
UNIT 8 SAMPLING DISTRIBUTION

Structure

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Basic Terminology
- 8.3 Introduction to Sampling Distribution
- 8.4 Standard Error
- 8.5 Central Limit Theorem
- 8.6 Law of Large Numbers
- 8.7 Sampling Distribution of Sample Mean
- 8.8 Sampling Distribution of Difference of Two Sample Means
- 8.9 Summary
- 8.10 Solutions /Answers
- 8.11 References

8.0 INTRODUCTION

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

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

This unit introduces the basic terminology and sampling distribution used in statistical inference. One of the important concepts in Statistics is the standard error. It has been discussed in this unit. In addition, this unit discusses the concept of the law of large numbers and sample size.

One of the most important sample statistics that is used to draw a conclusion about the population mean is the sample mean. This unit discusses the sampling distribution of the mean and the sampling distribution of the difference of two sample means. The unit ends by providing a summary of what we have discussed in this unit and the solution to the check your progress questions.

8.0 OBJECTIVES

After studying this unit, you should be able to:

- define statistical inference;
- define the basic terms as population, sample, parameter, statistic, estimator, estimate, etc. used in statistical inference;

- explain the concept and need of sampling distribution;
- explain the concept of standard error;
- describe the most important theorem of Statistics, “Central Limit Theorem”;
- explain the concept of the law of large numbers and adequate sample size; and
- explain the sampling distribution of sample mean and difference of two sample means.

8.2 BASIC TERMINOLOGY

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

Population

In Statistics, the term population has a broader meaning. A population is the collection or group of individuals /items /units/ observations under study. For example, the collection of books in a library, the particles in a salt bag, the rivers in India, the students in a classroom, etc., are considered as populations in Statistics.

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

We may classify the population into two types as:

(1) Finite and Infinite Population

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

If a population contains an infinite (uncountable) number of units or observations, it is called an **infinite population**. For example, the population of particles in a salt bag, the population of stars in the sky, the population of hairs on a human body, etc. Theoretically, sometimes, populations that are too large are assumed to be infinite.

(2) Real and Hypothetical Population

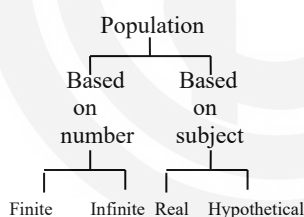
The population can also be classified as real and hypothetical. A population comprising the items or units which are all physically present is known as **real population**. All of the examples given for finite and infinite population are examples of a real population.

If a population consists of items or units which are not physically present but the existence of them can only be imagined or conceptualised then it is known as a **hypothetical population**. For example, the population of heads or tails in successive tosses of a coin a large number of times is considered as a hypothetical population.

Sample

To extract the information from all the elements or units or items of a large population may be, in general, time-consuming, expensive and difficult. And

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



also, if the elements or units of a population are destroyed under investigation, then gathering the information from all the units does not make sense. For example, to test the strength of the chalk, to test the quality of crackers, etc., the chalks and crackers are destroyed under the testing procedure. In such situations, a small part of the population is selected from the population, which is called a **sample**. Thus, the sample can be defined as given below:

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

The procedure for drawing a sample from the population is called **sampling**. A detailed discussion on sampling is beyond this course. You may learn about sampling from the references given in Section 8.11. The number of units selected in the sample is known as sample size and it is denoted by n .

Complete Survey and Sample Survey

Statistical data may be collected by two methods:

- (1) Complete Survey or Complete Enumeration or Census
- (2) Sample Survey or Sample Enumeration

(1) Complete Survey or Complete Enumeration or Census

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

(2) Sample Survey or Sample Enumeration

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

Simple Random Sampling or Random Sampling

The simplest and most common method of sampling is simple random sampling. In simple random sampling, the sample is drawn in such a way that each element or unit of the population has an equal and independent chance of being included in the sample. If a sample is drawn by this method, then it is known as a **simple random sample** or a **random sample**. The 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 value of this sample is denoted by $x_1, x_2, \dots, x_n$ or $y_1, y_2, \dots, y_n$. Some author 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, $X_1, X_2, \dots, X_n$. In addition, $X_1, X_2, \dots, X_n$ simply be assumed as the independent identically distributed (iid) random variables.

