UNIT 1 INTRODUCTION TO SAMPLING DISTRIBUTION

Structure

1.1 INTRODUCTION

In many situations, if we extract some information from all the elements or items of a large population in general, it may be time consuming and expensive. Also, if the elements or items of population are destroyed under investigation then the information gathering from all the elements is not making a sense because all the elements will be destroyed. Therefore, in most of the cases in daily life, business and industry the information is gathered by means of **sampling**. The results of a properly taken sample enable the investigator to arrive at generalisations that are valid for entire population. The process of generalising sample results to the population is called **Statistical Inference.**

For drawing inference about the population parameters, we draw all possible samples of same size and determine a function of sample values, which is called **statistic**, for each sample. The values of statistic are generally varied from one sample to another sample. Therefore, the sample statistic is a random variable and follows a distribution. The distribution of the statistic is called **sampling distribution of the statistic.**

This unit is divided into 8 sections. Section 1.1 is introductive in nature. In Section 1.2, we defined the basic terminology used in statistical inference. The introduction and needs of sampling distributions are described in the Section 1.3. In Section 1.4, we explored the concept of standard error. The most important theorem of Statistics "Central Limit Theorem" is described in Section 1.5. In Section 1.6, we explored the concept of law of large numbers. Unit ends by providing summary of what we have discussed in this unit in Section 1.7 and solution of exercises in Section 1.8.

Objectives

After studying this unit, you should be able to:

- define statistical inference;
- define the basic terms as population, sample, parameter, statistic, estimator, estimate, etc. used in statistical inference;
- explain the concept of sampling distribution;
- explore the importance and needs of sampling distribution;

- explain the concept of standard error;
- describe the most important theorem of Statistics "Central Limit Theorem";
- explain the concept of law of large numbers; and
- determine adequate sample size required for sample survey.

1.2 BASIC TERMINOLOGY

Before discussing the sampling distribution of a statistic, we shall be discussing basic definitions of some of the important terms which are very helpful to understand the fundamentals of statistical inference.

Population

In general sense "population" means a large group of people who live in a particular geographical area. For example, the group of people who live in New Delhi, the group of people working in IGNOU, students enrolled in PGDAST programme in IGNOU, etc.

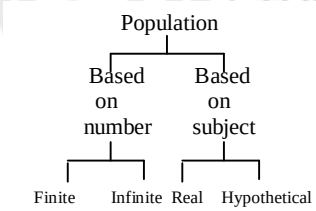
In Statistics, population has a much broader meaning. It does not only refer to people but also the group of elements or units under consideration by the analyst. Thus, population is the collection or group of individuals /items /units/ observations under study. For example, the collection of books in a library, the particles in a salt bag, the rivers in India, the students in a classroom, etc. are considered as populations in Statistics.

The group of individuals / items / units/ observations under study is known as population.

The total number of elements / items / units / observations in a population is known as population size and denoted by N. The characteristic under study may be denoted by X or Y.

We may classify the population into two types as:

(1) Finite and Infinite Population



If a population contains finite number of units or observations, it is called a **finite population**. For example, the population of students in a class, the population of bolts produced in a factory in a day, the population of electric bulbs in a lot, the population of books in a library, etc. are the finite populations because in these examples the number of units in the population is finite in numbers therefore these are all the examples of finite population.

If a population contains an infinite (uncountable) number of units or observations, it is called an **infinite population**. For example, the population of particles in a salt bag, the population of stars in the sky, the population of hairs on human body, etc. are infinite populations because in these examples the number of units in the population is not finite therefore these are all the examples of infinite population.

But theoretically sometimes, populations of too large in size are assumed infinite.

(2) Real and Hypothetical Population

The population can also be classified as real and hypothetical. A population comprising the items or units which are all physically present is known as **real population**. All of the examples given above are examples of a real population.

If a population consists the items or units which are not physically present but the existence of them can only be imagined or conceptualised then it is known

as **hypothetical population**. For example, the population of heads or tails in successive tosses of a coin a large number of times is considered as hypothetical population.

Sample

To extract the information from all the elements or units or items of a large population may be, in general, time consuming, expensive and difficult. And also if the elements or units of population are destroyed under investigation then gathering the information from all the units is not make a sense. For example, to test the strength of the chalk, to test the quality of crackers, etc. the chalks and crackers are destroyed under testing procedure. In such situations, a small part of population is selected from the population which is called a **sample.** Thus, the sample can be defined as below:

"A sample is a part / fraction / subset of the population."

The procedure of drawing a sample from the population is called **sampling.** You will learn about sampling in detail in Block 1 of course MST-005. The number of units selected in the sample is known as sample size and it is denoted by n.

Complete Survey and Sample Survey

Statistical data may be collected by two methods:

(1) Complete Survey or Complete Enumeration or Census

(2) Sample Survey or Sample Enumeration

(1) Complete Survey or Complete Enumeration or Census

When each and every element or unit of the population is investigated or studied for the characteristics under study then we call it **complete survey** or **census**. For example, suppose we want to find out the average height of the students of a study centre then if we measure the height of each and every student of this study centre to find the average height of the students then such type of survey is called complete survey.

(2) Sample Survey or Sample Enumeration

When only a part or a small number of elements or units (i.e. sample) of population are investigated or studied for the characteristics under study then we call it **sample survey** or **sample enumeration**. In the above example, if we select some students of this study centre and measure the height to find average height of the students then such type of survey is called sample survey.

Simple Random Sampling or Random Sampling

The simplest and most common method of sampling is simple random sampling. In simple random sampling, the sample is drawn in such a way that each element or unit of the population has an equal and independent chance of being included in the sample. If a sample is drawn by this method then it is known as a **simple random sample or random sample**. The random sample of size n is denoted by $X_1, X_2, ..., X_n$ or $Y_1, Y_2, ..., Y_n$ and the observed value of this sample is denoted by $x_1, x_2, ..., x_n$ or $y_1, y_2, ..., y_n$. Some author use $x_1, x_2, ..., x_n$ to represent the random sample instead of $X_1, X_2, ..., X_n$. But throughout the course, we use capital letters to represent the random sample, that is, $X_1, X_2, ..., X_n$ and $X_1, X_2, ..., X_n$ simply be assumed that these are the independent identically distributed(iid) random variables.

