limitation of the method is that it becomes less representative if we are dealing with populations having "hidden periodicities".
- Also, if the population is ordered in a systematic way with respect to the characteristics the investigator is interested in, then it is possible that only certain types of items will be included in the population, or at least more of certain types than others. For instance, in a

study of workers' wages the list may be such that every tenth worker on the list gets wages above 750 birr per month.

ii. Stratified Sampling

Stratified random sampling or simply stratified sampling is one of the random methods, which, by using the available information concerning the population, attempts to design a more efficient ban obtained by the simple random procedure.

While applying stratified random sampling technique, the procedure is given below:

a. The universe to be sampled is sub-divided (or stratified) into groups **(Strata** *plural for stratum***)**, which are mutually exclusive and include all items in the universe.
b. A simple random sample is then chosen independently from each group (**stratum**).

This sampling procedure differs from simple random sampling in that the sample items are chosen at random from the entire universe. In stratified random sampling the sampling is designed so that a designated number of items is chosen from each stratum. In simple random sampling the distribution of the sample among strata is left entirely to chance.

How to Select Stratified Random Sample?

- Some of the issues involved in setting up a stratified random sample are:

Basis of Stratification: What characteristic should be used to sub divide the universe into different strata? As a general rule, strata are created on the basis of a variable known to be correlated with the variable of interest and for which information on each universe element is known. Strata should be constructed in a way, which will minimize differences among sampling units within strata, and maximize difference among strata.

For example, if we are interested in studying the consumption pattern people of Addis Ababa, the city of Addis Ababa may be divided into various parts (such as zones or weredas) and from each part a sample may be taken at random. Before deciding on stratification, we must have knowledge of the traits of the population. Such knowledge may be based upon expert judgment, past data, preliminary observations from pilot studies, etc.

The purpose of stratification is to increase the efficiency of sampling by dividing a heterogeneous universe in such a way that *(i)* there is as great a homogeneity as possible within each stratum, and *(ii)* a marked difference is possible between the strata.

Number of Strata: How many strata should be constructed? The considerations limit the number of strata that is feasible; costs of adding more strata may soon outrun benefits. As a generalization more than six strata may be undesirable.

Sample size within Strata: How many observations should be taken from each stratum? When deciding this question, we can use either a proportional or a disproportional allocation. In proportional allocation, one samples each stratum in proportion to its relative weight. In disproportional allocation this is not the case. It may be pointed out that proportional allocation approach is simple and if all one knows about each stratum is the number of items in that stratum; it is generally also the preferred procedure. In disproportional sampling, the different strata are sampled at different rates. As a general rule when variability among observations within a stratum is high, one samples that stratum at a higher rate than for strata with less internal variation.

❖ **Merits of Stratified Sampling**

- Stratified sampling methods have the following advantages:
  - **More representative:** Since the population is first divided into various strata and then a sample is drawn from each stratum there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus

secured. C.J. Grohmann has rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.

- **Greater accuracy:** Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items.
- **Greater geographical concentration:** As compared with random sample, stratified samples can be more concentrated geographically, *i.e.,* the units from the different strata may be selected in such a way that all of them are localized in one geographical area. This would greatly reduce the time and expenses of interviewing.

❖ **Demerits of Stratified Sampling**

- Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous items as otherwise the results may not be reliable. If proper stratification of the population is not done, the sample may have the effect of bias.
- The items from each stratum should be selected at random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.
- Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random sample cost per observation may be quite high.

iii. Cluster (Multi-Stage) Sampling

Under this method the random selection is made of primary, intermediate and final (or the ultimate) units from a given population or stratum. There are several stages in which the sampling process is carried out. At first, the first stage units are sampled by some suitable method, such as simple random sampling. Then, a sample of second stage units is selected from each of the selected first stage units, again by some suitable method, which may be the same as, or different from the method employed for the first stage units. Further stages may be added as required. The procedure may be illustrated as follows:

Suppose we want to take a sample of 5,000 households from Addis Ababa. At the first stage, the State may be divided into a number of districts and a few districts selected at random. At the second stage, each district may be subdivided into a number of villages and a sample of villages may be taken at random. At the third stage, a number of households may be selected from each of the villages selected at the second stage. To take another example suppose in a particular survey, we wish to take a sample of 10.000 students from Arba Minch University. We may take department-primary units-as the first stage, then draw departments as the second stage, and choose students as the third and last stage.