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

(1) Simple Random Sampling without Replacement (SRSWOR)

In simple random sampling, if the elements or units are selected or drawn one by one in such a way that an element or unit drawn at a time is not replaced back to

Parameter Estimation and Hypothesis Testing

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

the population before the subsequent draws, is called **SRSWOR**. If we draw a sample of size n from a population of size N without replacement then the total number of possible samples is ${}^N C_n$. For example, consider a population that consists of three elements, A, B and C. Suppose we wish to draw a random sample of two elements then $N = 3$ and $n = 2$. The total number of possible random samples without replacement is ${}^N C_n = {}^3 C_2 = 3$. The possible sample would be: (A, B), or (A, C) and (B, C).

(2) Simple Random Sampling with Replacement (SRSWR)

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

Note 1: The statistical inference is based on the theoretical distribution so, in whole statistical inference, we assume that the random sample is selected from an infinite population or a finite population with replacement so that removal of the sample unit has no appreciable effect on the composition of the population. That is, we draw the random sample such that:

- (a) The n successive sample observations are independent and
- (b) The population composition remains constant from draw to draw.

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

Parameter

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

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

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

Sample Mean and Sample Variance

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

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

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

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 described earlier. The reason for taking $(n-1)$ in place of n will be discussed later in the unit.

Statistic

Any quantity calculated from sample values and 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 μ and variance σ^2 (both are unknown) then the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is a statistic whereas $\bar{X} - \mu$ and $\bar{X}/\sigma$ are not statistics because both are functions of unknown parameters.

Estimator and Estimate

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

So if a statistic is used to estimate an unknown population parameter then it is known as an **estimator** and the value of the estimator based on the observed value of the sample is known as an **estimate** of the parameter. For example, suppose the parameter λ of the Poisson population $f(x, \lambda)$ is unknown so we draw a random sample $X_1, X_2, \dots, X_n$ of size n from this Poisson population to

estimate λ . If we use the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ to estimate λ then $\bar{X}$ is

called estimator and any particular value of this estimator, say, $\bar{x}$ is called the estimate of unknown parameter λ . Consider another example, if we want to estimate the average height (μ) of students in a college with the help of sample mean $\bar{X}$ then $\bar{X}$ is the estimator and its particular value, say, 165 cm is the estimate of the population average height (μ).

In this course, we use capital letters for the estimator and small letters for the estimated value. 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, it is the right place to discuss the sampling distribution.

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

8.3 INTRODUCTION TO SAMPLING DISTRIBUTION

Consider an example, if we want to know the average income of 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 utilise sample observations to extract the necessary information about the population. **The results obtained from the sample are projected in such a way that they are valid for the entire population.** Therefore, the sample works like a “**Vehicle**” to reach (drawing) at valid conclusions about the population. Thus, statistical inference can be defined as:

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

For drawing the inference about the population, we analyse the sample data, that is, we calculate the value of a sample statistic such as the sample mean, sample proportion, sample variance, etc. Generally, if we want to draw the inference about the population mean we use sample mean, about population variance we use sample variance, etc. However, if you take two independent samples from the population, then it is possible that mean of sample-I is different from mean of sample-II. Thus, if we are using this sample mean to estimate the population mean then we get a different estimate of the mean of population. Similarly, we can take many samples of the same size and for each sample, we get sample mean which may or may not be distinct. From all of the sample means, we try to estimate the mean of the whole population.

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

Consider a population comprising four typists who type the sample page of a manuscript. The number of errors made by each typist is shown in Table 8.1, given below:

Table 8.1: Number of Errors per Typist

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

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

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

Now, let us assume that we do not know the average number of errors made by typists. So we decide to estimate the population mean on the basis of a sample of size $n = 2$. There are $N^n = 4^2 = 16$ possible simple random samples with replacement of size 2. All possible samples of size $n = 2$ are given below and for each sample the sample mean is calculated as shown in Table 8.2.

From table 8.2, we can see that the value of the sample statistic (sample mean) is varying from sample to sample and out of 16 samples only 4 samples (4, 7, 10 and 13) have their means equal to the population mean whereas other 12 samples result in some error in the estimation process. The error is the difference between the population mean and the sample mean used for the estimate.

Table 8.2: Calculation of Sample Mean

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

Now, consider all possible values of the sample mean and calculate their probabilities of occurrence. Then we arrange every possible value of the sample mean with their respective probabilities in the following Table 8.3 given below:

Table 8.3: Sampling Distribution of Sample Means

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

So the arrangement of all possible values of the sample mean with their corresponding probabilities is called the sampling distribution of the mean.