In simple random sampling elements or units are selected or drawn one by one therefore it may be classified into two types as:

(1) Simple Random Sampling without Replacement (SRSWOR)

In simple random sampling, if the elements or units are selected or drawn one by one in such a way that an element or unit drawn at a time is not replaced back to the population before the subsequent draws is called **SRSWOR**. If we draw a sample of size n from a population of size N without replacement then total number of possible samples is ${}^{N}C_{n}$. For example, consider a population that consists of three elements, A, B and C. Suppose we wish to draw a random sample of two elements then N = 3 and n = 2. The total number of possible random samples without replacement is ${}^{N}C_{n} = {}^{3}C_{2} = 3$ as (A, B), (A, C) and (B, C).

(2) Simple Random Sampling with Replacement (SRSWR)

In SRSWR each of the 1^{st}, 2^{nd}, .., n^{th} draw the elements are remain same N due to replacement so by rule of multiplication total number of possible samples is $N \times N \times ...n \text{ times} = N^{n}$.

In simple random sampling, if the elements or units are selected or drawn one by one in such a way that a unit drawn at a time is replaced back to the population before the subsequent draw is called **SRSWR**. In this method, the same element or unit can appear more than once in the sample and the probability of selection of a unit at each draw remains same i.e. 1/N. In this method, total number of possible samples is N^{n}. In above example, the total number of possible random samples with replacement is $N^{n} = 3^{2} = 9$ as (A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B) and (C, C).

Note1: The statistical inference is based on the theoretical distribution so in whole statistical inference, we assume that the random sample is selected from the infinite population or from a finite population with replacement so that removal of the sample unit has no appreciable effect on the composition of the population. That is, we draw the random sample such that:

(a) The n successive sample observations are independent and

(b) The population composition remains constant from draw to draw.

For example, if we draw a random sample $X_1, X_2, ..., X_n$ from normal population with mean μ and variance σ^2 then X_i's are independent and follow same normal distribution.

Parameter

A parameter is a function of population values which is used to represent the certain characteristic of the population.

A group of elements or units under study can be regarded as a population and the characteristics of population can be described with some measures such as total numbers of items, average, variance, etc. These measures are known as parameters of the population. Thus, we can define parameter as:

A parameter is a function of population values which is used to represent the certain characteristic of the population. For example, population mean, population variance, population coefficient of variation, population correlation coefficient, etc. are all parameters. Population parameter mean usually denoted by μ and population variance denoted by σ^2.

We know that the population can be described with the help of distribution and the distribution is fully determined with the help of its constants such as, in case of normal distribution, we need to know μ and σ^2 to determine the normal distribution, in case of Poisson distribution, we need to know λ, etc. These constants are known as parameter.

Sample Mean and Sample Variance

If $X_1, X_2, ..., X_n$ is a random sample of size n taken from a population whose probability density(mass) function f(x, θ) then sample mean is defined as

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

And sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Here, we divide $\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ by (n – 1) rather than n as our definition of the variance described in Unit 2 of MST-002. The reason for taking (n – 1) in place of n will become clear in the Section 5.4 of Unit 5 of this course.

Statistic

Any quantity calculated from sample values and does not contain any unknown parameter is known as **statistic**. For example, if $X_1, X_2, ..., X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2 (both are unknown) then sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic whereas

$\bar{X} - \mu$ and $\bar{X}/\sigma$ are not statistics because both are function of unknown parameters.

Estimator and Estimate

Generally, population parameters are unknown and the whole population is too large to find out the parameters. Since the sample drawn from a population always contains some or more information about the population, therefore in such situations, we guess or estimate the value of the parameter under study based on a random sample drawn from that population.

So if a statistic is used to estimate an unknown population parameter then it is known as **estimator** and the value of the estimator based on observed value of the sample is known as **estimate** of parameter. For example, suppose the parameter λ of the Poisson population f(x, λ) is unknown so we draw a random sample $X_1, X_2, ..., X_n$ of size n from this Poisson population to estimate λ. If we use sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate λ then $\bar{X}$ is called estimator and any particular value of this estimator, say, $\bar{x}$ is called estimate of unknown parameter λ. Consider another example, if we want to estimate the average height (μ) of students in a college with the help of sample mean $\bar{X}$ then $\bar{X}$ is the estimator and its particular value, say, 165 cm is the estimate of the population average height (μ).

Any statistic used to estimate an unknown population parameter is known as **estimator** and the particular value of the estimator is known as **estimate** of parameter. The estimated value of sample mean and sample variance are denoted by $\bar{x}$ and s^2 respectively.

In this course, we use capital letters for estimator and small letters for estimated value. In general, the estimator is denoted by $T_n = t(X_1, X_2, \ldots, X_n)$ where 'n' denotes the sample size and the estimator T is a function of the random sample $X_1, X_2, \ldots, X_n$.

Now, it is right place to do some examples.

Example 1: Categories the following populations as finite and infinite:

(i) Population of newly born babies during a year in a particular hospital.

(ii) Population of real numbers between 1 and 2.

(iii) Population of number of words on this page.

Solution: (i) and (iii) are the finite populations because the number of elements (babies or words) in both are finite whereas (ii) is infinite because there are infinite real numbers are possible between 1 and 2.

Example 2: If population size is 6 then how many samples of size of 4 are possible with replacement?

Solution: Here, we are given that

Population size = N = 6,

Sample size = n = 4

Since we know that all possible samples of size n taken from a population of size N with replacement are N^n so in our case $N^n = 6^4 = 1296$.

Now, you check your understanding of above discussion by answering the following exercises.

E1) In which situation we use only sample survey:

(i) Blood testing.

(ii) To know the average income of the people living in a colony.

(iii) To know the number of students enrolled in IGNOU.

(iv) To know the average life of electric bulbs.

E2) If $X_1, X_2, ..., X_n$ is a random sample of size n taken from normal population with mean μ and variance σ^2 both are unknown then find the statistic in the following:

(i) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$ (ii) $\sum_{i=1}^{n} X_i$

(iii) $\sum_{i=1}^{n} X_i / \sigma$ (iv) $\sum_{i=1}^{n} X_i^2$

E3) The weights (in kg.) of 5 workers in a factory are 56, 62, 74, 45 and 50. How many samples of size of 2 are possible with replacement? Also write all possible samples of size 2.

