

# UNIT 1

## BASIC CONCEPTS OF SAMPLING DISTRIBUTION

### Structure

---

|     |                                       |     |                      |
|-----|---------------------------------------|-----|----------------------|
| 1.1 | Introduction                          | 1.6 | Law of Large Numbers |
|     | Expected Learning Outcomes            | 1.7 | Summary              |
| 1.2 | Basic Terminology                     | 1.8 | Terminal Questions   |
| 1.3 | Introduction to Sampling Distribution | 1.9 | Solutions /Answers   |
| 1.4 | Concept of Standard Error             |     |                      |
| 1.5 | Central Limit Theorem                 |     |                      |

### 1.1 INTRODUCTION

---

In many situations, we have to extract some information from all units or items of a group/population. But, if

- the whole population is too large to study,
- the units of the population are destructive in nature,
- there are limited resources and manpower available, etc.

then gathering the information from all units is not practically inconvenient and sometimes units are destroyed under investigation. For example, as you know many of us use Facebook and if you are interested to know the average age of the Facebook user, then you have to survey every person in the world who uses Facebook. But it is not possible to survey everyone in the world.

Therefore, in most of the cases of daily life, business, industry, etc., the information about the whole group/population is gathered through a random **sample**. The results of a properly taken sample enable the investigator to arrive at generalisations that are valid for the entire group/population. The process of generalising sample results to the population is called **Statistical Inference**.

To draw inferences about the population characteristics (known as parameters) on the basis of a sample, we require the **sampling distribution** of the statistic (function of sample observations). This unit as the title is devoted to explaining the basic concepts required to understand the sampling distribution as well as statistical inference.

This unit is divided into 9 sections. Section 1.1 is introductory. In Section 1.2, we defined the basic terminology used in statistical inference such as population and sample, parameter and statistic, simple random sampling, estimator and estimate, etc. The concept and role of sampling distributions in statistical inference are described in Section 1.3. In Section 1.4, you will study the concept of standard error. The most important theorem “**central limit theorem**” and law “**law of large numbers**” of Statistics are described with their applications in Sections 1.5 and 1.6, respectively. The unit ends by providing a summary of what we have discussed in this unit in Section 1.7 and the terminal questions, and the solution of the SAQs/TQs are given in Sections 1.8 and 1.9, respectively.

In the next unit, we shall discuss the sampling distributions of sample means.

## Expected Learning Outcomes

After studying this unit, you should be able to:

- ❖ define statistical inference;
- ❖ define the basic terms such as population and sample, parameter and statistic, estimator and estimate, etc. used in statistical inference;
- ❖ explain the concept of the sampling distribution and standard error;
- ❖ describe the most important theorem of Statistics “central limit theorem”;
- ❖ apply the central limit theorem in the real world; and
- ❖ explain the concept of the law of large numbers with its application.

## 1.2 BASIC TERMINOLOGY

Before starting the discussion on the sampling distribution/statistical inference, you should understand the basic definitions of some of the important terms which are very helpful to understand the fundamentals of statistical inference.

### Population

In a general sense “population” means **a group of people who live in a particular geographical area or the entire group of individuals or objects that share a common characteristic and are of interest to a researcher.** For example, the group of people who live in New Delhi, the group of teachers working in IGNOU, students enrolled in the MSCAST programme in IGNOU, etc.

In Statistics, population need not consist only people but also the group of elements or units under consideration by the analyst. Thus, we can define a population as

**“A population is a collection/group of individuals /items /units/ observations under study.”**

For example, the collection of laptops of a company, the group of universities who are offering M.Sc. in Applied Statistics, the learners in a counselling session, etc. are considered as populations in Statistics.

In statistical inference, a population need not consist only of people or units

but also consider the measurements of the units. Thus, we can also define it as

**“A population is a group of measurements in the quantitative or qualitative form of the characteristic under study.”**

For example, salaries of employees in a company, weights of newly born babies in a hospital, haemoglobin levels of patients, marks of learners in MST-016 paper, the diameters of ball bearings produced by a company, etc.

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

### Sample

In general, collecting the information from all units of a large population is time-consuming and costly. Also, if the units of a population are destroyed under investigation, then gathering the information from all the units does not make sense. For example, test the blood of a patient, test the quality of crackers, etc. In such situations, a small part of the population is selected from the population which is called a **sample**. Thus, we can define a sample as

**“A sample is a part/fraction/subset of a population.”**

For example, a syringe full of blood taken from the vein of a patient is a sample of all blood in the patient's circulation at the moment. Similarly, a group of 26 learners of the MSCAST is a sample of the population of learners of the programme. The number of units selected in the sample is known as sample size and it is denoted by **n**. It is extremely important to choose a sample that is truly representative of the population so that the inferences derived from the sample can be generalised back to the population of interest.

### Sample Mean and Sample Variance

If  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  taken from a population, then the sample mean and is defined as

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

And sample variance is defined as

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

Here, we divide  $\sum_{i=1}^n (X_i - \bar{X})^2$  by  $(n - 1)$  rather than  $n$  as our definition of the

variance. The reason for taking  $(n - 1)$  in place of  $n$  will become clear in Unit 6 of this course.

Statistical inference is a technique of drawing conclusions about the population from data (sample). It is based on the assumption that the sample should be random. We have various techniques to select a random sample. In MST-013: Survey Sampling and Design of Experiments-I, you have studied various sampling methods so that we can maintain the randomness. In statistical inference, we use simple random sampling because for randomness it is sufficient. Let us look at this sampling.

Sampling is the statistical process of selecting a subset (called a “sample”) of a population of interest. In other words, the procedure of drawing a sample from the population is called **sampling**.

## Simple Random Sampling

Simple random sampling is the simplest and most common method of sampling used in statistical inference. In simple random sampling, the sample is drawn in such a way that each 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 a random sample**. The simple random sample of size  $n$  is denoted by  $X_1, X_2, \dots, X_n$  or  $Y_1, Y_2, \dots, Y_n$  and the observed values of this sample are denoted by  $x_1, x_2, \dots, x_n$  or  $y_1, y_2, \dots, y_n$ . Some authors use  $x_1, x_2, \dots, x_n$  to represent the random sample instead of  $X_1, X_2, \dots, X_n$ . But throughout the course, we use capital letters to represent the random sample, that is,  $X_1, X_2, \dots, X_n$  or  $Y_1, Y_2, \dots, Y_n$ .

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

### 1. Simple Random Sampling without Replacement (SRSWOR)

In simple random sampling, if the units are selected one by one in such a way that a unit drawn at a time is not replaced back to the population before the subsequent draws, is called **SRSWOR**. In this sampling, the same unit cannot appear more than once in the sample.

### 2. Simple Random Sampling with Replacement (SRSWR)

In simple random sampling, if the 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 sampling, the same element or unit can appear more than once in the sample.

Now a question may arise, **if we draw a sample of a specified (given) size from a population then how many samples are possible?** To understand the answer to this, we should know two things:

- First, whether the order of the unit in the sample matters or not, and
- Second, whether the drawn unit in the sample is replaced before the subsequent draw or not.

Based on these, the following four cases may arise:

- Ordered sampling with replacement
- Unordered sampling with replacement
- Ordered sampling without replacement
- Unordered sampling without replacement

To see the effect of ordering and replacement on the number of samples, we considered an example in which the size of the population is very small.

Suppose the statistical discipline of a university consisting of four typists (A, B, C and D) constitute a population. We select a sample of two typists from this population of typists for typing a manuscript. We now discuss how many samples may be possible in different cases.

### Ordered Sampling with Replacement

In this case,

- **Order matters.** The order of selecting the typists in the sample is

important. It means that if we select typist A first and then B (i.e. AB) differs from selecting B first and then A (i.e. BA) and each of these samples is regarded as a different possible sample that can be selected from this group of typists.

- **We replace the unit.** A unit drawn at a time is replaced back to the population before the subsequent draw. With replacement means a sample of selecting the typist A and then A again (AA) is possible.

In this case, the total number of possible samples will be 16 which are listed below:

|                        |                        |
|------------------------|------------------------|
| AA (repetition allows) | CA(order matters)      |
| AB                     | CB(order matters)      |
| AC                     | CC (repetition allows) |
| AD                     | CD                     |
| BA (order matters)     | DA(order matters)      |
| BB (repetition allows) | DB(order matters)      |
| BC                     | DC(order matters)      |
| BD                     | DD (repetition allows) |

For listing all possible samples, we list samples systematically. First, we list all of the possible samples with the first element of the population i.e. A as the first typist, then all of the possible samples with the second element of the population i.e. B, C and so on. In this way, we can be sure that we have all of the possible random samples.

In this case, we can determine the total number of possible samples of any size (n) that can be selected from a population of any size (N) using the following formula:

**Total number of possible samples =  $N^n$ .**

The ordered sampling with replacement is also known as **theoretical sampling** and it is used to develop the theories of sampling distribution. This case is also known as **simple random sampling with replacement (SRSWR)**.

### Unordered Sampling with Replacement

In this case,

- **Order does not matter.** The order of selecting the typists in the sample is not important. It means that if we select typist A first and then B (i.e. AB) is the same as selecting B first and then A (i.e. BA). These samples are considered as one and are not considered separate samples.
- **We replace the unit.** A unit drawn at a time is replaced back to the population before the subsequent draw. With replacement means a sample of selecting the typist A and then A again (AA) is possible.

In this case, the total number of samples will be 10 which are listed as follows:

|   |                                       |
|---|---------------------------------------|
| AA (repetition allows)                      | <del>CA</del> (order does not matter) |
| AB  | <del>CB</del> (order does not matter) |
| AC  | CC (repetition allows)                |
| AD  | CD                                    |
| <del>BA</del> (order does not matter AB=BA) | <del>DA</del> (order does not matter) |
| BB (repetition allows)                      | <del>DB</del> (order does not matter) |
| BC  | <del>DC</del> (order does not matter) |
| BD  | DD (repetition allows)                |

We can determine the total number of samples when order does not matter, and repetition is allowed using the following formula:

$$\text{Total number of possible samples} = {}^{N+n-1}C_n = {}^{N+n-1}C_{N-1}.$$

### Ordered Sampling without Replacement

In this case,

- **Order matters.** The order of selecting the typists in the sample is important. It means that if we select typist A first and then B (i.e. AB) differs from selecting B first and then A (i.e. BA) and each of these samples is regarded as a different possible sample that can be selected from this group of typists.
- **We do not replace the unit.** A unit drawn at a time is not replaced back to the population before the subsequent draw. This means that the same participant can never be sampled twice. So, the samples of AA, BB, CC and DD from the population in this example are not possible.

In this case, the total number of possible samples will be 12 which are listed as follows:

|   |   |
|---|---|
| <del>AA</del> (repetition does not allow) | CA (order matters)                        |
| AB  | CB (order matters)                        |
| AC  | <del>CC</del> (repetition is not allowed) |
| AD  | CD  |
| BA (order matters)                        | DA (order matters)                        |
| <del>BB</del> (repetition is not allowed) | DB (order matters)                        |
| BC  | DC (order matters)                        |
| BD  | <del>DD</del> (repetition is not allowed) |

We can determine the total number of possible samples when order matters, and repetition is not allowed using the following formula:

$$\text{Total number of possible samples} = N \times (N - 1) \times (N - 2) \times \dots \times (N - n + 1).$$

### Unordered Sampling without Replacement

In this case,

- **Order does not matter.** The order of selecting the typists in the sample is not important (order does not matter). It means that if we select typist A first and then B (i.e. AB) is the same as selecting B first and then A (i.e. BA). These samples are considered as one and are not considered separate samples.
- **We do not replace the unit.** A unit drawn at a time is not replaced back to the population before the subsequent draw. This means that the same participant can never be sampled twice. So, the samples of AA, BB, CC and DD from the population in this example are not possible.