Thus, we can define the sampling distribution of a statistic as:

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

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

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

The mean of sample means can also be calculated as

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

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

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

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

8.4 STANDARD ERROR

As we have seen in the previous section that the values of sample statistic may vary from sample to sample and all the sample values are not equal to the population parameter. Now, one can be interested to measure how much the values of sample statistic vary from the population parameter on average. You may use the standard deviation as a measure of variation.

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

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

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

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

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

The standard error is used to express the accuracy or precision of the estimate of population parameter because the reciprocal of the standard error is the measure of reliability or precision of the statistic. The standard error also determines the probable limits or confidence limits within which the population parameter may

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

be expected to lie with a certain level of confidence. The standard error is also applicable in the testing of hypothesis.

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

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

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

where P is the population proportion and $Q = 1-P$.

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

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

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

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

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

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

$$SE(p_1 - p_2) = \sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}$$

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

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

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

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

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

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

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

Similarly,

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

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

Let us do one example relating to standard error.

Example 1: Diameter of a steel ball bearing produced by a semi-automatic machine is known to be distributed normally with mean 12 cm and standard deviation 0.1 cm. If we take a random sample of size 10 with replacement then find the standard error of the sample mean for estimating the population mean of the diameter of steel ball bearing for the whole population.

Solution: Here, we are given that

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

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

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

Check Your Progress 1

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

.....

.....

.....

.....

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

.....

.....

.....

.....

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

.....

.....

.....

.....

8.5 CENTRAL LIMIT THEOREM

The central limit theorem is the most important theorem of Statistics. It was first introduced by De Moivre in the early eighteenth century. According to the central limit theorem, if $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a population with mean μ and variance σ^2 then the sampling distribution of the

sample mean tends to a normal distribution with mean μ and variance σ^2/n as the sample size tends to be large ($n > 30$), whatever may be the form of the parent population, that is:

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

and the variate

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

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

We do not intend to prove this theorem here, but merely show graphical evidence of its validity in Fig. 8.1. Here, we will also try to show how large must the sample size be for which we can assume that the central limit theorem applies?

In Fig. 8.1, we are trying to understand the sampling distribution of sample mean $\bar{X}$ for different populations and for varying sample sizes. We divide this figure into four parts A, B, C and D. The part 'A' of this figure shows four different populations as normal, uniform, binomial and exponential. The rest parts B, C and D represent the shape of the sampling distribution of mean of sizes $n = 2$, $n = 5$ and $n = 30$ respectively drawn from the populations shown in first row (Part-A). From the first column of this figure, we observed that when the parent population is normal then all the sampling distributions for varying sample sizes are also normal, having 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 $\bar{X}$ is symmetrical when $n = 2$ and tends to normal when $n = 30$.