1.3 INTRODUCTION TO SAMPLING DISTRIBUTION

As we have discussed in Section 1.1 that if the population is too large or the units or items of the population are destructive in nature or there is a limited resources such as men power, money, etc., then it is not possible practically to examine each and every unit of the population to obtain the necessary information about the population. For example, Suppose we want to know the average life of electric bulbs of a certain brand which are manufactured by a company. The company manufactures a lot of bulbs say 5,000 per day. Then gathering information about average life of all the bulbs is not making a sense

because the bulbs are destroyed under investigation. In another example, if we want to know the average income of the persons living in a big city then collecting the information from each and every person is very much time and manpower consuming. In such situations, one can draw sample from the population under study and utilize sample observations to extract the necessary information about the population. The results obtained from the sample are projected in such a way that they are valid for the entire population. Therefore, the sample works like a **"Vehicle"** to reach (drawing) at valid conclusions about the population. Thus, statistical inference can be defined as:

"The process of projecting the sample results for the whole population is known as **statistical inference**."

For drawing the inference about the population, we analyse the sample data, that is, we calculate the sample statistic as sample mean, sample proportion, sample variance, etc. Generally, if we want to draw the inference about the population mean we use sample mean, about population variance we use sample variance, etc.

For example, suppose we want to study the average life of electric bulbs produced by the company discussed in above example and we decide to estimate the average life of electric bulbs on the basis of a sample. So we take a sample, say, 10 bulbs and measure the life of each bulb selected in the sample. The observed values of life of bulbs are given in column 2 of the Table 1.1 given below:

Table 1.1: Life Times of 10 Bulbs in Sample-I and Sample-II

S.No.	Life of Bulbs (in hours)	
	Sample-I	Sample-II
1	2000	400
2	500	0
3	1000	1400
4	1500	1200
5	1200	800
6	200	700
7	700	1200
8	1400	800
9	800	600
10	1100	1500
Total	10400	8600
$\bar{X}$	1040	860

From the above table, the average (mean) life of these selected bulbs is 1040 hours. Now, if we are using this sample to make inference about the population average life then we say that, the sample mean is 1040 hours give an estimate of the average life of electric bulbs for the whole population. But if we take another sample of same size 10 from the population and calculate the sample mean as shown in the above table. The average life of bulbs selected in sample-II is 860 hours. This is different from the mean of sample-I. Thus, if we are using this sample mean to estimate the population mean then we gets a different estimate of the average life of bulbs for the whole population. Similarly, we can take many samples of same size 10 and for each sample we get sample mean which may or may not be distinct. From all of the sample means, we try to estimate the average life of bulbs of the whole population.

For better understanding of the process of generalising these sample results to the whole population, we consider the example in which size of the population is very small.

Consider a population comprising four typists who type the sample page of a manuscript. The number of errors made by each typist is shown in Table 1.2 given below:

Table 1.2: Number of Error per Typist

Typist	Number of Errors
A	4
B	2
C	3
D	1

The population mean (average number of errors) can be obtained as

$$\mu = \frac{4+2+3+1}{4} = 2.5$$

Now, let us assume that we do not know the average number of errors made by typists. So we decide to estimate the population mean on the basis of sample of size n = 2. There are $N^n = 4^2 = 16$ possible simple random samples with replacement of size 2. All possible samples of size n = 2 are given below and for each sample the sample mean is calculated as shown in Table 1.3 given below:

Table 1.3: Calculation of Sample Mean

Sample Number	Sample in Term of Typist	Sample Observation	Sample Mean $(\bar{X})$
1	(A, A)	(4, 4)	4.0
2	(A, B)	(4, 2)	3.0
3	(A, C)	(4, 3)	3.5
4	(A, D)	(4, 1)	2.5
5	(B, A)	(2, 4)	3.0
6	(B, B)	(2, 2)	2.0
7	(B, C)	(2, 3)	2.5
8	(B, D)	(2, 1)	1.5
9	(C, A)	(3, 4)	3.5
10	(C, B)	(3, 2)	2.5
11	(C, C)	(3, 3)	3.0
12	(C, D)	(3, 1)	2.0
13	(D, A)	(1, 4)	2.5
14	(D, B)	(1, 2)	1.5
15	(D, C)	(1, 3)	2.0
16	(D, D)	(1, 1)	1.0

From the above table, we can see that the value of the sample statistic (sample mean) is varying from sample to sample and out of 16 samples only 4 samples(4,7,10 and 13) have their means equal to population mean whereas other 12 samples result in some error in the estimation process. The error is the difference between the population mean and the sample mean used for estimate.

Now, consider all possible values of sample mean and calculate their probabilities (using relative frequency approach of probability described in Unit 2 of MST-003) of occurrence. Then we arrange every possible value of sample mean with their respective probabilities in the following Table 1.4 given below:

Table 1.4: Sampling Distribution of Sample Means

S. No.	$\bar{X}$	Frequency(f)	Probability(p)
1	1.0	1	1/16 = 0.0625
2	1.5	2	2/16 = 0.1250
3	2.0	3	3/16 = 0.1875
4	2.5	4	4/16 = 0.2500
5	3.0	3	3/16 = 0.1875
6	3.5	2	2/16 = 0.1250
7	4.0	1	1/16 = 0.0625

So the arrangement of all possible values of sample mean with their corresponding probabilities is called the sampling distribution of mean. Thus, we can define the sampling distribution of a statistic as:

"The probability distribution of all possible values of a sample statistic that would be obtained by drawing all possible samples of the same size from the population is called sampling distribution of that statistic."

The sampling distribution of sample means itself has mean, variance, etc. Therefore, the mean of sample means can be obtained by the formula

$$\text{Mean of sample means} = \bar{\bar{X}} = \frac{1}{K}\sum_{i=1}^{k}\bar{X}_i f_i \quad \text{where,} \quad K = \sum_{i=1}^{k} f_i$$

$$= \frac{1}{16}\left(1.0\times 1 + 1.5\times 2 + \ldots + 4.0\times 1\right) = 2.5 = \mu$$

The mean of sample means can also be calculated as

$$E(\bar{X}) = \bar{\bar{X}} = \sum_{i=1}^{k}\bar{X}_i p_i = 1.0\times\frac{1}{16} + 1.5\times\frac{2}{16} + \ldots + 4.0\times\frac{1}{16} = 2.5$$

Thus, we have seen for this population that mean of sample means is equal to the population mean, that is, $\bar{\bar{X}} = \mu = 2.5$. The fact that these two means are equal is not a chance as it can be shown that $\bar{\bar{X}}$ equals to population mean for any population and any given sample size.