In this case, the total number of possible samples will be 6 which are listed as follows:

|   |   |
|---|---|
| <del>AA</del> (repetition is not allowed) | <del>CA</del> (order does not matter)     |
| AB  | <del>CB</del> (order does not matter)     |
| AC  | <del>CC</del> (repetition is not allowed) |
| AD  | CD  |
| <del>BA</del> (order does not matter)     | <del>DA</del> (order does not matter)     |
| <del>BB</del> (repetition is not allowed) | <del>DB</del> (order does not matter)     |
| BC  | <del>DC</del> (order does not matter)     |
| BD  | <del>DB</del> (repetition is not allowed) |

We can determine the total number of samples when order does not matter, and repetition is not allowed using the following formula:

**Total number of possible samples =  ${}^N C_n$ .**

We may select diverse samples. But in practice, order does not matter because we do not care about the order in which units are selected. Also, we usually do not allow one individual to be chosen twice. Therefore, we often do the unordered sampling without replacement. Therefore, it is also called **experimental sampling**. This is commonly called **simple random sampling without replacement (SRSWOR)**. However, in statistical inference ordered simple random sampling with replacement (commonly called SRSWR) is used to develop the theory of statistical inference. On the other hand, in a large population, the probability of selecting the same individual twice is negligible, and it can be demonstrated that the results obtained from sampling with replacement are very close to the results obtained using sampling without replacement. The main advantage of sampling with replacement is that the sample observation will be independent, and this simplifies the analysis.

After understanding the various situations, let us take an example for illustration purposes.

**Example 1:** There are five sales associates at a car showroom. The number of cars they sold last week is as follows:

| Sales Associate | Cars Sold |
|-----------------|-----------|
| Vihaan          | 2         |
| Rohan           | 5         |
| Ritika          | 4         |
| Hassan          | 6         |
| Anita           | 10        |

A researcher wants to select a sample of 2 associates. How many samples of size 2 are possible with replacement? Also, write them.

**Solution:** Here, we are given that

Population size =  $N = 5$ , Sample size =  $n = 2$

Since we know that all possible samples of size  $n$  taken from a population of size  $N$  with replacement (order matter and replacement allow) are  $N^n$ .

Therefore, possible samples in our case  $N^n = 5^2 = 25$ . These 25 samples are given in Table 1.1 along with the car sold.

Table 1.1: Possible Samples of Sales Associates with Cars Sold

| Sample Number | Sample in Term of Associates | Sample Observations (car sold) | Sample Number | Sample in Term of Associates | Sample Observations (car sold) |
|---------------|------------------------------|--------------------------------|---------------|------------------------------|--------------------------------|
| 1             | (Vihaan, Vihaan)             | (2, 2)                         | 14            | (Ritika, Hassan)             | (4, 6)                         |
| 2             | (Vihaan, Rohan)              | (2, 5)                         | 15            | (Ritika, Anita)              | (4, 10)                        |
| 3             | (Vihaan, Ritika)             | (2, 4)                         | 16            | (Hassan, Vihaan)             | (6, 2)                         |
| 4             | (Vihaan, Hassan)             | (2, 6)                         | 17            | (Hassan, Rohan)              | (6, 5)                         |
| 5             | (Vihaan, Anita)              | (2, 10)                        | 18            | (Hassan, Ritika)             | (6, 4)                         |
| 6             | (Rohan, Vihaan)              | (5, 2)                         | 19            | (Hassan, Hassan)             | (6, 6)                         |
| 7             | (Rohan, Rohan)               | (5, 5)                         | 20            | (Hassan, Anita)              | (6, 10)                        |
| 8             | (Rohan, Ritika)              | (5, 4)                         | 21            | (Anita, Vihaan)              | (10, 2)                        |
| 9             | (Rohan, Hassan)              | (5, 6)                         | 22            | (Anita, Rohan)               | (10, 5)                        |
| 10            | (Rohan, Anita)               | (5, 10)                        | 23            | (Anita, Ritika)              | (10, 4)                        |
| 11            | (Ritika, Vihaan)             | (4, 2)                         | 24            | (Anita, Hassan)              | (10, 6)                        |
| 12            | (Ritika, Rohan)              | (4, 5)                         | 25            | (Anita, Anita)               | (10, 10)                       |
| 13            | (Ritika, Ritika)             | (4, 4)                         |               |                              |                                |

For listing all possible samples, we list samples systematically. First, we list all of the possible samples with the first element of the population i.e. A as the first typist, then all of the possible samples with the second element of the population i.e. B, C and so on. In this way, we can be sure that we have all of the possible random samples.

Let us try the following Self Assessment Questions to check your understanding of the number of possible samples.

### SAQ 1

A hospital administrator wishes to estimate the mean weight of babies born in her hospital. She collects the birth records of a day and the observes weights (in pounds) of 4 babies ( $B_1, B_2, B_3, B_4$ ) as 6, 8, 7, and 6 pounds. How many samples of size 2 are possible with replacement? Also, write them.

After understanding the concept of population, sample and how many samples are possible in different cases, let us move to the next concept.

### Parameter

The characteristics of a population can be described with some measures such as population mean, variance, etc. These measures are known as parameters of the population. Thus, we can define a parameter as:

**“A parameter or population parameter is a numerical value that summarises or measures or represents a specific characteristic of an entire population.”**

The parameters are derived from data collected from the entire population and taken as fixed constants.

For example, suppose the course coordinator of the MST-016 course calculates the average marks of all the learners in the MST-016 course then the obtained average mark is a parameter because it is based on all learners of the MST-016 course. Similarly, population variance, population coefficient of

variation, population correlation coefficient, etc. are all parameters. Population parameters are usually denoted by Greek letters, such as the population mean and variance are represented by Greek letters  $\mu$  and  $\sigma^2$ , respectively.

Generally, the parameters of a population are typically unknown and are estimated from sample data.

We know that the population of measurements such as height, marks, etc can be described with the help of distribution such as normal, Poisson, etc. and the distribution is fully determined with the help of its constants such as, in case of a normal distribution, we need to know  $\mu$  and  $\sigma^2$  to determine the normal distribution, in case of Poisson distribution, we need to know  $\lambda$ , etc. These constants are also known as the parameters.

### Statistic

As a parameter describes the characteristic of the population in a similar way a statistic describes the characteristic of the sample. We can define a statistic as:

**“A sample statistic or statistic is a numerical measure that summarizes or describes a characteristic of a sample.”**

A statistic is calculated using sample data. For example, suppose the course coordinator of the MST-016 course calculates the average marks of the learners in the MST-016 course by selecting some learners randomly instead of all learners then the obtained sample average mark is a statistic because it is based on the sample of the learners of the course. Similarly, sample variance, sample coefficient of variation, sample correlation coefficient, etc. are all statistics (plural of statistic). The statistic is usually denoted by Roman/Latin letters, such as the sample mean and variance are represented by  $\bar{X}$  and  $S^2$ , respectively. Generally, the statistic is used to estimate the population parameter. We may also define a statistic as

**Any quantity calculated from sample values that does not contain any unknown parameter is known as a statistic.**

For example, if  $X_1, X_2, \dots, X_n$  is a random sample of size  $n$  taken from a population with mean  $\mu$  and variance  $\sigma^2$  (both are unknown) then the sample

mean  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is a statistic whereas  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$  and  $\bar{X} / \sigma$  are not statistics

because both are functions of unknown parameters. If both  $\mu$  and  $\sigma^2$  are known then  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$  or  $\bar{X} / \sigma$  become a statistic because both  $\mu$  and  $\sigma^2$  are

known and are treated as constants. We use different symbols for parameters and statistics as follows:

|                    | Population parameter            | Sample statistic           |
|--------------------|---------------------------------|----------------------------|
| Mean               | $\mu$ (Greek letter “mu”)       | $\bar{X}$ (called “X-bar”) |
| Standard Deviation | $\sigma$ (Greek letter “sigma”) | S (Latin letter “S”)       |
| Variance           | $\sigma^2$                      | $S^2$                      |
| Proportion         | P                               | p                          |

## Estimator and Estimate

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

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

A statistic that is used to estimate an unknown parameter is known as an **estimator**. An estimator is a function of the sample observations. Common estimators are the sample mean and sample variance which are used to estimate the unknown population mean and variance, respectively. The estimator itself is a random variable because it is a function of the random sample observations. Its value varies from sample to sample due to the randomness inherent in the sampling process. A particular value of the estimator based on the observed value of the sample is known as an **estimate** of the parameter.

For example, let us suppose that a pharmacologist wants to test a new systolic blood pressure (SBP) medicine. For that, he selected a random sample of 100 SBP participants to try the medicine and prescribe the medicine for one month. After one month, he measured the SBP of the participants. Then, he calculated the mean of the SBP measurements. Suppose the mean SBP is found as 125 mm Hg. Then sample mean is called the estimator, and this number (125 mm Hg) allows him to estimate that a more general population is likely to maintain 125 mm Hg SPB while taking the medicine is an estimate. Consider another example, suppose we want to estimate the average height of students in a college. If we estimate the average height by selecting some students randomly from the college, then the sample average is the estimator and its particular value, say, 165 cm is the estimate of the average height of all students in the college.

In this course, we use capital letters for the estimator and small letters for the estimated value. For example, we represent the estimator sample mean by  $\bar{X}$  and its particular value as  $\bar{x}$  as an estimate. In general, the estimator is denoted by  $T_n = t(X_1, X_2, \dots, X_n)$  where 'n' denotes the sample size and the estimator T is a function of the random sample  $X_1, X_2, \dots, X_n$ .

Now, you check your understanding of the above discussion by answering the following Self Assessment Question.

### SAQ 2

If a quality control inspector takes a random sample of 10 ball bearings from the process of manufacturing the ball bearings and measures the internal diameter of these. The obtained results (in inches) are as follows:

29, 31, 30, 32, 30, 29, 30, 30, 29, 30

- (i) Obtain the standard deviation of the variation in the diameter.
- (ii) Find the estimator and estimate in this case.

After understanding the concept of basic terminology, you are now ready to learn the main concept of this unit, that is, sampling distribution. Let us discuss it in the next section.

## 1.3 INTRODUCTION TO SAMPLING DISTRIBUTION

As we have discussed in Section 1.1, if the population is too large or the units or items of the population are destructive or there are limited resources such as manpower, money, etc., then it is not possible practically to examine every unit of the population to obtain the necessary information or characteristics/parameters of the population. For example, suppose we want to know the average life of electric LED bulbs which are manufactured by a company. The company manufactures a lot of bulbs say 5,000 per day. Then gathering information about the average life of all bulbs does not make 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 person is very much time and manpower consuming. In such situations, one can draw a 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) valid conclusions about the population. This technique is known as **statistical inference**. We can define it as:

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

For drawing inferences about the population, we analyse the sample data, that is, we calculate the value of a sample statistic such as the sample mean, sample proportion, sample variance, etc. Generally, if we want to draw inferences about the population characteristic such as population mean, we use the sample mean, about population variance, we use sample variance, etc. For example, suppose a researcher of health science wants to know the average cholesterol levels of the persons living in a city. For practical reasons, he cannot reach out to each and every person in the city. So, he randomly selected 10 persons (Sample-I) from the city and measured their cholesterol levels. The observed values of the cholesterol levels are given in column 2 of Table 1.2.