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

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

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

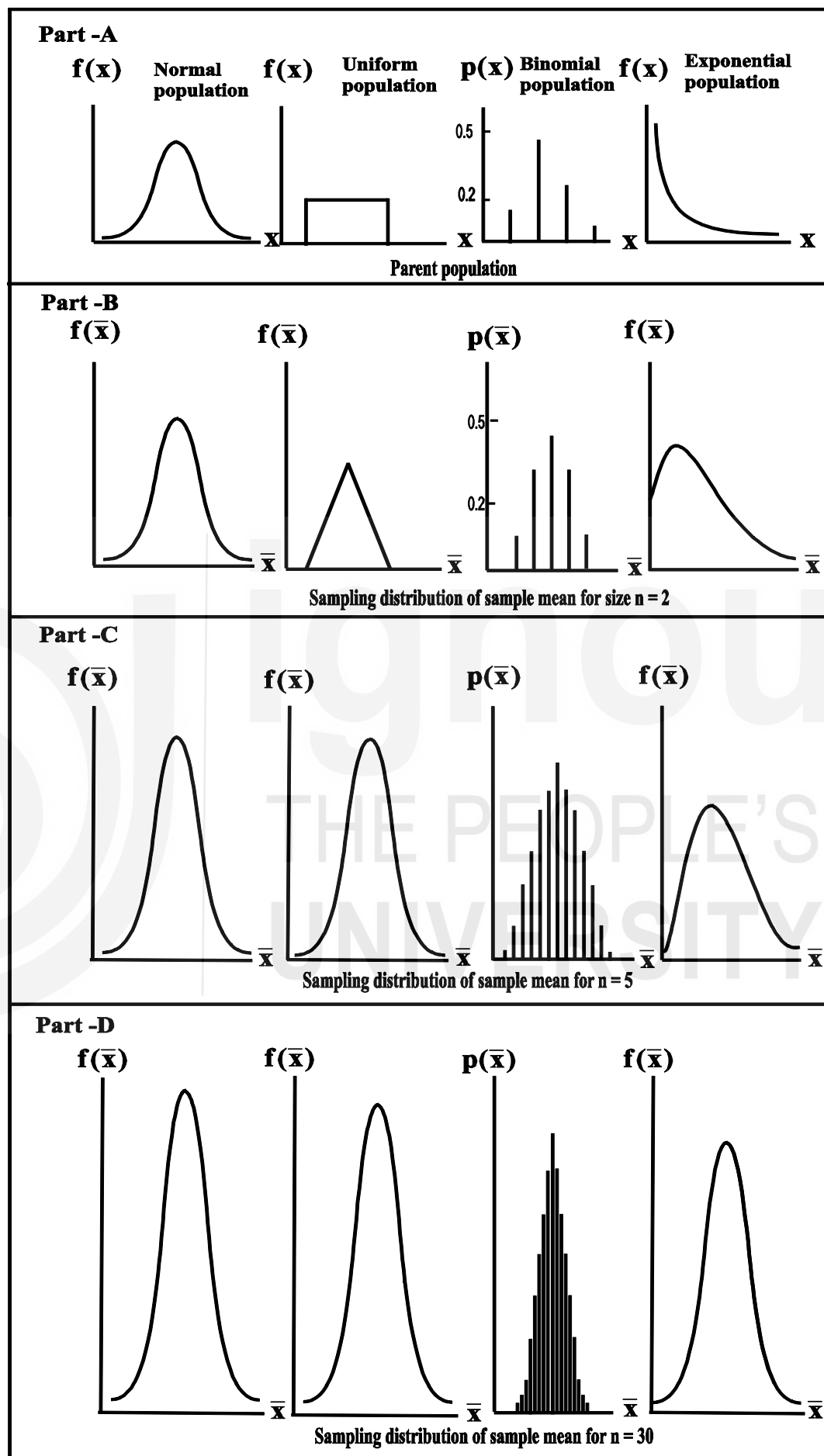


Fig. 8.1: Sampling distribution of sample means for various populations
when $n = 2$, $n = 5$ and $n = 30$

8.6 LAW OF LARGE NUMBERS AND SAMPLE SIZE

We have already discussed in the previous sections 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 form sampling distribution of that statistic. Then we draw inference about the population parameters.

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

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

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

Symbolically this can be written as

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

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

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

The proof of “law of large numbers” is beyond the scope of this Unit.

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

Suppose that the distribution of the weight of all the young men living in a city is close to a normal distribution with mean 65 kg and standard deviation of 25 kg. To understand the law of large numbers, we calculate the sample mean weight ($\bar{X}$) for varying sample size $n = 1, 2, 3, \dots$. Fig. 8.2 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 from $n = 1, 2, 3, \dots$

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

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

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

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

This is the third point on the graph.