But if the population is large say having 1,000 elements and we want to draw all samples of size 2 then there are $(1000)^2 = 1000000$ possible simple random samples with replacement. Then the listing of all such samples and finding the sampling distribution would be a difficult task empirically. Therefore, we may consider a theoretical experiment in which we draw all possible sample of a fixed size and obtain a sampling distribution.

Although in practice only a random sample is actually selected and the concept of sampling distribution is used to draw the inference about the population parameters.

It is now time for you to try the following exercise to make sure that you understand about sampling distribution.

E4) If lives of 3 Televisions of certain company are 8, 6 and 10 years then construct the sampling distribution of average life of Televisions by taking all samples of size 2.

1.4 STANDARD ERROR

As we have seen in the previous section that the values of sample statistic may vary from sample to sample and all the sample values are not equal to the population parameter. Now, one can be interested to measure how much the values of sample statistic vary from the population parameter on average. As we have already studied in Unit 2 of MST-002 that the standard deviation is used as a measure of variation.

Thus, for measuring the variation in the values of sample statistic around the population parameter we calculate the standard deviation of the sampling distribution. This is known as standard error of that statistic. Thus, standard error of a statistic can be defined as:

Standard deviation of sampling distribution of a statistic is called standard error.

"The standard deviation of a sampling distribution of a statistic is known as standard error and it is denoted by SE."

Therefore, the standard error of sample mean is given by

$$SE(\bar{X}) = \sqrt{\frac{1}{K}\sum_{i=1}^{k} f_i (\bar{x}_i - \mu)^2} \text{ where, } K = \sum_{i=1}^{k} f_i$$

In the previous example, we can calculate the standard error of sample mean as

$$SE(\bar{X}) = \sqrt{\frac{1}{16}\left[1\times(1.0-2.5)^2 + 2\times(1.5-2.5)^2 + ... + 1\times(4.0-2.5)^2\right]}$$

$$= \sqrt{\frac{1}{16}(2.25 + 2 + ... + 2.25)} = \sqrt{\frac{10}{16}} = 0.791$$

The computation of the standard error is a tedious process. There is an alternative method to compute standard error of the mean from a single sample as:

If $X_1, X_2, ..., X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2 then the standard errors of sample mean ($\bar{X}$) is given by

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The standard error is used to express the accuracy or precision of the estimate of population parameter because the reciprocal of the standard error is the measure of reliability or precision of the statistic. Standard error also determines the probable limits or confidence limits (described in Units 7 & 8 of this course) within which the population parameter may be expected to lie with certain level of confidence. Standard error is also applicable in testing of hypothesis (described in Block 3 of this course).

The standard errors of some other well known statistics are given below:

1. The standard error of sample proportion (p) is given by

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

where, P is population proportion and Q = 1 - P.

2. The standard error of sample median ($\tilde{X}$) is given by

$$SE(\tilde{X}) = \frac{\pi}{2}\frac{\sigma}{\sqrt{n}} = 1.25331\ SE\ (\bar{X})$$

3. The standard error of difference of two sample means is given by

$$SE\left(\bar{X} - \bar{Y}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where, σ_1^2 and σ_2^2 are the population variances and n_1 and n_2 are the sample sizes of two independent samples.

4. The standard error of difference between two sample proportions is given by

$$SE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$$

where, P_1 and P_2 are population proportions of two different populations and $Q_1 = 1 - P_1$ & $Q_2 = 1 - P_2$.

From all the above formulae, we can understand that standard error is inversely proportional to the sample size. Therefore, as sample size increases the standard error decreases.

Note 2: These standard errors discussed above were calculated under the assumption that sampling is done either from an infinite population or from a finite population with replacement. But, in real life sampling problems, most sampling plans do not permit an element to be selected twice in a given sample (i.e. sampling with replacement). Consequently, if the population is not large in relation to the size of the sample and sampling is done without replacement then we multiply the above standard errors by the correction factor $\sqrt{\frac{N-n}{N-1}}$.

where, N is the population size. This correction factor is known as finite population correction factor.

Therefore, in this case, the standard error of the sample mean is given by

$$SE\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Similarly,

$$SE(p) = \sqrt{\frac{N-n}{N-1}\frac{PQ}{n}}$$

In practice, if the sample size n is less than or equal to 10% of the population size N, this correction factor may be ignored for all practical purposes.

Let us do one example relating to standard error.

Example 3: Diameter of a steel ball bearing produced by a semi-automatic machine is known to be distributed normally with mean 12 cm and standard deviation 0.1 cm. If we take a random sample of size 10 with replacement then find standard error of sample mean for estimating the population mean of diameter of steel ball bearing for whole population.

Solution: Here, we are given that

$$\mu = 12,\ \sigma = 0.1,\ n = 10$$

Since the sampling is done with replacement therefore the standard error of sample mean for estimating population mean is given by

$$SE(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{10}} = 0.03$$

In the same way, you can try the following exercise.

E5) The average weight of certain type of tyres is 200 pounds and standard deviation is 4 pounds. A sample of 50 tyres is selected. Obtain the standard error of sample mean.

E6) A machine produces a large number of items of which 15% are found to be defective. If a random sample of 200 items is taken from the population, then find the standard error of sampling distribution of proportion.

1.5 CENTRAL LIMIT THEOREM

The central limit theorem is the most important theorem of Statistics. It was first introduced by De Movers in the early eighteenth century.

According to the central limit theorem, the sampling distribution of the sample means tends to normal distribution as sample size tends to large $(n > 30)$.

According to the central limit theorem, if $X_1, X_2, ..., X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2 then the sampling distribution of the sample mean tends to normal distribution with mean μ and variance σ^2/n as sample size tends to large $(n > 30)$ whatever the form of parent population, that is,

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

and the variate

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

follows the normal distribution with mean 0 and variance unity, that is, the variate Z follows standard normal distribution.

We do not intend to prove this theorem here, but merely show graphical evidence of its validity in Fig. 1.1. Here, we will also try to show that how large must the sample size be for which we can assume that the central limit theorem applies?