**Table 1.2: Cholesterol Levels of 10 Individuals in Sample-I and Sample-II**

| S. No.             | Cholesterol Level (in mg/dl) |             |
|--------------------|------------------------------|-------------|
|                    | Sample-I                     | Sample-II   |
| 1                  | 180                          | 200         |
| 2                  | 200                          | 200         |
| 3                  | 190                          | 180         |
| 4                  | 220                          | 200         |
| 5                  | 180                          | 190         |
| 6                  | 190                          | 180         |
| 7                  | 220                          | 240         |
| 8                  | 200                          | 220         |
| 9                  | 190                          | 200         |
| 10                 | 180                          | 210         |
| <b>Total</b>       | <b>1950</b>                  | <b>2020</b> |
| <b>Sample Mean</b> | <b>195</b>                   | <b>202</b>  |

From the table, the average (mean) cholesterol level of these selected persons is 195 mg/dl. Now, if you use this sample to make an inference about the population's average cholesterol level then we say that the sample average cholesterol level of 195 mg/dl is an estimate of the average cholesterol level of the persons living in this city (population). However, we do not know the precision (the estimate is accurate or not) of the estimate because the actual cholesterol level of the persons in the city is unknown.

If another random sample of 10 persons (Sample-II) is selected and measured their cholesterol levels (shown in column 3 of Table 1.2) then we get the average (mean) cholesterol level of these selected persons is 202 mg/dl. This is different from Sample-I because it contains different persons who have different cholesterol levels. Thus, if you use this sample mean to estimate the population mean then you get a different estimate of the average cholesterol level for the whole population. If other samples of 10 persons are selected, it is unlikely that exactly the same mean would be observed, therefore, we may expect different estimates of the population mean every time. Now, the following questions may arise:

- How well does a sample describe its population?
- How can we tell which sample gives the best estimate of the population parameter?
- What is the probability of selecting a sample with specific characteristics?

These questions can be answered once we establish the **sampling distribution** of the statistic such as mean, variance, proportion, etc. The sampling distribution of a statistic is important because it enables us to draw conclusions about the corresponding population parameter.

For a better understanding of the concept of the sampling distribution, we consider a population of very small size so that we can easily obtain unknown population parameters, say, mean ( $\mu$ ) and standard deviation ( $\sigma$ ) using all population observations. And see how the sampling distribution helps us to draw conclusions about the population.

Consider a small production industry which has five employees. The monthly salary (in thousands) of each employee is given as follows:

**Table 1.3: Monthly Salaries of Employees**

| Employee | Monthly Salary (in thousands) |
|----------|-------------------------------|
| Lavnik   | 25                            |
| Avishi   | 30                            |
| Aman     | 15                            |
| Tanishq  | 25                            |
| Harsh    | 10                            |

We can calculate the mean of the monthly salary (population) as

$$\mu = \frac{25 + 30 + 15 + 25 + 10}{5} = 21$$

Similarly, we can obtain the standard deviation of the monthly salary (population) as

$$\begin{aligned}\sigma &= \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2 = \sqrt{\frac{(25-21)^2 + (30-21)^2 + (15-21)^2 + (25-21)^2 + (10-21)^2}{5}} \\ &= \sqrt{\frac{16 + 81 + 36 + 16 + 121}{5}} = \sqrt{54} = 7.35\end{aligned}$$

We can plot the graph of the population (monthly salary) to know the form of the population as follows in Fig. 1.1.

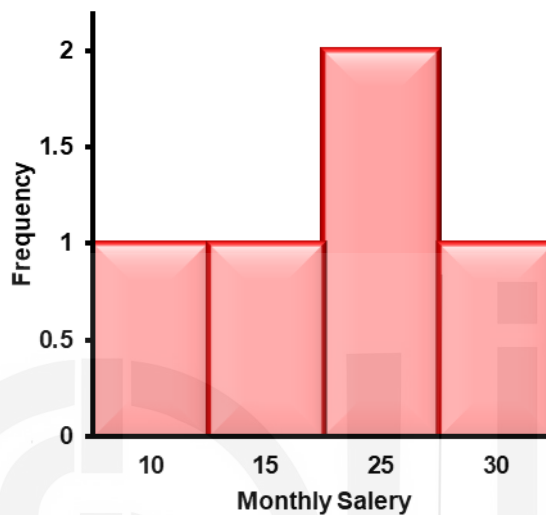


Fig. 1.1: Monthly salary of the employees of a small industry.

From the above figure, we observe that the monthly salary of the employees does not follow a known distribution (especially normal) and it is left-skewed.

Let us assume that we do not know the average salary of the employees. So we decide to estimate the population mean on the basis of a sample of size  $n = 2$ . As you have studied in Section 1.2, there are  $N^n = 5^2 = 25$  possible simple random samples with replacement of size 2. All possible samples of size  $n = 2$  are given in the second column of Table 1.4. We also calculate the means of each sample which are given in the last column of the same table.

Table 1.4: Samples and Sample Means

| Sample number | Sample in Term of Employees | Sample Observation (monthly salary) | Sample Mean |
|---------------|-----------------------------|-------------------------------------|-------------|
| 1             | (Lavnik, Lavnik)            | (25, 25)                            | 25          |
| 2             | (Lavnik, Avishi)            | (25, 30)                            | 27.5        |
| 3             | (Lavnik, Aman)              | (25, 15)                            | 20          |
| 4             | (Lavnik, Tanishq)           | (25, 25)                            | 25          |
| 5             | (Lavnik, Harsh)             | (25, 10)                            | 17.5        |
| 6             | (Avishi, Lavnik)            | (30, 25)                            | 27.5        |
| 7             | (Avishi, Avishi)            | (30, 30)                            | 30          |
| 8             | (Avishi, Aman)              | (30, 15)                            | 22.5        |
| 9             | (Avishi, Tanishq)           | (30, 25)                            | 27.5        |
| 10            | (Avishi, Harsh)             | (30, 10)                            | 20          |

| Sample | Sample in Term of Employees | Sample Observation (monthly salary) | Sample Mean |
|--------|-----------------------------|-------------------------------------|-------------|
| 11     | (Aman, Lavnik)              | (15, 25)                            | 20          |
| 12     | (Aman, Avishi)              | (15, 30)                            | 22.5        |
| 13     | (Aman, Aman)                | (15, 15)                            | 15          |
| 14     | (Aman, Tanishq)             | (15, 25)                            | 20          |
| 15     | (Aman, Harsh)               | (15, 10)                            | 12.5        |
| 16     | (Tanishq, Lavnik)           | (25, 25)                            | 25          |
| 17     | (Tanishq, Avishi)           | (25, 30)                            | 27.5        |
| 18     | (Tanishq, Aman)             | (25, 15)                            | 20          |
| 19     | (Tanishq, Tanishq)          | (25, 25)                            | 25          |
| 20     | (Tanishq, Harsh)            | (25, 10)                            | 17.5        |
| 21     | (Harsh, Lavnik)             | (10, 25)                            | 17.5        |
| 22     | (Harsh, Avishi)             | (10, 30)                            | 20          |
| 23     | (Harsh, Aman)               | (10, 15)                            | 12.5        |
| 24     | (Harsh, Tanishq)            | (10, 25)                            | 17.5        |
| 25     | (Harsh, Harsh)              | (10, 10)                            | 10          |

From the above table, you can see that the value of the sample statistic (sample mean) varies from sample to sample and out of 25 samples there is no sample whose mean is equal to the population mean. It means that out of 25 samples, no one estimates the population mean exactly. So, what do we do now? How can we estimate it? The sampling distribution helps us in such situations. We now try to understand the concept of it.

Since the sample mean varies from sample to sample, therefore, it is treated as a random variable. As you have studied in MST-012, a random variable has a distribution, so the sample statistic (sample mean) has a distribution. To obtain the distribution of a statistic (sample mean) we arrange the values of it in ascending or descending order and calculate the frequency of each value as shown in Table 1.5. We can also obtain the probability distribution (using the relative frequency approach of probability) described in MST-012 of the occurrence of each value which is given in the last column of Table 1.5.

**Table 1.5: Sampling Distribution of Sample Means**

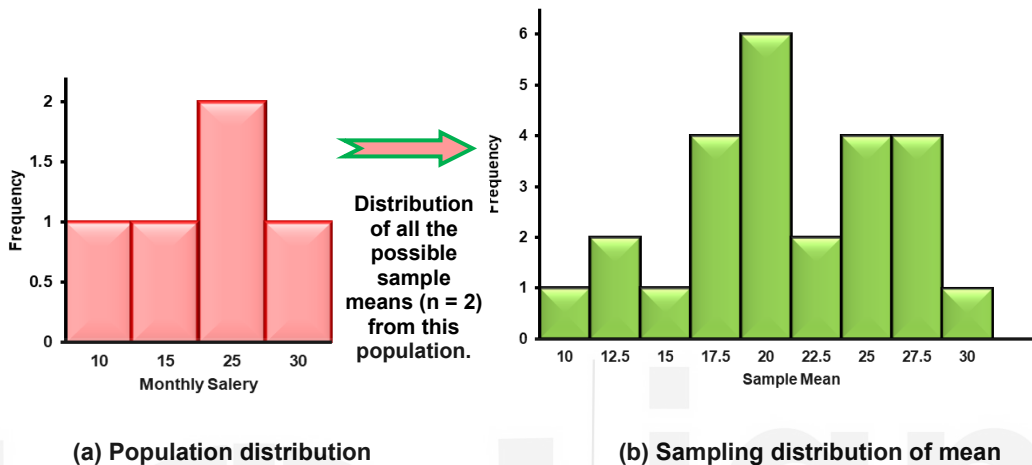
| S. No. | $\bar{X}$ | Frequency(f) | Probability(p) |
|--------|-----------|--------------|----------------|
| 1      | 10        | 1            | $1/25 = 0.04$  |
| 2      | 12.5      | 2            | $2/25 = 0.08$  |
| 3      | 15        | 1            | $1/25 = 0.04$  |
| 4      | 17.5      | 4            | $4/25 = 0.16$  |
| 5      | 20        | 6            | $6/25 = 0.24$  |
| 6      | 22.5      | 2            | $2/25 = 0.08$  |
| 7      | 25        | 4            | $4/25 = 0.16$  |
| 8      | 27.5      | 4            | $4/25 = 0.16$  |
| 9      | 30        | 1            | $1/25 = 0.04$  |

We can also construct the sampling distribution of the sample median instead of mean in the same way but the sampling distribution median is generally not normal.

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

**“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 the sampling distribution of that statistic.”**

To get an idea of the shape of the sampling distribution of mean, we plot the graph (frequency bar) of the values of the sample mean taking the sample mean on the X-axis and corresponding frequencies on the Y-axis as shown in Fig. 1.2(b).