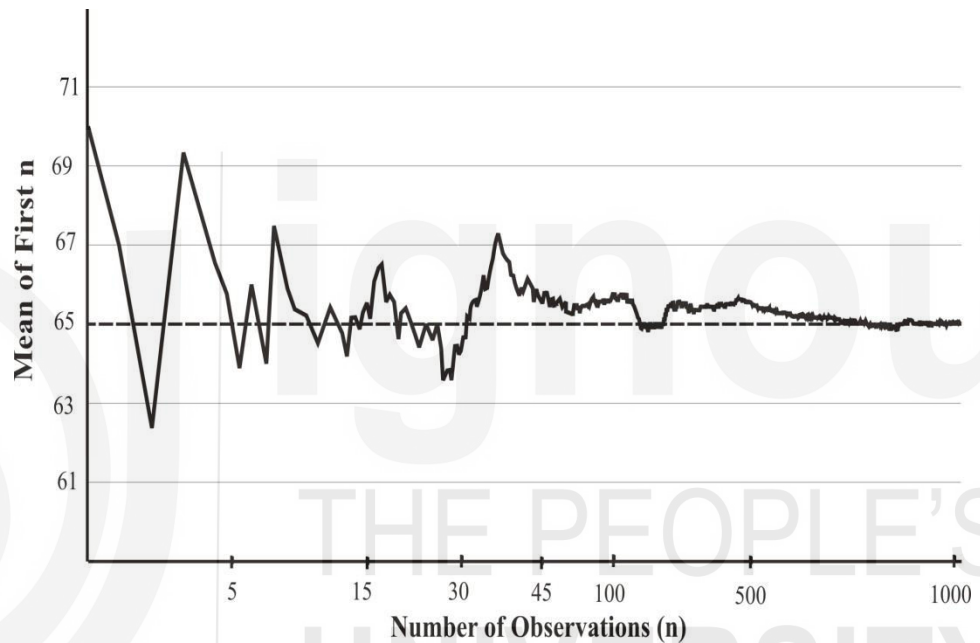


Fig. 8.2: Mean for Varying Sample Size

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

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

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

Solution: Here, we are given

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

By the law of large numbers, we have

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

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

That is, $n \geq 125$

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

Example 3: An investigator wishes to estimate the mean of a population using a sample large enough that the probability will be 0.95 that the sample mean will not differ from the population mean by more than 25 per cent of the standard deviation. How large a sample should be taken?

Solution: We have

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

By the law of large numbers, we have

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

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

That is, $n \geq 320$

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

8.7 SAMPLING DISTRIBUTION OF SAMPLE MEAN

One of the most important sample statistics, which is used to draw a conclusion about the population mean, is the sample mean. For example, an investigator may want to estimate the average income of the people living in a particular geographical area, a product manager may want to estimate the average life of electric bulbs manufactured by a company, a pathologist may want to estimate the mean time required to complete a certain analysis, etc. In the above cases, an estimate of the population mean is required, and one may estimate this on the basis of a sample taken from that population. For this, the sampling distribution of sample mean is required.

We have already given you the flavour of the sampling distribution of sample mean with the help of an example in earlier section in which we draw all possible samples of the same size from the population and calculate the sample mean for each sample. After calculating the value of sample mean for each sample we observed that the values of sample mean vary from sample to sample. Then the sample mean is treated as a random variable and a probability distribution is constructed for the values of sample mean. This probability distribution is known as the sampling distribution of sample mean. Therefore, the sampling distribution of sample mean can be defined as:

“The probability distribution of all possible values of the sample mean that would be obtained by drawing all possible samples of the same size from the population is called the sampling distribution of sample mean or simply sampling distribution of mean.”

The sampling distribution of sample means is a theoretical probability distribution of sample means that would be obtained by drawing all possible samples of the same size from the population.

Now, the question may arise in your mind “what is the shape of the sampling distribution of mean? Is it normal, exponential, binomial, Poisson, etc.?”

With the help of Fig. 8.1, we have described the shape of the sampling distribution of mean for different populations and for varying sample sizes. From this figure, we observed that when the parent population is normal then all the sampling distributions for varying sample sizes are also normal whereas when parent population is uniform, binomial, or exponential then the shapes of the sampling distributions of mean are not in the form of specifying distribution when the sample size is small ($n = 2$ or 5).

Generally, when samples are drawn from non-normal populations then it is not possible to specify the shape of the sampling distribution of mean when the sample size is small. Although when sample size is large (> 30) then we observed that sampling distribution of mean converges to normal distribution whatever the form of the population i.e. normal or non-normal.

After knowing the shapes of the sampling distribution of mean in different situations, you may be interested to know the mean and variance of the sampling distribution of mean when samples are drawn from a normal population.

In practice, only **one random sample** is selected and the concept of the sampling distribution is used to draw the inference about the population parameters. If $X_1, X_2, \dots, X_n$ is a random sample of size n taken from a normal population with mean μ and variance σ^2 then it has also been established that the sampling distribution of sample mean $\bar{X}$ is also normal. The mean and variance of the sampling distribution of $\bar{X}$ can be obtained as

$$\begin{aligned}\text{Mean of } \bar{X} &= E(\bar{X}) = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \quad [\text{By definition of } \bar{X}] \\ &= \frac{1}{n}[E(X_1) + E(X_2) + \dots + E(X_n)]\end{aligned}$$

Since $X_1, X_2, \dots, X_n$ are randomly drawn from the same population so they also follow the same distribution as the population. Therefore,

$$E(X_1) = E(X_2) = \dots = E(X_n) = E(X) = \mu$$

and

$$\text{Var}(X_1) = \text{Var}(X_2) = \dots = \text{Var}(X_n) = \text{Var}(X) = \sigma^2$$

Thus,

$$\begin{aligned}E(\bar{X}) &= \frac{1}{n} \left(\underbrace{\mu + \mu + \dots + \mu}_{n\text{-times}} \right) \\ &= \frac{1}{n}(n\mu) = \mu\end{aligned}$$

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

and variance

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] \\ &= \frac{1}{n^2}[\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)]\end{aligned}$$

The statistical inference is based on the theoretical distribution so if we say that a random sample, say, $X_1, X_2, \dots, X_n$ is taken from a population it means that the random sample is selected from the infinite population or from a finite population with replacement such that X_i 's are independent and followed same population distribution.