❖ **Merits of Cluster Sampling**

1. Multi-stage sampling introduces flexibility in the sampling method, which is lacking in the other methods.
2. It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the fieldwork to be concentrated and yet large area to be covered.
3. Another advantage of the method is that subdivision into second stage units *(i.e.,* the construction of the second stage frame) need be carried out for only those first stage units, which are included in the sample. It is, therefore, particularly valuable in surveys of underdeveloped areas where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

❖ **Demerit of Cluster Sampling**

However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units, which have been selected by some single stage process. We have discussed above the various random procedures in independent designs. In practice we often combine two or more of these methods into a single design.

2. **Non-probability Sampling Methods**

Non-random sampling is a process of sample selection without the use of randomization. In other words, a non-random sample is selected basis other than the probability consideration. Such as convenience, judgment, etc,

The most important difference between random and non-random sampling is that the pattern of sampling variability can be ascertained in case of random sampling. Whereas in non-random sampling, there is no way of knowing the patterns of variability in the process.

i. Judgment Sampling

In this method of sampling the choice of sample items depends exclusively on the judgment of the investigator. In other words, the investigator exercises his judgment in the choice and includes those items -in the sample which he thinks are most typical of the universe with regard to the characteristics under investigation. For example. if sample of ten students is to be selected from a class of sixty for analyzing the spending habits of students, the investigator would select 10 students who, in his opinion, are representative of the class.

ii. Quota Sampling

Quota sampling is a type of Judgment sampling and is perhaps the commonly used sampling technique in non-probability category. In a quota sample quota are set up according to some specified characteristics such as so many in each of several income groups, so many in each age, so many with certain political or religious affiliations, and so on. Each interviewer is then told to interview a certain number of persons, which constitute his quota. Within the quota, the selection of sample items depends on personal judgment. For example, in a radio listening survey, the interviewers may be told interview 500 people living in a certain area and that out of every 100 persons interviewed 60 are to be housewives, 25 farmers and 15 children above the age of 15. Within these quotas the interviewer is free to select the people to be interviewed. Because of the risk of personal prejudice and bias entering the process of selection, the quota sampling is not widely used in practical work.

Quota sampling and stratified random sampling are similar in as much both methods the universe is divided into parts and the total is allocated among the parts. However, the two procedures diverge radically. In stratified random sampling the sample with each stratum is chosen at random. In quota sampling, the sampling within each cell is not done at random; the field representatives are given wide latitude in selection of respondents to meet their quotas.

Quota sampling is often used in public opinion studies. It occasionally provides satisfactory results if the interviewers are carefully trained and if they follow their instructions closely. It is often found that since the choices of respondents within a cell is left to the field representatives; the more accessible and articulate people within a cell will usually be the ones who are interviewed.

iii. Convenience Sampling

A convenience sample is obtained by selecting ‘convenient’ population units. The method of convenience sampling is also called the **chunk**. A chunk refers to that fraction of the population being investigated which is selected neither by probability nor by judgment but by convenience. A sample obtained from readily available lists such as automobile registrations; telephone directories, etc., is a convenience sample and not a random sample even if the sample is drawn at random from the lists. If a person is to submit a project report on labor-management relations in textile industry and he takes a textile mill close to his office and interviews some people over there, he is following the convenience sampling method. Convenience samples are prone to bias by their very nature-selecting population elements that are convenient to choose almost always make them special or different from the best of the elements in the population in some way.