(a) Population distribution

(b) Sampling distribution of mean

**Fig. 1.2: Population distribution and sampling distribution of mean for  $n = 2$ .**

From Fig. 1.2, we observe that the shape of the sampling distribution has been changed.

As the probability distribution such as normal, Poisson, binomial, etc. allows us to gain an understanding of the summary (mean, standard deviation, etc), likelihood and probabilities of different values occurring in the outcome, in a similar way, a sampling distribution is a probability distribution that does the same job. The sampling distribution of the sample means itself has mean, variance, etc. Therefore, we can calculate the mean of sample means as

$$\begin{aligned} \text{Mean of the sample means} &= (\bar{\bar{X}}) = \frac{1}{K} \sum_{i=1}^k \bar{X}_i f_i \quad \text{where, } K = \sum_{i=1}^k f_i \\ &= \frac{1}{25} (10 \times 1 + 12.5 \times 2 + \dots + 30 \times 1) = 21 = \mu \end{aligned}$$

We can also calculate the mean of the sample means as

$$E(\bar{X}) = \bar{\bar{X}} = \sum_{i=1}^k \bar{X}_i p_i = 10 \times \frac{1}{25} + 12.5 \times \frac{2}{25} + \dots + 30 \times \frac{1}{25} = 21$$

You saw! The mean of the sampling distribution of means is equal to the population mean. This is what we wanted. However, we did not get the same using a single sample. We find it as the mean of the sampling distribution of means.

Thus, we can say that to find the exact estimate of the unknown population parameter, we first find the sampling distribution of the corresponding statistic and then compute the mean of the obtained sampling distribution of that statistic.

We now calculate the standard deviation of the sampling distribution of mean as

$$SD(\bar{X}) = \sqrt{\frac{1}{K} \sum_{i=1}^k f_i (\bar{X}_i - \bar{\bar{X}})^2}$$

$$SD(\bar{X}) = \sqrt{\frac{1}{25} [1 \times (10 - 21)^2 + 2 \times (12.5 - 21)^2 + \dots + 1 \times (30 - 21)^2]}$$

$$= \sqrt{\frac{1}{25} (121 + 144.5 + \dots + 121)} = 5.20$$

$$SD(\bar{X}) = 5.20$$

We now compare the distribution of the population with the sampling distribution of mean. In Fig. 1.2, we plot them.

Let us now see what happens as we increase the sample size. So we decide to estimate the population mean on the basis of a sample of size  $n = 3$ . In this case, there are  $N^n = 5^3 = 125$  possible simple random samples with replacement of size 3. We can list all possible samples of size  $n = 3$  and for each sample, we calculate the sample means which are shown in Table 1.6.

Table 1.6: Samples and Sample Means

| Sample number | Sample in Term of Employees | Sample Observation (monthly salary) | Sample Mean |
|---------------|-----------------------------|-------------------------------------|-------------|
| 1             | (Lavnik, Lavnik, Lavnik)    | (25, 25, 25)                        | 25          |
| 2             | (Lavnik, Lavnik, Avishi)    | (25, 25, 30)                        | 26.67       |
| 3             | (Lavnik, Lavnik, Aman)      | (25, 25, 15)                        | 21.67       |
| 4             | (Lavnik, Lavnik, Tanishq)   | (25, 25, 25)                        | 25          |
| 5             | (Lavnik, Lavnik, Harsh)     | (25, 25, 10)                        | 20          |
| 6             | (Lavnik, Avishi, Lavnik)    | (25, 30, 25)                        | 26.67       |
| 7             | (Lavnik, Avishi, Avishi)    | (25, 30, 30)                        | 28.33       |
| 8             | (Lavnik, Avishi, Aman)      | (25, 30, 15)                        | 23.33       |
| 9             | (Lavnik, Avishi, Tanishq)   | (25, 30, 25)                        | 26.67       |
| 10            | (Lavnik, Avishi, Harsh)     | (25, 30, 10)                        | 21.66       |
| ...           | ...                         | ...                                 | ...         |
| 125           | (Harsh, Harsh, Harsh)       | (10, 10, 10)                        | 10          |

We now prepare the sampling distribution of mean as in the case when  $n = 2$ .

| $\bar{X}$ | Frequency(f) | Probability(p)   |
|-----------|--------------|------------------|
| 10        | 1            | $1/125 = 0.008$  |
| 11.67     | 3            | $3/125 = 0.024$  |
| 13.33     | 3            | $3/125 = 0.024$  |
| 15        | 7            | $7/125 = 0.056$  |
| 16.67     | 15           | $15/125 = 0.12$  |
| 18.33     | 12           | $12/125 = 0.096$ |
| 20        | 15           | $15/125 = 0.12$  |
| 21.67     | 24           | $24/125 = 0.192$ |
| 23.33     | 15           | $15/125 = 0.12$  |

|       |    |                  |
|-------|----|------------------|
| 25    | 11 | $11/125 = 0.088$ |
| 26.67 | 12 | $12/125 = 0.096$ |
| 28.33 | 6  | $6/125 = 0.048$  |
| 30    | 1  | $1/125 = 0.008$  |

We now calculate the mean of the sample distribution of mean for  $n = 3$  as

$$\begin{aligned} \text{Mean of the sample means} &= (\bar{\bar{X}}) = \frac{1}{K} \sum_{i=1}^k \bar{X}_i f_i \\ &= \frac{1}{125} (10 \times 1 + 11.67 \times 3 + \dots + 30 \times 1) = 21 = \mu \end{aligned}$$

Similarly, we calculate the standard deviation of the sampling distribution of mean when  $n = 3$  as

$$\begin{aligned} \text{SD}(\bar{X}) &= \sqrt{\frac{1}{K} \sum_{i=1}^k f_i (\bar{X}_i - \bar{\bar{X}})^2} \\ \text{SD}(\bar{X}) &= \sqrt{\frac{1}{125} [1 \times (10 - 21)^2 + 3 \times (11.67 - 21)^2 + \dots + 1 \times (30 - 21)^2]} = 4.24 \end{aligned}$$

We now plot the sampling distribution of mean when the sample size  $n = 3$  in Fig. 1.3.

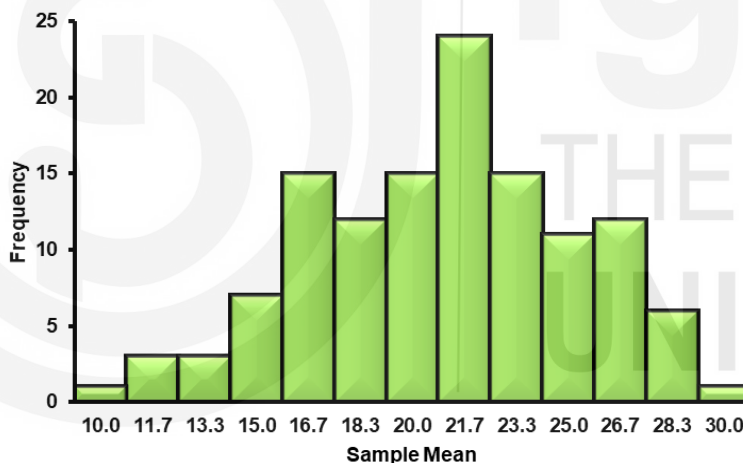


Fig. 1.3: Sampling distribution of mean when sample size  $n = 3$

From Fig. 1.3, you can observe that it is similar to the normal distribution.

From the above discussion, we observe some features of the sampling distribution:

- The shape of the sampling distribution of mean may be quite different from the shape of the population.
- The sample means “**pile up**” around the population mean and “**tail off**” towards the extremes. For this example, the population mean is  $\mu = 21$ , and the sample means are clustered around a value of 21. It should not surprise you that the sample mean tends to approximate the population mean.
- The sampling distribution of sample mean tends to bell-shaped normal probability distribution as sample size  $n$  increases.

- The standard deviation of the sampling distribution of mean decreases as sample size increases for  $n = 2$  it is 5.20 and for  $n = 3$  it is 4.24.

### Factors that influence sampling distribution

The form of sampling distribution depends on the following factors:

- the distribution of the population,
- the statistic being considered,
- the sampling procedure employed, and
- the sample size used.

In the above example, you saw that when the population size is 5 there are only 25 and 125 samples of size  $n = 2$  and 3, respectively. In more realistic situations, when the population is large then the number of possible samples increases dramatically. For example, when the population has 1,000 units then there will be  $(1000)^2 = 1000000$  possible simple random samples of size 2 and it is virtually impossible to obtain every possible random sample. In such situations, we draw a random sample from the population to draw inferences about the population parameters and use the concept of theoretical sampling distribution which are already developed.

It is now time for you to try the following Self Assessment Question to make sure that you understand the sampling distribution.

### SAQ 3

A Municipal Corporation Office has four typists. An officer gave the same sample page of a manuscript to all four typists. The number of errors made by each typist is shown in Table 1.7 as given below:

Table 1.7: Number of Errors per Typist

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

Answer the following questions.

- (i) Is the distribution of population values normally distributed (bell-shaped)?
- (ii) What is the population mean and standard deviation?
- (iii) How many samples of size  $n = 2$  are possible with replacement?
- (iv) List all possible samples of 2 from the population and compute their means.
- (v) Organize the means into a sampling distribution.
- (vi) Does the distribution of the sample mean computed in part (v) show some tendency towards being bell-shaped?
- (vii) What is the mean of the sampling distribution? What observations can

be made about the population and the sampling distribution?

- (viii) Compare the dispersion (SD) in the population with that in the distribution of the sample mean.

## 1.4 CONCEPT OF STANDARD ERROR

In Section 1.3, we describe the sampling distribution of a statistic (mean) and observe that the mean of the sampling distribution of mean is equal to the population mean. But in the real world, the population size is too large, and you see that when the population size is large then the number of possible samples increases dramatically, and it is virtually impossible to actually obtain every possible random sample and then observe the sampling distribution. In such situations, we draw a random sample from the population to draw inferences about the population parameters and use the concept of theoretical sampling distribution. For example, consider the example of estimating the average cholesterol levels of the persons living in a city. Due to cost and time constraints, the researcher selected 10 individuals randomly from the same city and obtained the average (mean) cholesterol level of these selected individuals was 195 mg/dl. Can you expect that it is somewhere close to or equal to the average cholesterol level of the whole population? I think the answer is no. Then the question may arise:

- How can we judge that an estimate observed from a single sample is a reliable estimate of the population parameter?
- How can know the precision (the estimate is accurate or not) of the estimate when it is not possible to compute the exact value?

The standard error does such jobs for us. It measures how far the sample mean (average) of the data is likely to be from the true population mean. The standard error provides a measure of how much distance is expected on average between a sample mean and the population mean. It gives the accuracy of a sample mean by measuring the sample-to-sample variability of the sample means. We now formally define the standard error (SE) as

**“The standard deviation of a sampling distribution of a statistic (mean, proportion, standard deviation) is known as standard error”.**

The standard error is not denoted by any symbol whereas it is denoted by its abbreviation SE.