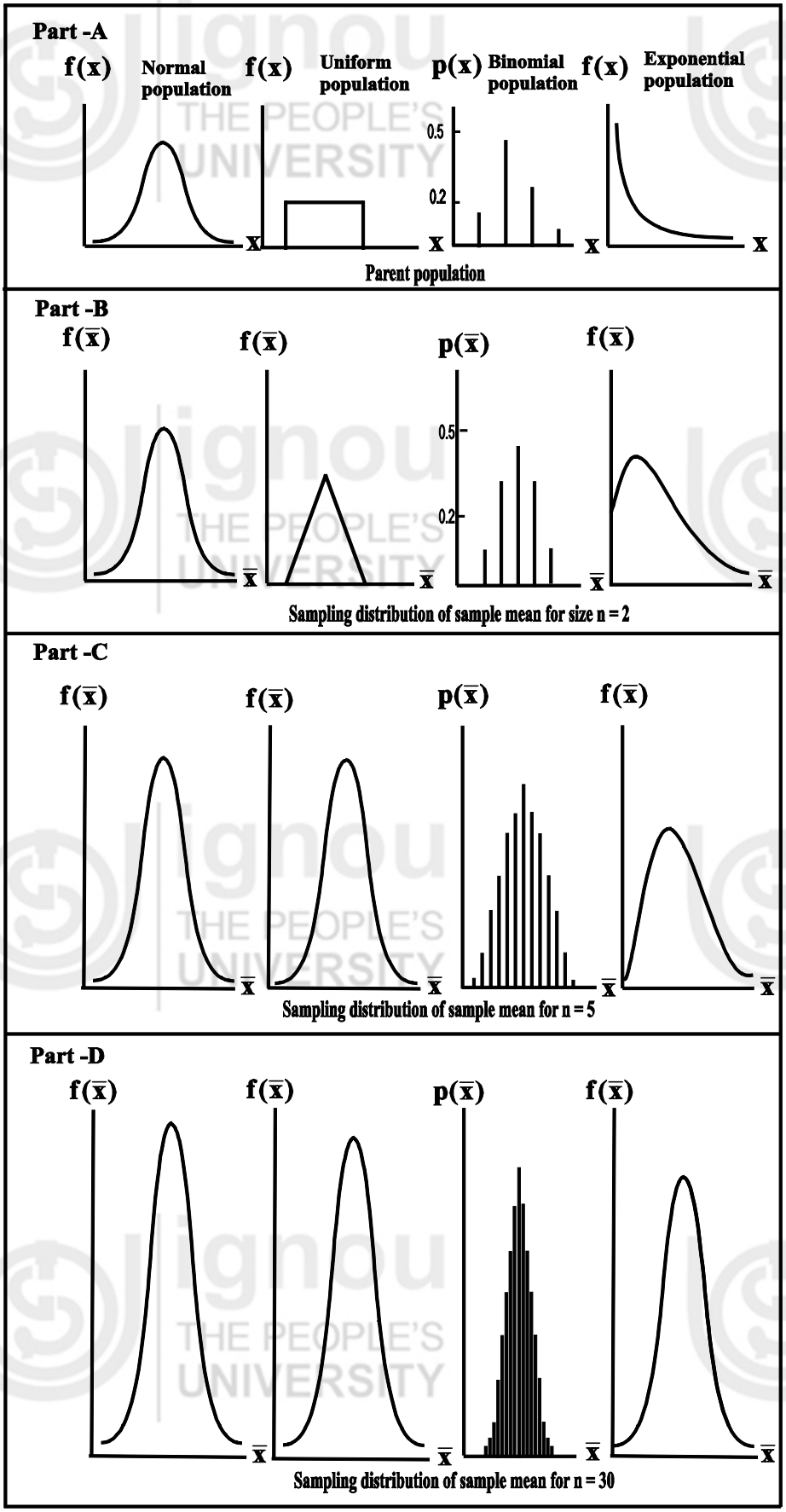


Fig. 1.1: Sampling distribution of sample means for various populations when n = 2, n = 5 and n = 30

In this figure, we are trying to understand the sampling distribution of sample mean $\bar{X}$ for different populations and for varying sample sizes. We divide this figure into four parts A, B, C and D. The part 'A' of this figure shows four different populations as normal, uniform, binomial and exponential. The rest parts B, C and D represent the shape of sampling distribution of mean of sizes n = 2, n = 5 and n = 30 respectively drawn from the populations shown in first row (Part-A). From the first column of this figure, we observed that when the parent population is normal then all the sampling distributions for varying sample sizes are also normal, having same mean but their variances decrease as n increases.

The second column of this figure represents the uniform population. Here, we observe that the sampling distribution of $\bar{X}$ is symmetrical when n = 2 and tends to normal when n = 30.

However, the third column of this figure represents the binomial population (discrete). Again when n = 2, the sampling distribution of $\bar{X}$ is symmetrical and for n = 5 it is quite bell shaped and tends to normal when n = 30. The last column of this figure represents the exponential population which is highly skewed. Here, we observe that for n =30, the distribution of $\bar{X}$ is symmetrical bell shaped normal.

Here, we conclude that if we draw a random sample of large size n > 30 from the population then the sampling distribution of $\bar{X}$ can be approximated by a normal probability distribution, whatever the form of parent population.

Note 3: The central limit theorem can also be applicable in the same way for the sampling distribution of sample proportion, sample standard deviation, difference of two sample means, difference of two sample proportions, etc. that is, if we take a random sample of large size $(n = 36 > 30)$ from the population then the sampling distribution of sample proportion, sample standard deviation, difference of two sample means, difference of two sample proportions, etc. approximated by a normal probability distribution, whatever the form of parent population. The mean and variance of these sampling distributions shall be discussed in the next unit.

Let us do one example based on central limit theorem.

Example 4: Suppose in IGNOU, the average (mean) weight of all male students is 60 kg and standard deviation is 25 kg. If a sample of 36 male students is selected at random, find the probability that the male students having average weight

(i) more than 70 kg
(ii) less than 55 kg
(iii) between 50 kg and 65 kg

Solution: Here, we are given that

$$\mu = 60 \text{ kg}, \sigma = 25 \text{ kg}, n = 36$$

Since, we draw a large sample (n > 30), therefore by central limit theorem the sampling distribution of the sample means will also follow a normal distribution with mean

$$E(\bar{X}) = \mu = 60$$

and variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{(25)^2}{36} = 17.36$$

(i) The probability that the male students having average weight more than 70 kg is given by

$$P[\bar{X} > 70] \qquad [\text{See Fig. 1.2}]$$

To get the value of this probability, we convert the variate $\bar{X}$ into a standard normal variate Z by the transformation

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - 60}{\sqrt{17.36}} = \frac{\bar{X} - 60}{4.17}$$