Hence the results obtained by following convenience sampling method can hardly be representative of the population-they are generally biased and unsatisfactory. However, convenience sampling is often used for making pilot studies. Questions may be tested and the chunk may obtain preliminary information before the final sampling design is decided upon.

7.1.2 Errors in sample survey:

- There are two types of errors these are **sampling error** and **non- sampling error**.

a) Sampling error:

It refers to the differences between the sample estimate and the actual value of the characteristics of the population. It may arise due to inappropriate sampling techniques applied.

Sampling errors can be of two types.

- **Biased error:** An error that arises on account of some biases or imbalances on the part of the investigators, informants, or instruments of counting, measuring, or experimenting.
- **Unbiased error:** An error that does not take place on account of any bias with anybody but occurs accidentally may be due to a chance or due to an arithmetic error. Such errors arise automatically without any motive.

- The magnitude of sampling error can be reduced by taking a larger sample.

b) Non-sampling error

These errors occur in acquiring, recording, or tabulating statistical data. These are more serious than sampling errors because a sampling error can be minimized by taking a larger sample. However, a non-sampling error cannot be minimized even by taking a larger sample.

Such errors are errors due to procedure bias such as:

- Due to incorrect responses

- Measurement
- Errors at different stages in processing the data.

❖ **The Needs for Sampling**

- Reduced cost
- Greater speed
- Greater accuracy
- Greater scope
- More detailed information can be obtained.

7.1.3 Size of Sample

An important decision that has to be taken in adopting a sampling technique is about the size of the sample. Size of sample means the number of sampling units selected from the population for investigation. Experts have expressed different opinions on this point. For example, some have suggested that the sample size should be 5 per cent of the size of population while others are of the opinion that sample size should be at least 10 per cent. However, these views are of little use in practice because no hard and fast rule can be laid down that sample size should be 5 per cent, 10 per cent or 25 per cent of the universe size. It may point out that more size alone does not ensure representativeness. A smaller sample, but well selected sample, may be superior to a larger but badly *selected* sample. If the size of the sample is small it may not represent the universe and the inference drawn about the population may be misleading. On the other hand, if the size of sample is very large, it may be too burdensome financially, require a lot of time and may have serious problems of managing it. Hence the sample size should neither be too small nor too large. It should be 'optimum'. Optimum size, according to Parten, is one that fulfils the requirements of ***efficiency***, ***representativeness***, ***reliability*** and ***flexibility.***

What is the Factors Need to be Considered While Deciding about the Sample Size?

The following factors should be considered while deciding the sample size:

i. **The size of the Universe:** The larger the size of the universe, the bigger should be the sample size.

i. **The Resources Available:** If the resources available are vast a larger sample size could be taken. However, in most cases resources constitute a big constraint on sample size.

ii. **The Degree of Accuracy or Precision Desired:** The greater the degree of accuracy desired, the larger should be the sample size. However, it does not necessarily mean that bigger samples always ensure greater accuracy. If a sample is selected by experts by following scientific method, it may ensure better results even when it is small compared to a situation in which large sample size is selected by inexperienced people.

iii. **Homogeneity or Heterogeneity of the Universe:** If the universe consists of homogeneous units a small sample may serve the purpose but if the universe consists of heterogeneous units a large sample may be inevitable.

iv. **Nature of Study:** For an intensive and continuous study a small sample may be suitable. But for studies, which are not likely to be repeated and are quite extensive in nature, it may be necessary to take a larger sample size.

v. **Method of Sampling Adopted:** The size of sample is also influenced by the type of sampling plan adopted. For example, if the sample is a simple random sample, it may necessitate a bigger sample size. However, in a properly drawn stratified sampling plan, even a small sample may give better results.

vi. **Nature of Respondents:** Where it is expected a large number of respondents will not co-operate and send back the questionnaire, a large sample should be selected.

The above factors have to be properly weighted before arriving at the sample size. However, the selection of optimum sample size is not that simple as it might seem to be. If the sample is used which is larger than necessary, resources are wasted, if the sample is smaller than required the objectives of the analysis may not be achieved.