The standard error is a standard deviation of the sampling distribution, so it serves the same two purposes for the sampling distribution as the standard deviation of the data which are listed as follows:

1. The standard error provides a measure of how much discrepancy is expected from one sample to another. When the standard error is small, then all of the sample means are close together and have similar values. If the standard error is large, then the sample means are scattered over a wide range and there are major differences from one sample to another.
2. Standard error measures how well an individual sample mean represents the entire population. Specifically, it gives an indication of the reasonable deviation that can be expected between a sample mean and the

Standard deviation and standard error of the mean are both statistical measures of variability. While the standard deviation of a sample depicts the spread of observations within the given sample regardless of the population mean where as the standard error of the mean measures the degree of dispersion of sample means around the population mean.

population mean.

The standard error is an extremely valuable measure because it specifies precisely how well a sample mean estimates its population mean, that is, how much error you should expect between the sample mean and population mean. However, you do not expect a sample to provide a perfectly accurate picture of the population. There is always some discrepancy, or error, between a sample statistic and the corresponding population parameter. But with the help of standard error, we are able to calculate exactly how much error to expect.

By the definition of the standard error, “the standard deviation of a sampling distribution of a statistic is known as standard error”. In the previous section, we calculate the standard deviation of the sampling distribution of the sample mean as follows:

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

$$\begin{aligned} SE(\bar{X}) &= \sqrt{\frac{1}{25} [1 \times (10 - 21)^2 + 2 \times (12.5 - 21)^2 + \dots + 1 \times (30 - 21)^2]} \\ &= \sqrt{\frac{1}{25} (121 + 144.5 + \dots + 121)} = 5.20 \end{aligned}$$

Therefore, it is the standard error of the mean.

The computation of the standard error using the sampling distribution is a tedious process. Therefore, there is an alternative method to compute the standard error of the mean from a single sample. We can calculate the magnitude of the standard error using the following formula:

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

We can calculate the standard error of the mean using the standard deviation of the population (the amount of variability among the units of the population) and the sample size.

But in real life, the population standard deviation is generally unknown. In such situations, it is more common to estimate the standard error by substituting the sample standard deviation in place of the population standard deviation in the formula for standard error. Therefore, the formula for standard error becomes as

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

where S is the standard deviation of the sample and we can calculate it using the following formula:

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

These are the expressions of the standard error of the mean. We give the expressions of standard error for other statistics such as proportion, standard deviation, etc. in the subsequent units.

Let us try one example relating to standard error.

**Example 2:** The diameter of a steel ball bearing produced by a semi-automatic machine is known to be distributed normally with a mean of 12 cm and a standard deviation of 0.1 cm. If we take a random sample of size 10 then find the standard error of the sample mean for estimating the population mean of the diameter of all ball bearings produced.

**Solution:** Here, we are given that

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

Since, population standard deviation ( $\sigma$ ) is given, therefore, we can calculate the standard error of the sample mean as

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

We now discuss the applications of standard error.

### Applications of Standard Error

Standard error plays a very crucial role:

1. The standard error is used to find the accuracy or precision or reliability of the sample estimate of a population parameter. The precision of a sample estimate can be defined as the reciprocal of the standard error of the statistic. We can calculate the precision as

$$\text{Precision} = \frac{1}{\text{Standard error of the estimate}}$$

2. It is used to construct confidence intervals (CI) within which the population parameter may be expected to lie with a certain level of confidence. The standard error also determines the probable limits or confidence limits. We will discuss confidence intervals in more detail in Units 12, 13 and 14.
3. The standard error is also used to test whether the difference between the sample statistic and the population parameter is significant or is due to sampling fluctuations. It means that the standard error is also applicable in the testing of hypothesis. We will discuss testing of hypothesis in more detail in the last block of this course.

Since the precision is reciprocal to the standard error, we try to decrease it.

### Factors that Decrease Standard Error

The formula for calculating of standard error of the sample mean is given as follows:

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

From the above formula, the magnitude of the standard error depends on two factors:

1. The standard deviation of the population from which the sample is selected; and
2. The size of the sample.

We examine each of these factors one at a time.

### Effect of population standard deviation

The standard error depends directly on the population standard deviation. Therefore, as the population standard deviation decreases, the standard error also decreases. It means that if the population units deviate from the population mean then the less possibility that the sample means will deviate from the population mean. To visualise the effect of SD of the population, we consider different populations having SD as shown in the table given below and we select samples of the same size 9 ( $n = 9$ ) from each population. We compute the standard error for each sample as follows:

| Sample Size | Standard Deviation of the Population ( $\sigma$ ) | Standard Error |
|-------------|---|----------------|
| 9           | 1   | 0.33           |
| 9           | 4   | 1.33           |
| 9           | 9   | 3              |
| 9           | 16  | 5.33           |
| 9           | 25  | 8.33           |
| 9           | 36  | 12             |
| 9           | 49  | 16.33          |
| 9           | 64  | 21.33          |
| 9           | 81  | 27             |
| 9           | 100   | 33.33          |

We now plot the standard error corresponding to the standard deviation of the population as shown in Fig. 1.4.

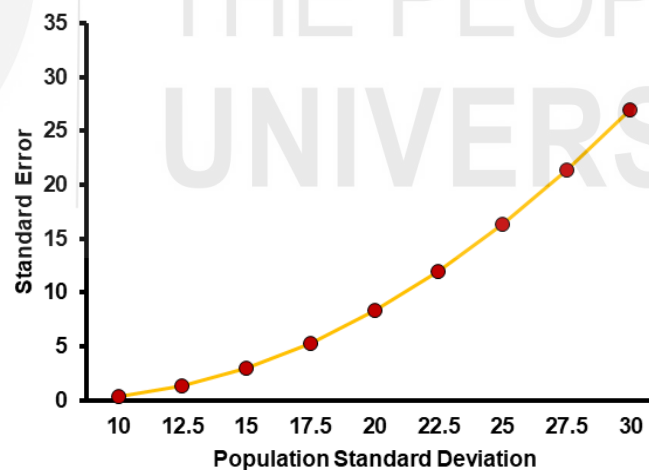


Fig. 1.4: The standard error with the population standard deviation.

From the above figure, we observe that as the standard deviation of the population decreases, the standard error or the average distance of the sample mean deviating from the population mean also decreases.

### Effect of sample size (n)

As we noted in the previous section, as we increase the sample size  $n$  from 2 to 3 then standard error decreases. Also, from the formula for calculating standard error, we observe that there is an inverse relationship between the sample size and the standard error. As we increase the sample size, the

standard error decreases. It means that if we collect more data in a sample, our estimate of the population mean will be more accurate. To illustrate the general relationship between standard error and sample size, we visualize the effect of the sample size in Fig. 1.5. For that, we select samples of different sizes from a single population with a standard deviation equal to 9 and compute the standard error for each sample as follows:

| Sample Size | Standard Deviation of Population ( $\sigma$ ) | Standard Error |
|-------------|---|----------------|
| 1           | 9   | 9              |
| 4           | 9   | 4.5            |
| 9           | 9   | 3              |
| 16          | 9   | 2.25           |
| 25          | 9   | 1.8            |
| 36          | 9   | 1.5            |
| 49          | 9   | 1.3            |
| 64          | 9   | 1.13           |
| 81          | 9   | 1              |
| 100         | 9   | 0.9            |

We now plot the standard error corresponding to the sample size as shown in Fig. 1.5.

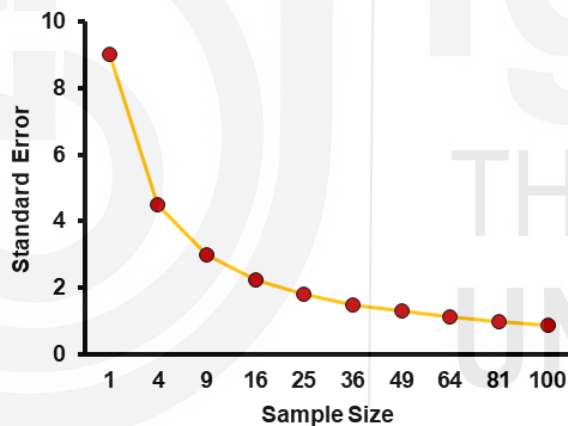


Fig. 1.5: The standard error with the sample size.

From the above figure, we observe that the standard error of the mean will approach zero with the increasing number of observations in the sample. It happens because when the sample size becomes large then the sample becomes more and more representative of the population, and the sample mean approaches the actual population mean.

**Note:** The 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 comparison to the size of the sample and sampling is done without replacement then we multiply the above standard errors by the correction

$$\text{factor } \sqrt{\frac{N-n}{N-1}}.$$

where  $N$  is the population size. This correction factor is known as a **finite population correction factor**.

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

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

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.

You should try the following Self Assessment Question before studying next section.

### SAQ 4

A juice company packages the juice in 500 mL pouches. Suppose you are incharge of monitoring that pouches that they are being filled correctly and you randomly select a sample of 25 pouches from the thousands of pouches filled during a shift. Given that the standard deviation of the juice-packaging process is 12 mL,

- (i) Calculate the standard error of the sample mean.
- (ii) If you select a sample of 100 pouches, what will be the standard error?
- (iii) What will be the size of the standard error of the mean as the sample size is increased from 25 to 100?

We now introduce you to some of the most powerful data laws, principles, and rules that are based on statistical theories. These will help you become a better data scientist or data analyst in general. The central limit theorem and law of large numbers are the foundation of statistical inference. Understanding these fundamental concepts can help you analyse and interpret data more effectively, enabling you to make confident decisions based on solid evidence. In the coming session, we will discuss the central limit theorem and in the next session, the law of large numbers.

## 1.5 CENTRAL LIMIT THEOREM

In the previous sections, you saw that when the population size is large then the number of possible samples increases dramatically, and it is virtually impossible to actually obtain every possible random sample. In such situations a question may arise, is it possible to determine exactly what the sampling distribution of the sample means without considering all samples? The answer to the question is given by the central limit theorem. It gives the form of the sampling distribution of mean under some conditions.

The first concept of the central limit theorem was given by French Mathematician Abraham De Moivre in 1733. At that time, the central limit theory proposed by De Moivre was not popular. But another well-known French mathematician, Pierre-Simon Laplace, revived the idea in 1812. The central limit theorem discoveries made by Laplace at that time caught the interest of numerous academics and thinkers. The central limit theorem was extended later in 1901 by Russian mathematician Aleksandr Lyapunov. He



(1667-1754)

Abraham De Moivre was a French mathematician. He was the first who gave the central limit theorem.

disproved the idea in general and provided mathematical evidence for its validity.

The central limit theorem is one of the most important theorems of Statistics. It states that

**The sampling distribution of the mean approaches to a normal distribution as the size of the sample increases, regardless of the shape of the original population distribution.**

We can explain the central limit theorem as

If a random sample of size  $n$  is taken from a population with mean  $\mu$  and finite variance  $\sigma^2$  then the sampling distribution of the sample mean tends to a normal distribution with mean  $\mu$  and variance  $\sigma^2/n$  as the sample size tends to be large ( $n \geq 30$ ) whatever may be the form of the parent population, that is,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ when } n \geq 30$$

We do not intend to prove this theorem here but merely show graphical evidence of its validity in Fig. 1.6. Here, we will also try to show how large must the sample size be for which we can apply the central limit theorem.

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 present this figure in four parts A, B, C and D. Part 'A' of this figure shows four different populations as normal, uniform, binomial and exponential.

The rest of the parts B, C and D represent the shape of the sampling distributions of mean for the sample size  $n = 2$ ,  $n = 5$  and  $n = 30$ , respectively drawn from the populations shown in the first row (Part-A).