Therefore, subtract 60 from each term and then divide each term by 4.17 in the above inequality. Thus, the probability becomes

$$P\left[\frac{\bar{X} - 60}{4.17} > \frac{70 - 60}{4.17}\right] = P[Z > 2.40]$$

$$= 0.5 - P[0 \le Z \le 2.40]$$

$$= 0.5 - 0.4918 \qquad [\text{Using table area under normal curve}]$$

$$= 0.0082$$

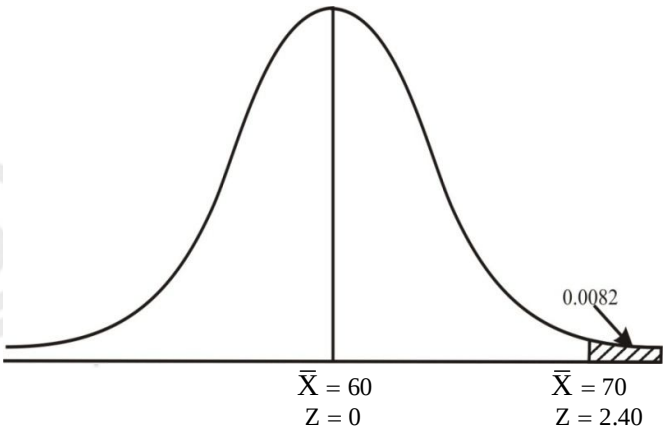


Fig. 1.2

(ii) The probability that the male students having average weight less than 55 kg is given by

$$P[\bar{X} < 55] = P\left[\frac{\bar{X} - 60}{4.17} < \frac{55 - 60}{4.17}\right]$$

$$= P[Z < -1.20] \qquad [\text{See Fig. 1.3}]$$

$$= P[Z > 1.20] \qquad [\because \text{normal curve is symmetrical about the line } Z = 0.]$$

$$= 0.5 - P[0 \le Z \le 1.20]$$

$$= 0.5 - 0.3849 \qquad [\text{Using table area under normal curve}]$$

$$= 0.1151$$

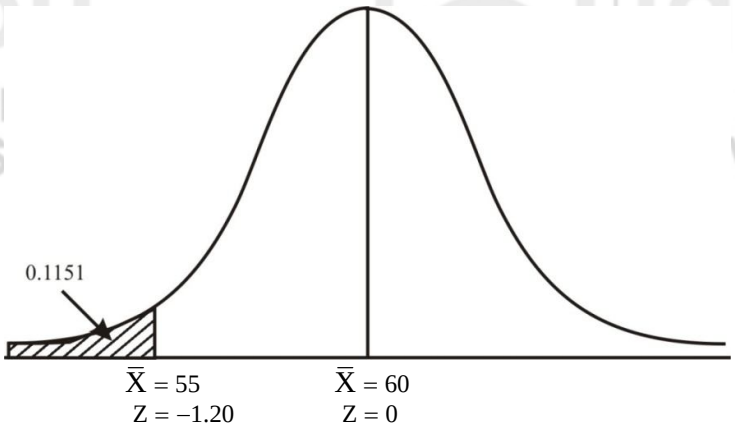


Fig. 1.3

(iii) The probability that the male students having average weight between 50 kg and 65 kg is given by

$$P\left[50 < \bar{X} < 65\right] = P\left[\frac{50-60}{4.17} < \frac{\bar{X}-60}{4.17} < \frac{65-60}{4.17}\right]$$

$$= P\left[-2.40 < Z < 1.20\right] \qquad \text{[See Fig.1.4]}$$

$$= P\left[-2.40 < Z < 0\right] + P\left[0 < Z < 1.20\right]$$

$$= P\left[0 < Z < 2.40\right] + P\left[0 < Z < 1.20\right]$$

$$= 0.4918 + 0.3849 \qquad \text{[Using table area under normal curve]}$$

$$= 0.8767$$

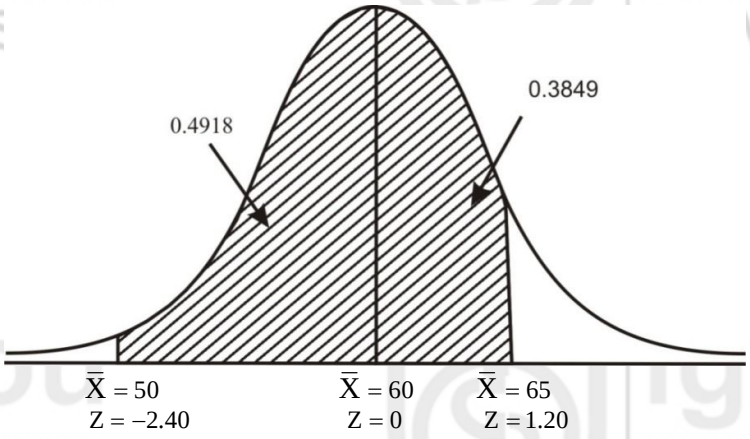


Fig. 1.4

Now, you can try the following exercise which will make you more user friendly with the use of central limit theorem.

E7) Average height of the students of science group in a college is 65 inches with a standard deviation of 2.2 inches. If a sample of 40 students is selected at random, what is the probability that the average height of these 40 students lies between 64 and 65.5 inches?

1.6 LAW OF LARGE NUMBERS

We have already discussed in Section 1.2 of this unit that the population parameters are generally unknown and for estimating parameters, we draw all possible random samples of same size from the population and calculate the values of sample statistic such as sample mean, sample proportion, sample variance, etc. for all samples and with the help of these values we form sampling distribution of that statistic. Then we draw inference about the population parameters.

But in real-world the sampling distributions are never really observed. The process of finding sampling distribution would be very tedious because it would involve very large number of samples. So in real-world problems, we draw a random sample from the population to draw inference about the population parameters. A very crucial question then arises: "Using a random sample of finite size say n, can we make a reliable inference about population parameter?" The answer is "yes", reliable inference about population parameter can be made by using only a finite sample and we shall demonstrate this by **"law of large numbers"**. This law of large numbers can be stated in words as:

"A positive integer can be determined such that if a random sample of size n or lager is taken from a population with parameter, say, population mean (μ), the probability that the sample mean $\bar{X}$ will deviate from population mean μ can be made to be as close to 1 as desired."

This law states that for any two arbitrary small numbers ε and η where $\varepsilon > 0$ and $0 < \eta < 1$, there exists an integer n such that if a random sample of size n or larger is drawn from the population and calculate the sample mean $\bar{X}$ for this sample, then the probability that $\bar{X}$ deviates from μ by less than ε is greater than $1-\eta$ (as close to 1 as desired) , that is, $\bar{X}$ arbitrarily close to μ.