❖ **Determination of Sample Size**

A number of formulae have been devised for determining the sample size depending upon the availability of information. As an example, a few formulae are given below:

$$n=\left(\frac{zs}{d}\right)^2$$, Where

n = Sample size

z= Value at a specified level of confidence or desired degree of precision.

s= Standard deviation of the population

d= Difference between population mean and sample mean.

The steps in computing the sample size from the above formula are:

i. Select the desired degree of precision, *i.e.,* specified level of confidence and designates it as small 'z' (at 1% level of significance or 99% confidence level the value of 'z' is 2.58, and at 5% level of significance or 95% confidence level 1.96).

ii. Multiply the 'z' selected in step 1 by the standard deviation of the universe, which may be assumed.

iii. Divide the product of the preceding step by the standard error of mean or difference between population and sample mean. Square the resultant quotient. The result is the size of sample required.

Example 7.1

Determine the sample size if s = 6, population mean = 25, sample mean = 23 and the desired degree of precision is 99 per cent.

n=$\left(\frac{ZS}{d}\right)^2$, s = 6, d = 25-23 = 2, z = 2.576 (at 1% level the z value is 2.576)

Substituting the values: $n = \left[\frac{2.576 \times 6}{2}\right]^2 = 7.728^2 = 59.72 \; or \; 60$

Similarly, the sample size can be determined from the formula for determining the standard error of mean, *i.e.*, $s_x = \frac{1}{\sqrt{n}} = s^2{}_x = \frac{s^2}{n}$.

If s is 10 and S_x = 2.25, n shall be

$n = \left(\frac{10}{2.25}\right)^2 = (4)^2 = 16$

- **Sampling Distribution**

We discussed several different measures of central tendency and variability in Chapter 3 and 4 and distinguished between numerical descriptive measures of a population (parameters) and numerical descriptive measures of a sample (statistics). Thus, μ and σ are parameters, whereas $\bar{x}$ and s are statistics.

The numerical value of a sample statistic cannot be predicted exactly in advance. Even if we knew that a population mean μ was $216.37 and that the population standard deviation σ was $32.90—even if we knew the complete population distribution—we could not say that the sample mean $\bar{x}$ would be exactly equal to $216.37. A sample statistic is a random variable; it is subject to random variation because it is based on a random sample of measurements selected from the population of interest. Also, like any other random variable, a sample statistic has a probability distribution. We call the probability distribution of a sample statistic the **sampling *distribution*** of that statistic.

- The sampling distribution of a statistic depends on the distribution of the population, the size of the sample and the method of sample selection.

7.2 Sampling Distribution of Sample Statistic

Sampling distribution: the sampling distribution of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n. It is a theoretical idea we do not actually build it.

Basically,

- Sampling distribution of ***sample mean*** $(\bar{x})$.
- Sampling distribution of ***sample proportion*** $(\hat{p})$.
- Sampling distribution of ***sample variance*** (s^2). But in this course, we will consider only sampling distribution for sample mean.

7.2.1 Sampling distribution of Sample a Mean

In addition to knowing how individual data values vary about the mean for a population, researchers are interested in knowing how the means of samples of the same size taken from the same population vary about the population mean.

Suppose a researcher selects a sample of 30 adult males and finds the mean of the measure of the triglyceride levels for the sample subjects to be 187 milligrams/deciliter. Then suppose a second sample is selected, and the mean of that sample is found to be 192 milligrams/deciliter. Continue the process for 100 samples. What happens then is that the mean becomes a random variable, and the sample means 187, 192, 184. . . 196 constitute a ***sampling distribution of sample means.***

Definition

A **sampling distribution of sample means** $\bar{x}$ is a probability distribution using the means computed from all possible random samples of a specific size n taken from a population with mean μ and standard deviation σ.

If the samples are randomly selected with replacement, the sample means, for the most part, will be somewhat different from the population mean μ. These differences are caused by ***sampling error***.