From the first column of this figure, we observed that when the parent population is normal then all sampling distributions of mean for varying sample sizes  $n = 2$ ,  $5$  and  $30$  are also normal, having the 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 mean is symmetrical and does not follow any standard distribution when  $n = 2$  and tends to be normal when  $n = 30$ .

However, the third column of this figure represents the binomial population (discrete). Again, when  $n = 2$ , the sampling distribution of mean is symmetrical and for  $n = 5$  it is quite bell-shaped and tends to be normal when  $n = 30$ .

The last column of this figure represents the exponential population which is highly skewed. Here, we also observe that when the sample size  $n = 2$  then the sampling distribution of mean does not follow any standard distribution but for  $n = 30$ , the distribution of mean is symmetrical bell-shaped normal.

From Fig. 1.6, we also observe that the rate at which the distribution approaches a normal distribution depends on the shape of the population. If the population itself is normally distributed, the sampling distribution of mean is also normal for any sample size  $n$ , as stated earlier. On the other hand, for

We can apply the central limit theorem to almost all types of probability distributions, but the population must have a finite variance. That restriction rules out the Cauchy distribution because it has infinite variance.

population distributions that are very different from a normal distribution, a relatively large sample size is required to achieve a good normal approximation for the distribution.

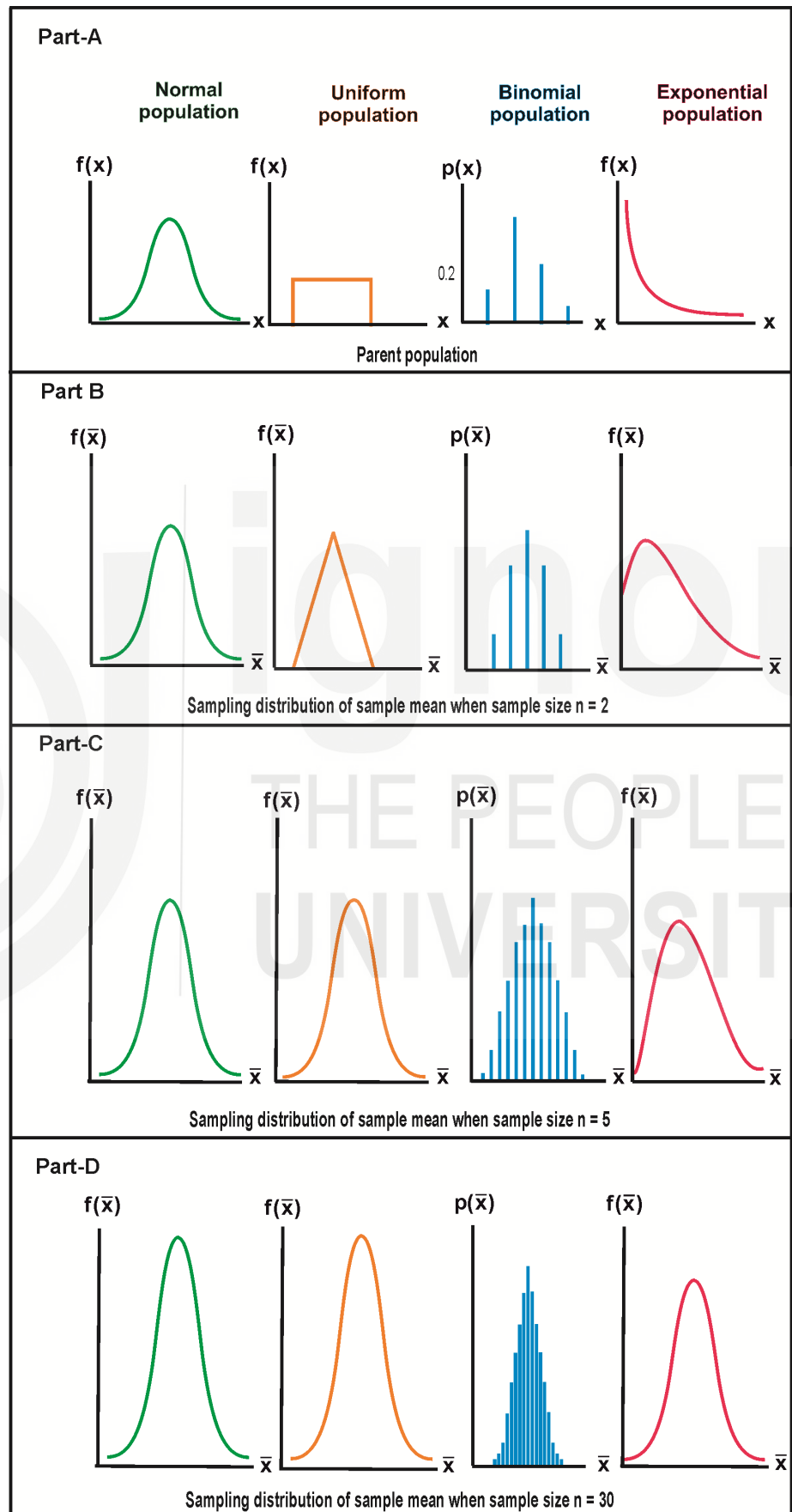


Fig. 1.6: Sampling distribution of sample means for various populations when  $n = 2$ ,  $n = 5$  and  $n = 30$ .

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

From the above discussion, we draw the following results:

1. When the population/study variable is normally distributed, then the sampling distribution of mean will be normally distributed, for any sample size  $n$ .
2. When the distribution of the population/study 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.
3. If the population distribution is fairly symmetrical, the sampling distribution of the mean is approximately normal for samples as small as 5.

Let us see the applications of the central limit theorem using an example.

**Example 3:** The average breaking strength of a certain brand of steel cable is 2500 pounds, with a standard deviation of 160 pounds. A sample of 40 cables is randomly selected and tested. What is the sampling distribution of mean? Also, find the probability that the average breaking strength of the sample cables:

- (i) more than 2550 pounds
- (ii) less than 2480 pounds
- (iii) between 2450 pounds and 2550 pounds

**Solution:** Here, we are given that

The average breaking strength of the steel cable is 2500 pounds with a standard deviation of 160 pounds. These are either population or sample information but till that no sample is taken so it is the mean of the whole steel cables (population), therefore,

$$\mu = 2500 \text{ pounds, } \sigma = 160 \text{ pounds and } n = 40$$

To find out the required probability, we require the probability distribution as you have seen in MST-012 but the distribution of the breaking strength of the steel cable is not given however the sample size is 40 large ( $n > 30$ ).

Therefore to find the sampling distribution, we use the central limit theorem which tells us the distribution of the same. According to the central limit theorem, when the sample size is large ( $n \geq 30$ ), the sampling distribution of the sample means will follow a normal distribution with mean

$$E(\bar{X}) = \mu = 2500$$

and variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{(160)^2}{40} = 640$$

Therefore, the sampling distribution of the average breaking strength of the steel cables is normally distributed with a mean of 2500 pounds and a variance of 640 pounds<sup>2</sup>.

- (i) The probability that the average breaking strength of the sample cables is more than 2550 pounds is given by

$$P[\bar{X} > 2550] \quad [\text{see Fig. 1.7}]$$

To get the value of this probability, we convert the sample mean  $\bar{X}$  into a standard form. Since the sampling distribution of the average breaking strength ( $\bar{X}$ ) of the steel cables is normally distributed with mean of 2500 pounds and a variance of 640 pounds<sup>2</sup>, therefore, we transform the sample mean ( $\bar{X}$ ) into the standard normal Z-score as

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - 2500}{\sqrt{640}} = \frac{\bar{X} - 2500}{25.30}$$

Therefore, subtract 2500 from each term and then divide each term by 25.30 in the above inequality. Thus, the probability expression becomes

$$\begin{aligned} P\left[\frac{\bar{X} - 2500}{25.30} > \frac{2550 - 2500}{25.30}\right] &= P[Z > 1.98] \\ &= 1 - P[Z \leq 1.98] \end{aligned}$$

From the standard normal table (Table IV) given in the Appendix of this volume, we get

$$P[\bar{X} > 2550] = 1 - P[Z \leq 1.98] = 1 - 0.9761 = 0.0239$$

It means that 2.39% of steel cables have an average breaking strength of more than 2550 pounds.

- (ii) We can obtain the probability that the average breaking strength of the sample cables is less than 2480 pounds as

$$\begin{aligned} P[\bar{X} < 2480] &= P\left[\frac{\bar{X} - 2500}{25.30} < \frac{2480 - 2500}{25.30}\right] \\ &= P[Z < -0.79] \quad [\text{see Fig. 1.8}] \\ &= 0.2148 \quad [\text{using Table III}] \end{aligned}$$

It means that 21.48% of the steel cables have an average breaking strength less than 2480 pounds.

- (iii) We can compute the probability that the average breaking strength of the sample cables lies between 2450 pounds and 2550 pounds as

$$\begin{aligned} P[2450 < \bar{X} < 2550] &= P\left[\frac{2450 - 2500}{25.30} < \frac{\bar{X} - 2500}{25.30} < \frac{2550 - 2500}{25.30}\right] \\ &= P[-1.98 < Z < 1.98] \quad [\text{see Fig. 1.9}] \\ &= P[Z < 1.98] - P[Z < -1.98] \\ &= 0.9761 - 0.0239 = 0.9522 \end{aligned}$$

It means that 95.22% of the steel cables have an average breaking strength between 2450 pounds and 2550 pounds.

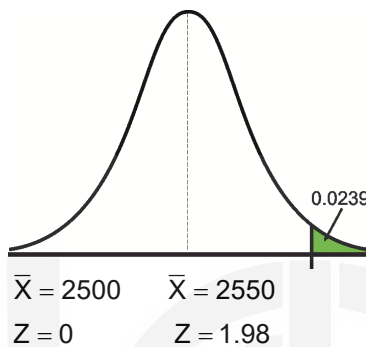


Fig. 1.7

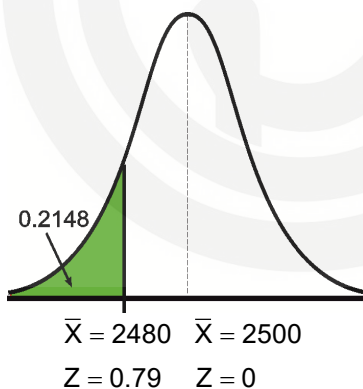


Fig. 1.8

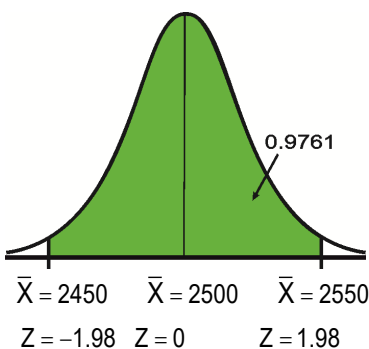


Fig. 1.9

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

### SAQ 5

A survey of a city found that the mean number of days per month that people who suffer from migraine headaches is 12.4 days with a standard deviation 3 days. Find the probability that if a random sample of 50 people who suffer from migraine headaches is selected, the mean of the sample will be between 12 and 13 days.

## 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 the 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 prepare the sampling distribution of that statistic which helps us to draw inferences about the population parameters.