Symbolically this can be written as

For any $\varepsilon > 0$ and $0 < \eta < 1$ there exists an integer n such that $n \geq \dfrac{\sigma^2}{\varepsilon^2 \eta}$, where σ^2 is the finite variance of population then

$$P\left[\left|\bar{X} - \mu\right| < \varepsilon\right] \geq 1 - \eta$$

or $$P\left[-\varepsilon < \bar{X} - \mu < \varepsilon\right] \geq 1 - \eta \quad \left[\because |X| < a \Rightarrow -a < X < a\right]$$

The proof of "law of large numbers" is beyond the scope of this course.

Here, with an example we have to try to understand the law of large numbers in action, as we take more observations the sample mean approaches to the population mean.

Suppose that the distribution of the weight of all the young men living in a city is close to normal distribution with mean 65 kg and standard deviation 25 kg. To understand the law of large numbers, we calculate the sample mean weight ($\bar{X}$) for varying sample size n = 1, 2, 3, …. Fig. 1.5 shows the behaviour of the sample mean weight $\bar{X}$ of men chosen at random from this city. The graph plots the values of $\bar{X}$ (along vertical axis and sample size along horizontal axis) as sample size varying from n =1, 2, 3…

First we start a sample of size n = 1, that is, we select a man randomly from the city. Suppose selected man had weight 70 kg, therefore, the line of the graph starts from this point. Now, select second man randomly and suppose his weight is 55 kg. So for n = 2 the sample mean is

$$\bar{X} = \frac{70+55}{2} = 62.5 \text{ kg}$$

This is the second point on the graph. Now, select third man randomly from the city and suppose his weight is 83 kg. Therefore, for n = 3 the sample mean is

$$\bar{X} = \frac{70+55+83}{3} = 69.3 \text{ kg}$$

This is the third point on the graph.

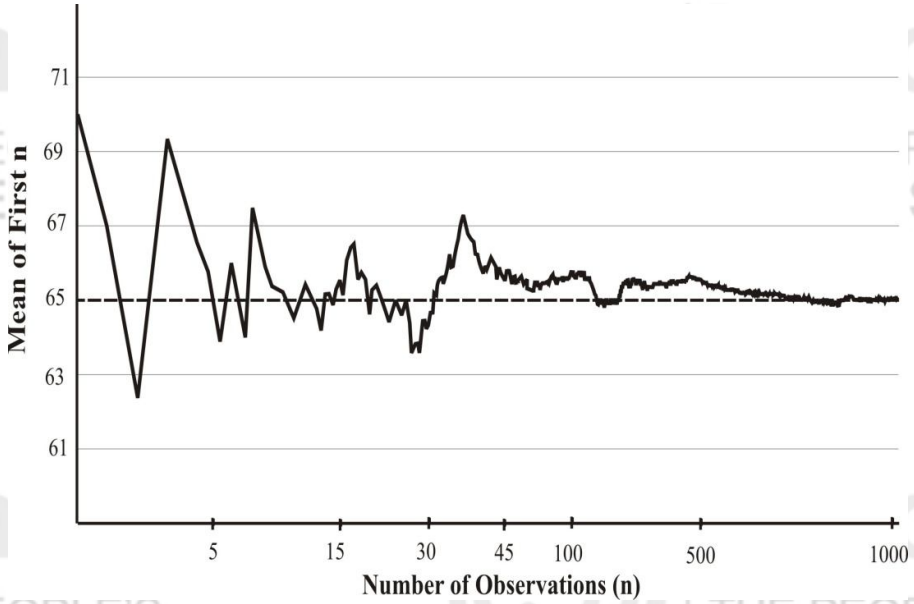


Fig. 1.5

This process will continue. From the graph we can see that the mean of the sample changes as we make more observations, eventually, however, the mean of the observations get close and close to the population mean 65 kg as the sample size increases.

Note 4: The law of large numbers can also be applied for the sample proportion, sample standard deviation, difference of two sample means, difference of two sample proportions, etc.

Example 5: A pathologist wants to estimate the mean time required to complete a certain analysis on the basis of sample study so that he may be 99% confident that the mean time may remain with ± 2 days of the mean. As per the available records, the population variance is 5 days^2. What must be the size of the sample for this study?

Solution: Here, we are given

$$1 - \eta = 0.99 \Rightarrow \eta = 0.01, \quad \varepsilon = 2, \quad \sigma^2 = 5$$

By the law of large numbers, we have

$$n \geq \frac{\sigma^2}{\varepsilon^2 \eta}$$

$$= \frac{5}{2^2 \times 0.01} = 125$$

That is, $n \geq 125$

Hence, at least 125 units must be drawn in a sample.

Example 6: An investigator wishes to estimate the mean of a population using a sample large enough that the probability will be 0.95 that the sample mean will not differ from the population mean by more than 25 percent of the standard deviation. How large a sample should be taken?

Solution: We have

$$1 - \eta = 0.95 \Rightarrow \eta = 0.05, \ \varepsilon = 0.25\sigma$$

By the law of large numbers, we have

$$n \geq \frac{\sigma^2}{\varepsilon^2 \eta}$$

$$= \frac{\sigma^2}{(0.25\sigma)^2 \times 0.05} = 320$$

That is, $n \geq 320$

Therefore, at least 320 units must be taken in the sample so that the required probability will be 0.95.

Now, try the following exercise for your practice.

E8) The mean of a population is unknown and having a variance equal to 2. Find out that how large a sample must be taken, so that the probability will be at least 0.95 that the sample mean will lie within the range of 0.5 of the population mean?

We now end this unit by giving a summary of what we have covered in it.

1.7 SUMMARY

In this unit, we have covered the following points:

1. The statistical procedure which is used for drawing conclusions about the population parameter on the basis of the sample data is called **"statistical inference"**.
2. The group of units or items under study is known as **"Population"** whereas a part or a fraction of population is known as **"sample"**.
3. A "**parameter"** is a function of population values which is used to represent the certain characteristic of the population and any quantity calculated from sample values and does not contain any unknown population parameter is known as "**statistic"**.
4. Any statistic used to estimate an unknown population parameter is known as "**estimator"** and the particular value of the estimator is known as "**estimate"** of parameter.
5. The probability distribution of all possible values of a sample statistic that would be obtained by drawing all possible samples of the same size from the population is called "**sampling distribution"** of that statistic.