There are commonly three properties of interest of a given sampling distribution.

- Its Mean
- Its Variance
- Its functional form.

❖ **Steps for the construction of Sampling Distribution of the sample mean**

Step 1. From a finite population of size *N,* randomly draw all possible samples of size *n.*

Step 2. Calculate the mean for each sample $\bar{x}$.

Step 3. Summarize the mean obtained in step 2 in terms of frequency distribution or relative frequency distribution.

Note: once a particular sample is obtained, it cannot be obtained a second time.

When all possible samples of a specific size n are selected with replacement from a population, the distribution of the sample means for a variable has two important properties, these properties are:

I. The mean of the sample means will be the same as the population mean.

$$[\,i.e.\;\; \mu_{\bar{x}} = \mu]$$

II. The standard deviation of the sample means will be smaller than the standard deviation of the population, and it will be equal to the population standard deviation divided by the square root of the sample size.

$$[\,i.e.\;\; \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}]$$

- **The following examples illustrate these two properties.**

Example 7.2

Suppose an instructor gave an 8-point quiz to a small class of four students. The results of the quiz were 2, 6, 4, and 8. For the sake of discussion, assume that the four students constitute the population. Hence the mean of the population will be:

$$\mu = \frac{2+6+4+8}{4} = 5$$

And the population standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{4}(x_i-\mu)^2}{\sum_{i=1}^{n} f_i}} = \sqrt{\frac{(2-5)^2+(6-5)^2+(4-5)^2+(8-5)^2}{4}} \approx 2.236$$

Now, if all samples of size 2 are taken with replacement and the mean of each sample is found, the distribution is as shown below.

Sample	Mean **($\bar{x}$)**	Sample	Mean **($\bar{x}$)**
2,2	2	6,2	4
2,4	3	6,4	5
2,6	4	6,6	6
2,8	5	6,8	7
4,2	3	8,2	5
4,4	4	8,4	6
4,6	5	8,6	7
4,8	6	8,8	8

A frequency distribution of sample means is as follows.

$\bar{x}$	F
2	1
3	2
4	3
5	4
6	3
7	2
8	1

This implies the mean of the sample mean $\mu_{\bar{x}}$ will be:

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{n}\bar{x}_i f_i}{\sum_{i=1}^{n} f_i} = \frac{(2*1)+(3*2)+\cdots+(8*1)}{16} = \frac{80}{16}$$

$$= 5 = population\ meam\ (\mu)$$

$$\Rightarrow \mu_{\bar{x}} = \mu$$

The standard deviation of sample means, denoted by $\sigma_{\bar{x}}$, is

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x}_i-\mu_{\bar{x}})^2}{\sum_{i=1}^{n} f_i}} = \sqrt{\frac{(2-5)^2+(3-5)^2+\cdots+(8-5)^2}{16}} \approx 1.581$$

This is the same as the population standard deviation, divided by $\sqrt{n}$ $(\sqrt{2})$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{\sqrt{2}} \approx 1.581$$

Remark:

1. In general, if sampling is ***with replacement***,

$$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{n} \quad \Rightarrow \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

2. If sampling is ***without replacement***,

$$\sigma_{\bar{x}}{}^2 = \frac{\sigma^2}{n}[\frac{N-n}{N-1}] \quad \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

3. In any case the samples mean $\bar{x}$.is unbiased estimator of the population mean μ.

   i.e. $\mu_{\bar{x}} = \mu \Rightarrow E(\bar{x}) = \mu$

- Sampling may be from a normally distributed population or from a non-normally distributed population.
- When sampling is from a normally distributed population, the distribution of $\bar{x}$ will possess the following property.

i. The distribution of $\bar{x}$ will be normal.

ii. The mean of $\bar{x}$ is equal to the population mean , i.e.$\mu_{\bar{x}} = \mu$

iii. The variance of $\bar{x}$ is equal to the population variance divided by the sample size,

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

iv. The functional form of the sample mean also be;

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

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma^2/_n} \sim N(0, 1)$$

Example 7.3

A population of size 20 is sampled without replacement. The standard deviation of the population is 0.35. We require the standard error of the mean to be no more than 0.15. What is the minimum sample size?