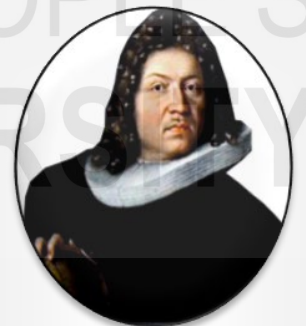
But in the real-world the population size is too large and you see that when the population size is large then the number of possible samples increases dramatically and it is virtually impossible to actually obtain every possible random sample and then observe the sampling distributions. In such situations, we use the concept of theoretical sampling distribution such as the central limit theorem and we draw a random sample from the population to draw inferences about the population parameters. A very crucial question then arises: **“Using a random sample of finite size, can we make a reliable inference about population parameters?”** The answer is “yes”, reliable inference about population parameters can be made by using only a finite sample and we shall demonstrate this by **“law of large numbers”**. The law of large numbers is a powerful principle that helps statisticians and analysts make sense of data. After understanding how it works and how it applies to real-world situations, you can make better decisions and draw more accurate conclusions from your data.

The Law of large numbers was initially known as **“the Golden Theorem”** or **“Bernoulli’s theorem”**. It was first coined by the Swiss mathematician **Jacob Bernoulli** in 1713. The theorem later became known as the **“Law of large numbers”**. There is also a more general version of the law of large numbers for averages, proved more than a century later by the Russian mathematician **Pafnuty Chebyshev**.

This law of large numbers can be stated in words as:

**“The law of large numbers states that as a sample size increases, its mean gets closer to the average of the whole population. In other words, as the sample size increases, the average of the observed results will become more and more representative of the true parameter value.”**

This is the same intuition behind the idea that if we collect more data, our sample of data will be more representative of the population.



(1654-1705)

Jacob Bernoulli (also known as James' or Jacques) was a Swiss mathematician. He mainly contributed to Analytic geometry, Probability theory, Variable calculus.

However, his most important contribution was in the field of probability, where he derived the first version of the law of large numbers in his work **Ars Conjectandi**.

Here, try to understand the law of large numbers in action with the help of an example.

Suppose the distribution of the weight of all the young men living in a city is close to a normal distribution with a mean weight of 65 kg and a standard deviation of 5 kg. To understand the law of large numbers, we calculate the sample mean weight ( $\bar{X}$ ) for varying sample sizes  $n = 1, 2, 3, \dots$ . Fig. 1.10 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 the vertical axis and sample size along the horizontal axis as sample size varying  $n = 1, 2, 3, \dots$

First, we start with a sample of size  $n = 1$ , that is, we select a man randomly from the city. Suppose the selected man has a weight of 70 kg, therefore, the line of the graph starts from this point. We now select the second man randomly and suppose his weight is 62 kg. So for  $n = 2$ , the sample mean is

$$\bar{X} = \frac{70 + 62}{2} = 66 \text{ kg}$$

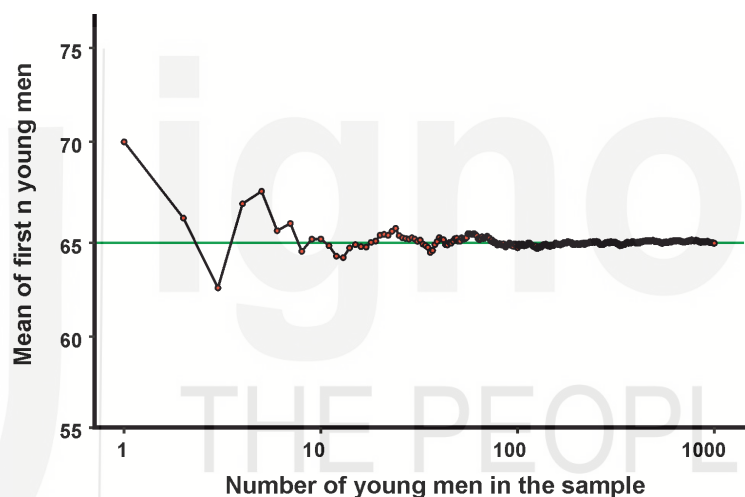


Fig. 1.10: Law of Large numbers

This is the second point on the graph. Now, we select the third man randomly from the city and suppose his weight is 55 kg. Therefore, for  $n = 3$ , the sample mean is

$$\bar{X} = \frac{70 + 62 + 55}{3} = 62.33 \text{ kg}$$

This is the third point on the graph.

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 gets closer and closer to the population mean of 65 kg as the sample size increases.

Hence from Fig. 1.10, we can observe that every additional data point gathered has the potential to move the sample to the true population mean.

We use the law of large numbers in a variety of fields including statistics, market research, finance, healthcare, insurance, engineering, etc. Here we mention some examples of the use of the law of large numbers:

1. In market research, let us suppose you are working for a marketing company, and you want to know what percentage of people in a certain city

prefer your brand of mobile. It is not possible for you to survey every single person in the city, so you take a random sample of 100 people and ask them which brand of mobile they prefer. If you find that 40% of them prefer your brand. Does that mean that 40% of the entire population prefers your brand? Not necessarily there is always a chance that your sample was biased in some way.

However, if you were to take a larger sample, say, 10,000 people and found that 45% of them prefer your brand, that would be a stronger indication that a majority of the population really does prefer your brand. And if you were to take an even larger sample say, 100,000 people and found that 48% of them prefer your brand, that would be an even stronger indication that your brand is really popular in the city.

2. In financial analysis, we can also use the law of large numbers. For example, suppose you are a stock analyst and you want to know the average rate of return for a certain stock. You could calculate the rate of return for a small sample of investors and get a rough estimate, but it would not be very accurate. However, if you take a larger sample of investors to calculate the rate of return, you will get a more accurate estimate of the true rate of return.
3. Similarly, the law of large numbers is also used in medicine to study the effectiveness of treatments. If you want to know whether a certain medication is effective for a particular condition then you could study a small sample of patients and get a rough idea. However, if you study a larger sample of patients, you will get a more accurate estimate of the effectiveness of the medication.

After studies where we can use the law of large numbers, we should note the following important points:

- The law of large numbers does not tell that the sample mean will always reflect the true population characteristics, especially for small samples.
- If a given sample mean deviates from the true population mean, then the law of large numbers does not guarantee that successive samples will move the observed average towards the population mean.

I hope you have understood the concept of the law of large numbers and how you can apply it to real-world situations. Hence, you can make better decisions and draw more accurate conclusions from your data.

With the help of the law of large numbers, we can also determine the minimum sample size to get a reliable inference about the population.

This law states that for any two arbitrary small numbers  $\epsilon$  ( $\epsilon > 0$ ) and  $\eta$  ( $0 < \eta < 1$ ), there exists an integer  $n = \frac{\sigma^2}{\epsilon^2 \eta}$  such that if a random sample of size  $n$  or

larger is drawn from the population and the sample mean ( $\bar{X}$ ) is calculated for this sample, then the sample mean ( $\bar{X}$ ) arbitrarily close to the population mean ( $\mu$ ) where  $\sigma^2$  is the finite variance of the population. In statistical inference  $\epsilon$  is called margin of error and  $1 - \eta$  is called confidence level.

Hence, to get a reliable inference about the population, we can determine the

You have also studied the law of large numbers in Unit 18 of MST-012: Probability and Probability Distribution. Where you studied the same in the context of probability.

\minimum sample size with the help of the law of large numbers. For that, the minimum sample size should be  $n \geq \frac{\sigma^2}{\varepsilon^2 \eta}$ .

Let us take an example to understand the same.

**Example 4:** A hospital administrator wants to estimate the mean weight of babies born in his/her hospital. How large a sample of birth records should be taken if he/she wants a 99 per cent confidence that the estimate is within the range of 0.4 pounds? As per the available records, the population SD is 0.5 pounds.

**Solution:** Here, we are given that

$$\varepsilon = \text{margin of error} = 0.4, \text{ confidence level} = 0.99 \text{ and } \sigma = 0.5$$

Also, for 99% confidence,  $1 - \eta = 0.99 \Rightarrow \eta = 0.01$

We can calculate the minimum sample size for estimating the mean weight of babies born in her hospital as

$$\begin{aligned} n &\geq \frac{\sigma^2}{\varepsilon^2 \eta} = \frac{(0.5)^2}{(0.4)^2 \times 0.01} \\ &= 156.25 \sim 157 \end{aligned}$$

Hence, the hospital administrator should take a random sample of at least 157 babies.

Now, try the following Self Assessment Question for your practice.

---

### SAQ 6

A pathologist wants to estimate the mean time required to complete a certain analysis on the basis of a sample study so that he may be 99% confident that the mean time may remain within  $\pm 2$  days of the mean. As per the available records, the population standard deviation is 5 days. What must be the size of the sample for this study?

---

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:

- The statistical procedure which is used for drawing conclusions about the population parameter on the basis of the sample data is called “**statistical inference**”.
- A population is a group of measurements in the quantitative or qualitative form of the characteristic under study.
- A parameter or population parameter is a numerical value that summarises or measures or represents a specific characteristic of an entire population.
- A sample statistic or a statistic is a numerical measure that summarizes or describes a characteristic of a sample.
- 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 the parameter.

- 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 the “**sampling distribution**” of that statistic.
- The standard deviation of the sampling distribution of a statistic is known as “**standard error**”.
- The **central limit theorem** states that the sampling distribution of the sample means tends to a normal distribution as the sample size tends to be large ( $n > 30$ ).
- The **law of large numbers** states that as a sample size increases, the sample mean gets closer to the average of the whole population.

## 1.8 TERMINAL QUESTIONS

1. In the ball bearings question (SAQ 2), Estimate the standard error of the sample mean.
2. Describe the Central Limit theorem.

## 1.9 SOLUTIONS / ANSWERS

### Self Assessment Questions (SAQs)

1. Here, we are given that

Population size =  $N = 4$

Sample size =  $n = 2$

Since we know that all possible samples of size  $n$  taken from a population of size  $N$  with replacement (order and replacement allowed) are  $N^n$ . Therefore, possible samples in this case  $N^n = 4^2 = 16$ . These 16 samples are given in Table 1.8 along with the weights (in pounds) of the babies.