6. The standard deviation of the sampling distribution of a statistic is known as **"standard error"**.
7. The most important theorem of Statistics is **"central limit theorem"** which state that the sampling distribution of the sample means tends to normal distribution as sample size tends to large (n > 30).
8. According to law of large numbers, a positive integer can be determined such that if a random sample of size n or larger is taken from a population with parameter, say, population mean (μ), the probability that the sample mean will be very closed to population mean, can be made as close to 1 as desired.

1.8 SOLUTIONS / ANSWERS

E1) In (i) and (iv) we use only sample survey because the units (persons or blubs) are destroyed under investigation and the information gathering from all the elements is not making a sense whereas in (ii) and (iii) we can use sample as well as complete survey.

E2) (ii) and (iv) are the statistics because both do not contain the unknown parameters (i.e. μ and σ^2) whereas (i) and (iii) are not statistics because they contain unknown parameters.

E3) Here, we are given that

$$N = 5, n = 2$$

All possible samples (with replacement) are $N^n = 5^2 = 25$ which are shown in the Table 1.5 given below:

Table 1.5: Possible samples

Sample Number	Sample Observation	Sample Number	Sample Observation
1	56, 56	14	74, 45
2	56, 62	15	74, 50
3	56, 74	16	45, 56
4	56, 45	17	45, 62
5	56, 50	18	45, 74
6	62, 56	19	45, 45
7	62, 62	20	45, 50
8	62, 74	21	50, 56
9	62, 45	22	50, 62
10	62, 50	23	50, 74
11	74, 56	24	50, 45
12	74, 62	25	50, 50
13	74, 74		

E4) Here, we are given that

$$N = 3, n = 2$$

Since we have to estimate the average life of Televisions on the basis of samples of size n = 2 therefore, all possible samples (with replacement) are $N^n = 3^2 = 9$ and for each sample we calculate the sample mean as shown in Table 1.6 given below:

Table 1.6: Calculation of Sample Mean

Sample Number	Sample Observation	Sample Mean ($\bar{X}$)
1	8, 8	8
2	8, 6	7
3	8,10	9
4	6, 8	7
5	6, 6	6
6	6, 10	8
7	10, 8	9
8	10, 6	8
9	10, 10	10

Since the arrangement of all possible values of sample mean with their corresponding probabilities is called the sampling distribution of mean thus, we arrange every possible value of sample mean with their respective probabilities in the following Table 1.7 given below:

Table 1.7: Sampling distribution of sample means

S.No.	Sample Mean ($\bar{X}$)	Frequency	Probability
1	6	1	1/9
2	7	2	2/9
3	8	3	3/9
4	9	2	2/9
5	10	1	1/9

E5) Here, we are given that

$$\mu = 200,\ \sigma = 4,\ n = 50$$

Since the sampling is done with replacement therefore we know that the standard error of sample mean for estimating population mean is given by

$$SE\left(\bar{X}\right) = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{50}} = 0.57$$

E6) Here, we are given that

$$P = \frac{15}{100} = 0.15,\ n = 200$$

Therefore, we know that the standard error of sample proportion is given by

$$SE\left(p\right) = \sqrt{\frac{PQ}{n}} = \sqrt{\frac{0.15\times\left(1-0.15\right)}{200}} = 0.025$$

E7) Here, we are given that

$$\mu = 65,\ \sigma = 2.2,\ n = 40$$

Since we draw a large sample $n = 40 > 30$, therefore, by central limit theorem, the sampling distribution of the sample mean will follow a normal distribution with mean

$$E\left(\bar{X}\right) = \mu = 65$$

and variance

$$\text{Var}\left(\bar{X}\right) = \frac{\sigma^2}{n} = \frac{(2.2)^2}{40} = 0.12$$

Therefore, the probability that the average height of these 40 students will lie between 64 and 65.5 inches is given by

$$P\left[64 < \bar{X} < 65.5\right] \qquad \left[\text{See Fig.1.6}\right]$$

To get the value of this probability, we convert the normal variate $\bar{X}$ into a standard normal variate Z by the transformation

$$Z = \frac{\bar{X} - E\left(\bar{X}\right)}{\sqrt{\text{Var}\left(\bar{X}\right)}} = \frac{\bar{X} - 65}{\sqrt{0.12}} = \frac{\bar{X} - 65}{0.35}$$

Therefore, subtract 65 from each term and then divide each term by 0.35 in the above inequality. Thus, the probability becomes

$$P\left[\frac{64-65}{0.35} < \frac{\bar{X}-65}{0.35} < \frac{65.5-65}{0.35}\right] = P\left[-2.86 < Z < 1.43\right]$$

$$= P\left[-2.86 < Z < 0\right] + P\left[0 < Z < 1.43\right]$$

$$= P\left[0 < Z < 2.86\right] + P\left[0 < Z < 1.43\right]$$

$$= 0.4979 + 0.4236 = 0.9215$$

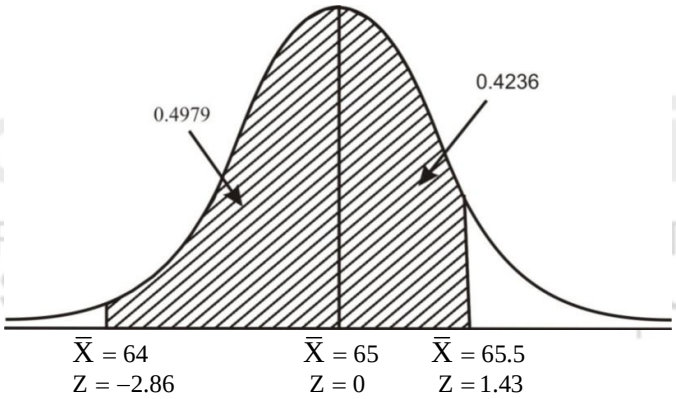


Fig. 1.6

E8) Here, we are given that

$$\sigma^2 = 2,\ \varepsilon = 0.5,\ 1 - \eta = 0.95 \Rightarrow \eta = 0.05$$

By the law of large numbers, we have

$$n \geq \frac{\sigma^2}{\varepsilon^2 \eta}$$

$$= \frac{2}{(0.5)^2 \times (0.05)} = 160$$

That is, $n \geq 160$

Therefore, at least 160 units must be taken in the sample so that the probability will be at least 0.95 that the sample mean will lie within 0.5 of the population mean.