Answer: *Equation 8.6 gives the relationship*

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

In this case σ is 0.35 and *N* is 20. What value of *n* is required if $\sigma_{\bar{x}}$ is at the limiting value of 0.15?

Substituting,

$$0.15 = \frac{0.35}{\sqrt{n}}\sqrt{\frac{20-n}{20-1}}$$

$$\Rightarrow \sqrt{\frac{20-n}{n}} = \frac{0.15\sqrt{19}}{0.35}$$

$$\Rightarrow 20 - n = 3.490\,\boldsymbol{n}$$

$$then\ n = \frac{20}{4.490} = 4.45$$

But the sample size, the number of observations in the sample, must be an *integer*. It must be at least 4.45, so the minimum sample size is 5. A sample size of 4 would not satisfy the requirement.

Example 7.

The standard deviation of measurements of a linear dimension of a mechanical part is 0.14 mm. What sample size is required if the standard error of the mean must be no more than

(a) 0.04 mm,

(b) 0.02 mm?

<u>Soln</u>.

Since the dimension can be measured as many times as desired, the population size is effectively infinite. Then

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(a) For $\sigma_{\bar{x}}$ = 0.04 mm and σ = 0.14 mm,

$$\sqrt{n} = \frac{0.14}{0.04} = 3.50$$

$$\Rightarrow n = 12.25$$

Then for $\sigma_{\bar{x}}$ ≤ 0.04 mm, the minimum sample size is 13.

(b) For $\sigma_{\bar{x}}$ = 0.02 mm and σ = 0.14 mm,

$$\sqrt{n} = \frac{0.14}{0.02} = 7.0$$

$$\Rightarrow n = 49$$

Then for $\sigma_{\bar{x}}$ ≤ 0.02 mm, the minimum sample size is 49.

The standard deviation of the sample means is also called the **standard error of the mean.**

- A third property of the sampling distribution of sample means pertains to the shape of the distribution and is explained by the **central limit theorem.**

The Central Limit Theorem

As the sample size *n* increases without limit, the shape of the distribution of the sample means taken with replacement from a population with mean μ and standard deviation σ will approach a normal distribution. As previously shown, this distribution will have a mean μ and a standard deviation $\frac{\sigma}{\sqrt{n}}$. If the sample size is sufficiently large, the central limit theorem can be used to answer questions about sample means in the same manner that a normal distribution can be used to answer questions about individual values. The only difference here is that a new formula must be used for the *z* values. It is:

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

If a large number of samples of a given size are selected from a normally distributed population, or if a large number of samples of a given size that is greater than or equal to 30 are selected from a population that is not normally distributed, and the sample means are computed, then the distribution of sample means will look like the normal distribution.

- It's important to remember two things when you are using the central limit theorem:

I. When the original variable is normally distributed, the distribution of the sample means will be normally distributed, for any sample size *n.*

II. When the distribution of the original variable is not normal, a sample size of 30 or more is needed to use a normal distribution to approximate the distribution of the sample means. The larger the sample, the better the approximation will be.

- The following examples show you how the standard normal distribution can be used to answer questions about sample means.

Example 7.5

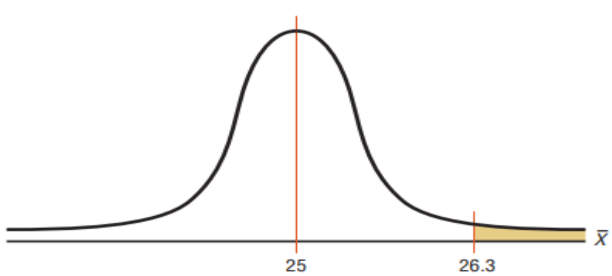
A. C. Nielsen: a research group reported that children between the ages of 2 and 5 watch an average of 25 hours of television per week. Assume the variable is normally distributed and the standard deviation is 3 hours. If 20 children between the ages of 2 and 5 are randomly selected, find the probability that the mean of the number of hours they watch television will be greater than 26.3 hours.