Table 1.8: Possible Samples of Babies

| Sample Number | Sample in Term of Associates       | Sample Observations (car sold) | Sample Number | Sample in Term of Associates       | Sample Observations (car sold) |
|---------------|------------------------------------|--------------------------------|---------------|------------------------------------|--------------------------------|
| 1             | (B <sub>1</sub> , B <sub>1</sub> ) | (6, 6)                         | 9             | (B <sub>3</sub> , B <sub>1</sub> ) | (7, 6)                         |
| 2             | (B <sub>1</sub> , B <sub>2</sub> ) | (6, 8)                         | 10            | (B <sub>3</sub> , B <sub>2</sub> ) | (7, 8)                         |
| 3             | (B <sub>1</sub> , B <sub>3</sub> ) | (6, 7)                         | 11            | (B <sub>3</sub> , B <sub>3</sub> ) | (7, 7)                         |
| 4             | (B <sub>1</sub> , B <sub>4</sub> ) | (6, 6)                         | 12            | (B <sub>3</sub> , B <sub>4</sub> ) | (7, 6)                         |
| 5             | (B <sub>2</sub> , B <sub>1</sub> ) | (8, 6)                         | 13            | (B <sub>4</sub> , B <sub>1</sub> ) | (6, 6)                         |
| 6             | (B <sub>2</sub> , B <sub>2</sub> ) | (8, 8)                         | 14            | (B <sub>4</sub> , B <sub>2</sub> ) | (6, 8)                         |
| 7             | (B <sub>2</sub> , B <sub>3</sub> ) | (8, 7)                         | 15            | (B <sub>4</sub> , B <sub>3</sub> ) | (6, 7)                         |
| 8             | (B <sub>2</sub> , B <sub>4</sub> ) | (8, 6)                         | 16            | (B <sub>4</sub> , B <sub>4</sub> ) | (6, 6)                         |

For listing all possible samples, we list samples systematically. First, we list all of the possible samples with the first element of the population i.e. B<sub>1</sub> as the first typist, then all of the possible samples with the second element of the population i.e. B<sub>2</sub>, B<sub>3</sub> and so on. In this way, we can be sure that we have all of the possible random

2. We know the formula of sample variance as

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

To calculate it, first, we have to calculate the sample mean ( $\bar{X}$ ). We calculate the sample mean as

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{29 + 31 + 30 + 32 + 30 + 29 + 30 + 30 + 29 + 30}{10}$$

$$\bar{X} = \frac{300}{10} = 30$$

We now calculate the sample standard deviation as

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

$$= \sqrt{\frac{(29-30)^2 + (31-30)^2 + (30-30)^2 + (32-30)^2 + (30-30)^2 + (29-30)^2 + (30-30)^2 + (30-30)^2 + (29-30)^2 + (30-30)^2}{9}}$$

$$= \sqrt{\frac{1+1+0+4+0+1+0+0+1+0}{9}} = \sqrt{\frac{8}{9}} = 0.94$$

In this case, we are estimating the population standard deviation from the sample standard deviation, therefore, the sample standard deviation is the estimator for the population standard deviation and 0.94 pounds is the estimated value of the population standard deviation.

3. First of all, we check the shape of the population. For that, we plot the population values (number of errors) as follows:

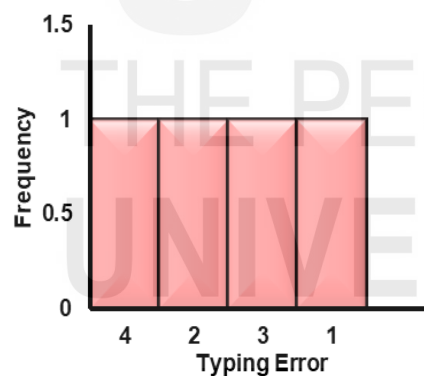


Fig. 1.11: Distribution of the population of typing error.

The above figure shows that it is a uniform distribution instead of bell-shaped (normal distribution).

We can calculate the population mean (average number of errors) as

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

Similarly, we can compute the standard deviation of the population as

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2} = \sqrt{\frac{(4-2.5)^2 + (2-2.5)^2 + (3-2.5)^2 + (1-2.5)^2}{4}} = 1.12$$

Here, the population size (N) is 4. Therefore, 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 1.9.

Table 1.9: Samples and Sample Means

| Sample Number | Sample in Terms 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                       |

To obtain the sampling distribution of all sample means, we arrange these values in ascending order and calculate the frequency of each value as shown in Table 1.10. We can also obtain the probability distribution using the relative frequency approach of probability in the last column of Table 1.10.

Table 1.10: Sampling Distribution of Sample Means

| $\bar{X}$ | Frequency(f) | Probability(p) |
|-----------|--------------|----------------|
| 1.0       | 1            | 1/16 = 0.0625  |
| 1.5       | 2            | 2/16 = 0.1250  |
| 2.0       | 3            | 3/16 = 0.1875  |
| 2.5       | 4            | 4/16 = 0.2500  |
| 3.0       | 3            | 3/16 = 0.1875  |
| 3.5       | 2            | 2/16 = 0.1250  |
| 4.0       | 1            | 1/16 = 0.0625  |

After finding the sampling distribution of mean, we check the shape of the sampling distribution of mean. For that, we put the sample means on the X-axis with the corresponding frequency on the Y-axis as shown in Fig. 1.12.

From the figure, we observe that it is a bell-shaped (normal) distribution.

If we compare the population distribution with the sampling distribution of mean then we observe that even though the shape of the population is uniform the sampling distribution of the mean is normal.

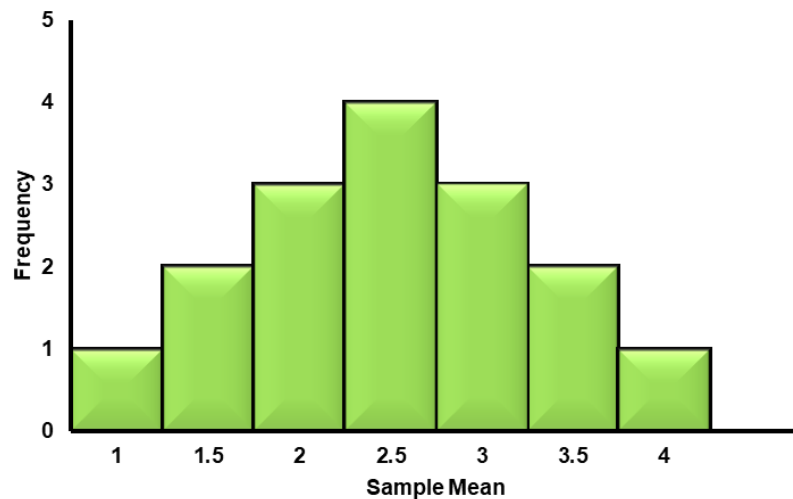
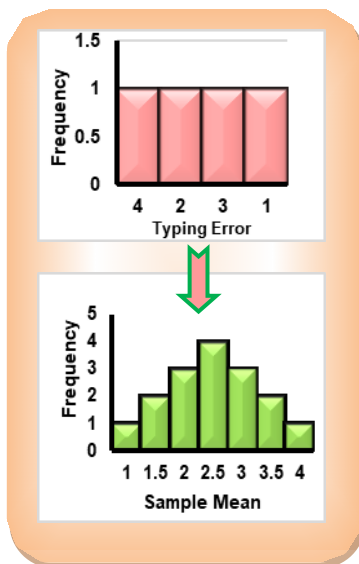


Fig. 1.12: Sampling distribution of mean.

The sampling distribution of the sample means itself has mean, variance, etc. Therefore, we can compute the mean of this distribution as

$$\begin{aligned}\text{Mean of samples means} &= \bar{\bar{X}} = \frac{1}{K} \sum_{i=1}^k \bar{X}_i f_i \text{ where, } K = \sum_{i=1}^k f_i \\ &= \frac{1}{16} (1.0 \times 1 + 1.5 \times 2 + \dots + 4.0 \times 1) = 2.5 = \mu\end{aligned}$$

Similarly, we can calculate the standard deviation of the sample distribution of mean as

$$\begin{aligned}\text{SD}(\bar{X}) &= \sqrt{\frac{1}{K} \sum_{i=1}^k f_i (\bar{X}_i - \mu)^2} \text{ where, } K = \sum_{i=1}^k f_i \\ &= \sqrt{\frac{1}{16} [1 \times (1.0 - 2.5)^2 + 2 \times (1.5 - 2.5)^2 + \dots + 1 \times (4.0 - 2.5)^2]} \\ &= \sqrt{\frac{1}{16} (2.25 + 2 + \dots + 2.25)} = \sqrt{\frac{10}{16}} = 0.791\end{aligned}$$

Hence, we can conclude that the mean of the sampling distribution of mean is the same as the population mean whereas the dispersion of the sampling distribution (0.791) is less in comparison to the population (1.12).

4. Here, we are given that

$$n = 25, \sigma = 12$$

Since, population standard deviation ( $\sigma$ ) is given, therefore, we can calculate the standard error of the sample mean as

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{25}} = 2.4$$

From the above result, we conclude that the variation in the sample means for samples of  $n = 25$  is much less than the variation in individual

pouches of juice ( $\sigma = 15$ ).

When we select a sample of 100 pouches, then the standard error will be

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$$

When we increased the sample size from 25 to 100, then we observed that the standard error became half. Thus, we conclude that if we want to half the standard error, we must increase the sample size 4 times.

5. Here, we are given that

The mean number of days per month that people who suffer from migraine headaches is 12.4 days with the standard deviation is 3 days. These are either population or sample information but till that, no sample is taken so it is the information of the whole group (population), therefore,

$$\mu = 12.4 \text{ days, } \sigma = 3 \text{ days and } n = 50$$

To find the required probability, we require the probability distribution but is not given however the sample size is 50 large sample ( $n > 30$ ). Therefore, to find the sampling distribution we use the central limit theorem. According to the central limit theorem, the sampling distribution of mean number of days per month that people who suffer from migraine headaches follows a normal distribution with mean

$$E(\bar{X}) = \mu = 12.4 \text{ days}$$

and variance

$$\begin{aligned} \text{Var}(\bar{X}) &= \frac{\sigma^2}{n} \\ &= \frac{(3)^2}{50} = 0.18 \end{aligned}$$

We have to find the probability that the sample mean will lie between 12 and 13 days is given by

$$P[12 < \bar{X} < 13] \quad [\text{see Fig. 1.13}]$$

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

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - 12.4}{\sqrt{0.18}} = \frac{\bar{X} - 12.4}{0.42}$$

Therefore, subtract 12.4 from each term and then divide each term by 0.42 in the above inequality. Thus, the probability expression becomes:

$$\begin{aligned} P\left[\frac{12 - 12.4}{0.42} < \frac{\bar{X} - 12.4}{0.42} < \frac{13 - 12.4}{0.42}\right] &= P[-0.95 < Z < 1.43] \\ &= P[Z < 1.43] - P[Z < -0.95] = 0.9236 - 0.1711 = 0.7525 \end{aligned}$$

It means that 75.25% of the mean of the sample will lie between 12 and

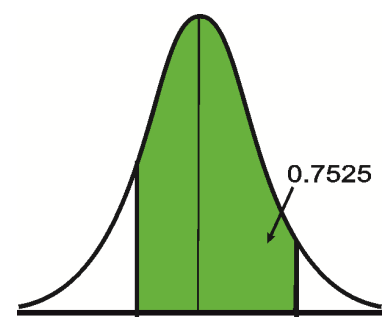


Fig. 1.13

13 days.

6. Here, we are given

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

By the law of large numbers, we can calculate the minimum size as

$$n \geq \frac{\sigma^2}{\epsilon^2 \eta} = \frac{5^2}{2^2 \times 0.01} = 625$$

That is,  $n \geq 625$

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

### **Terminal Questions (TQs)**

1. Here, we are given that

$$n = 10$$

Since, the population standard deviation ( $\sigma$ ) is not given, therefore, we can estimate the standard error of the sample mean using the formula given as follows:

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

To use the above formula, first, we have to calculate the sample standard deviation. But we have already calculated it in SAQ 2 as 0.94. Therefore, we can compute the estimate of standard error as

$$SE(\bar{X}) = \frac{S}{\sqrt{n}} = \frac{0.94}{\sqrt{10}} = 0.30$$

2. Refer to Section 1.5.