If X and Y are two independent random variables and a & b are two constant then by the addition theorem of expectation, we have

$$E(aX + bY) = aE(X) + bE(Y)$$

and

$$\begin{aligned}\text{Var}(aX + bY) \\ = a^2 \text{Var}(X) + b^2 \text{Var}(Y)\end{aligned}$$

$$= \frac{1}{n^2} \left(\underbrace{\sigma^2 + \sigma^2 + \dots + \sigma^2}_{n\text{-times}} \right) = \frac{1}{n^2} (n\sigma^2)$$

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

Hence, we conclude that if the samples are drawn from a normal population with mean μ and variance σ^2 then the sampling distribution of mean $\bar{X}$ is also normal distribution with mean μ and variance σ^2/n , that is,

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

The standard error (SE) of sample mean can be obtained by the definition of SE as

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

Note 5: The statistical inference is based on the theoretical distribution. Therefore, in whole statistical inference, we assume that the random sample is selected from the infinite population or a finite population with replacement.

If the sampling is done without replacement from a finite normal population of size N then the sample mean is distributed normally with mean

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

and variance

$$\text{Var}(\bar{X}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

The proof of this formula is beyond the scope of this course.

Therefore, the standard error of sample mean is given by

$$\text{SE}(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \sqrt{\frac{N-n}{N-1} \frac{\sigma^2}{n}}$$

If the sample is drawn from any population other than normal then by central limit theorem, the sampling distribution of $\bar{X}$ tends to normal as the sample size n increases, that means, the sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n when the sample size is large (> 30).

Let us see an application of the sampling distribution of mean with the help of an example.

Example 4: Diameter of a steel ball bearing produced on a semi-automatic machine is known to be distributed normally with mean 12 cm and standard deviation 0.1 cm. If we take a random sample of size 10 then find

- Mean and variance of the sampling distribution of mean.
- The probability that the sample mean lies between 11.95 cm and 12.05 cm.

Solution: Here, we are given that

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

- Since the population is normally distributed therefore the sampling distribution of the sample mean also follows a normal distribution with mean

$$E(\bar{X}) = \mu = 12$$

and variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{(0.1)^2}{10} = 0.001$$

- (ii) The probability that the sample mean lies between 11.95 cm and 12.05 cm is given by

$$P[11.95 < \bar{X} < 12.05]$$

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

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{\text{Var}(\bar{X})}} = \frac{\bar{X} - 12}{\sqrt{0.001}} = \frac{\bar{X} - 12}{0.03}$$

Therefore, by subtracting 12 from each term and then dividing each term by 0.03 in the above inequality, we get the required probability as

$$\begin{aligned} P\left[\frac{11.95 - 12}{0.03} < \frac{\bar{X} - 12}{0.03} < \frac{12.05 - 12}{0.03}\right] \\ &= P[-1.67 < Z < 1.67] \\ &= P[-1.67 < Z < 0] + P[0 < Z < 1.67] \\ &= P[0 < Z < 1.67] + P[0 < Z < 1.67] \\ &= 2P[0 < Z < 1.67] \\ &= 2 \times 0.4525 \quad \left[\begin{array}{l} \text{Using table for area} \\ \text{under normal curve} \end{array} \right] \\ &= 0.9050 \end{aligned}$$

Therefore, the probability that the sample mean lies between 11.95 and 12.05 cm is 0.9050.