Solution

Since the variable is approximately normally distributed, the distribution of sample means will be approximately normal, with a mean of 25. The standard deviation of the sample means is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{20}} = 0.671$$

Step 1: Draw a normal curve and shade the desired area. The distribution of the means is shown in the Figure below, with the appropriate area shaded.



Step 2: Convert the value to a *z* value. The *z* value is

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{26.3 - 25}{3/\sqrt{20}} = \frac{1.3}{0.671} = 1.94$$

Step 3: Find the corresponding area for the *z* value. The area to the right of 1.94 is 1.000 - 0.9738 = 0.0262 or 2.62%

Step 4: Conclusion

One can conclude that the probability of obtaining a sample mean larger than 26.3 hours is 2.62% [that is, *P* ($\bar{x} > 26.3$) = 0.0262]. Specifically, the probability that the 20 children selected between the ages of 2 and 5 watches more than 26.3 hours of television per week is 2.62%.

Example 7.6

A plant manufactures electric light bulbs with a burning life that is approximately normally distributed with a mean of 1200 hours and a standard deviation of 36 hours. Find the probability that a random sample of 16 bulbs will have a sample mean less than 1180 burning hours. ***(Exercise!)***

Example 7.7

An astronomer wants to measure the distance from her observatory to a distant star. However, due to atmospheric disturbances, any measurement will not yield the exact distance d. As a result, the astronomer has decided to make a series of measurements and then use their average value as an estimate of the actual distance. If the astronomer believes that the values of the successive measurements are independent random variables with a mean of d light years and a standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within ± .5 light years?***(Exercise!)***

Example7.8

If the uric acid values in normal adult males are approximately normally distributed with mean 5.7mgs and a standard deviation 1mg, find the probability that a sample of size 9 will yield a mean;

i. Greater than 6
ii. Between 5 & 6
iii. Less than 5.2

Solution

Let $\bar{x}$ be the amount of uric acid in normal adult males

$\mu = 5.7$, $\sigma = 1$ and n = 9

$\Rightarrow \bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

$\sim N(5.7, \frac{1}{9})$

$\Rightarrow\ Z = \frac{\bar{x}-\mu}{sd} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Then,

i. P (X > 6) =?

$\Rightarrow$ P (X > 6) = P $(\frac{x-\mu}{\sigma/\sqrt{n}} > \frac{6-5.7}{1/\sqrt{9}}) = \mathrm{P}\left(z > \frac{0.3}{0.3333}\right)$

$= \mathrm{P}\,(z > 0.90) = 1 - \mathrm{P}\,(z < 0.90) = 1 - 0.8159 = \mathbf{0.1841}$

ii. P (5 < X < 6) =?

$\Rightarrow$ P (5 < X < 6) = P $(\frac{5-5.7}{1/\sqrt{9}} < \frac{x-\mu}{\sigma/\sqrt{n}} > \frac{6-5.7}{1/\sqrt{9}}) = \mathrm{P}\left(\frac{-0.7}{0.3333} < z < \frac{0.3}{0.3333}\right)$

$= \mathrm{P}\,(-2.10 < z < 0.90) = \mathrm{P}\,(z < 0.90) - \mathrm{P}\,(z < -2.10) = 0.8159 - 0.01786 = \mathbf{0.798}$

iii. P (X < 5.2) =?

$\Rightarrow$ P (X < 5.2) = P $(\frac{x-\mu}{\sigma/\sqrt{n}} < \frac{5.2-5.7}{1/\sqrt{9}}) = \mathrm{P}\left(z < \frac{-0.5}{0.3333}\right) = \mathrm{P}\,(z < -1.50) = \mathbf{0.06681}$