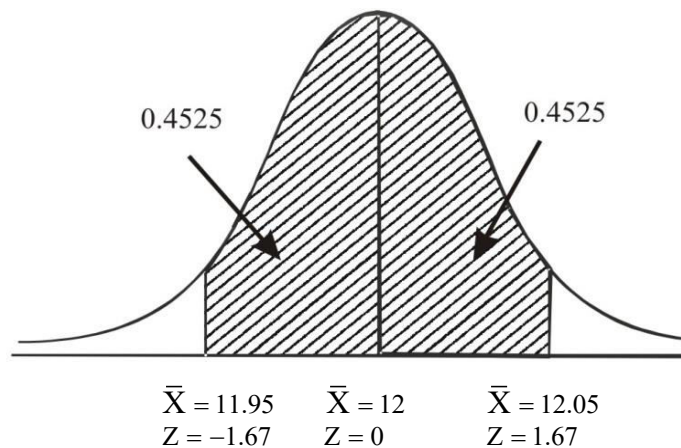


Fig. 8.3: Area Computation in Standard Normal Curve

Now, continuing our discussion about the sampling distribution of mean, we consider another situation.

In the earlier examples, we assume that variance σ^2 of the normal population is known but in general, it is not known and in such a situation the only alternative

left is to estimate the unknown σ^2 . The value of sample variance (S^2) is used to estimate the σ^2 where,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

Thus, in this case, it has been established that the sampling distribution of mean is not normal and the variate $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ is known as t-variate and follows t-distribution with $(n-1)$ df instead of normal distribution. This distribution has major applications in Statistics.

If n is sufficiently large (> 30) then we know that almost all the distributions are very closely approximated by a normal distribution. Thus, in this case, t-distribution is also approximated as a normal distribution. So for large sample size (> 30), the sampling distribution of mean also follows a normal distribution with mean μ and variance S^2/n .

Now, try some exercises for your practice

Check Your Progress 2

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

.....

.....

.....

.....

- 2) If parent population is normal with mean μ and variance σ^2 then, what is sampling distribution of sample means?

.....

.....

.....

.....

- 3) The weight of a certain type of a truck tyre is known to be distributed normally with mean 200 pounds and standard deviation 4 pounds. A random sample of 10 tyres is selected. What is the sampling distribution of sample mean? Also, obtain the mean and variance of this distribution.

.....

.....

.....

.....

- 4) The mean life of electric bulbs produced by a company is 2550 hours. A random sample of 100 bulbs is selected and the standard deviation is found to be 54 hours. Find the mean and variance of the sampling distribution of mean.

.....

.....

.....

8.8 SAMPLING DISTRIBUTION OF DIFFERENCE OF TWO SAMPLE MEANS

The sampling distribution of difference of two sample means is the theoretical sampling distribution of difference of sample means $(\bar{x} - \bar{y})$ that would be obtained by drawing all possible samples from both the populations.

There are so many problems where someone may be interested to draw the inference about the difference of two population means. For example, two manufacturing companies of bulbs are produced the same type of bulbs and one may be interested to know which one is better than the other, an investigator may want to know the difference of average income of the peoples living in two cities, say, A and B, two different types of drugs, were tried on a certain number of patients for controlling blood pressure and one may be interested to know which one has a better effect on controlling blood pressure, etc. Therefore, in such situations, to draw the inference we require the sampling distribution of difference of two sample means.

Let the same characteristic measures from two populations be represented by X and Y variables and the variation in the values of these constitute two populations, say, population-I for variation in X and population-II for variation in Y. Suppose population-I is having mean μ_1 and variance σ_1^2 and population-II is having mean μ_2 and variance σ_2^2 . Then we take all possible samples of same size n_1 from population-I and then the sample mean, say, $\bar{X}$ is calculated for each sample. Similarly, all possible samples of same size n_2 are taken from the population-II and the sample mean, say, $\bar{Y}$ is calculated for each sample. Then we consider all possible differences of means $\bar{X}$ and $\bar{Y}$. The difference of these means may or may not differ from sample to sample and so we construct the probability distribution of these differences. The probability distribution thus obtained is known as the sampling distribution of the difference of sample means. Therefore, the sampling distribution of difference of sample means can be defined as:

“The probability distribution of all values of the difference of two sample means would be obtained by drawing all possible samples from both the populations is called the sampling distribution of difference of two sample means.”

If both the parent populations are normal, that is,

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

then as we discussed in the previous section that

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Thus, as we have discussed in earlier unit that if two independent random variables X and Y are normally distributed then the difference $(X - Y)$ of these also normally distributed. Therefore, in our case, the sampling distribution of difference of two sample means $(\bar{X} - \bar{Y})$ also follows a normal distribution with mean

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

and variance

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If X and Y are independent random variables then
 $E(X - Y) = E(X) - E(Y)$
 and
 $\text{Var}(X - Y)$
 $= \text{Var}(X) + \text{Var}(Y)$

Therefore, the standard error of difference of two sample means is given by

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

If the parent populations are not normal but n_1 and n_2 are large (> 30), then by central limit theorem, the sampling distribution of $(\bar{X} - \bar{Y})$ is very closely normally distributed with mean $(\mu_1 - \mu_2)$ and variance $\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$.

Let us see an application of the sampling distribution of difference of two sample means with the help of an example.

Example 5: Electric CFLs manufactured by company A have mean lifetime of 2400 hours with standard deviation 200 hours, while CFLs manufactured by company B have mean lifetime of 2200 hours with the standard deviation of 100 hours. If random samples of 125 electric CFL of each company are tested, find

- The mean and standard error of the sampling distribution of the difference of mean lifetime of electric CFLs.
- The probability that the CFLs of company A will have a mean lifetime at least 160 hours more than the mean lifetime of the CFLs of company B.

Solution: Here, we are given that

$$\mu_1 = 2400, \sigma_1 = 200, \mu_2 = 2200, \sigma_2 = 100$$

$$n_1 = n_2 = 125$$

- Let $\bar{X}$ and $\bar{Y}$ denote the mean lifetime of CFLs taken from companies A and B, respectively. Since n_1 and n_2 are large ($n_1, n_2 > 30$) therefore, by the central limit theorem, the sampling distribution of $(\bar{X} - \bar{Y})$ follows a normal distribution with mean

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 = 2400 - 2200 = 200$$

and variance

$$\begin{aligned} \text{Var}(\bar{X} - \bar{Y}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \\ &= \frac{(200)^2}{125} + \frac{(100)^2}{125} = 400 \end{aligned}$$

Therefore, the standard error is given by

$$SE(\bar{X} - \bar{Y}) = \sqrt{\text{Var}(\bar{X} - \bar{Y})} = \sqrt{400} = 20$$

- Here, we want to find out the probability which is given by

$$P[\bar{X} \geq 160 + \bar{Y}] = P[(\bar{X} - \bar{Y}) \geq 160]$$

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

$$Z = \frac{(\bar{X} - \bar{Y}) - E(\bar{X} - \bar{Y})}{\sqrt{\text{Var}(\bar{X} - \bar{Y})}} = \frac{(\bar{X} - \bar{Y}) - 200}{\sqrt{400}} = \frac{(\bar{X} - \bar{Y}) - 200}{20}$$

Therefore, by subtracting 200 from each term and then dividing each term by 20 in the above inequality, we get the required probability

$$\begin{aligned}
 P\left[\frac{(\bar{X} - \bar{Y}) - 200}{20} \geq \frac{160 - 200}{20}\right] &= P[Z \geq -2] \\
 &= P[-2 < Z < 0] + P[0 < Z < \infty] \\
 &= P[0 < Z < 2] + P[0 < Z < \infty] \\
 &= 0.4772 + 0.5 = 0.9772
 \end{aligned}$$

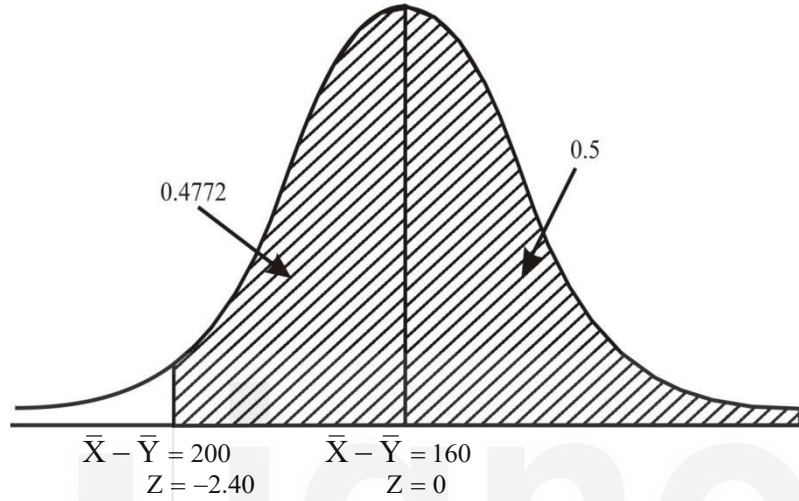


Fig. 8.4: Computation of Probability

Now, continuing our discussion about the sampling distribution of difference of two means, we consider another situation.

If population variances σ_1^2 & σ_2^2 are unknown then we estimate σ_1^2 & σ_2^2 by the values of the sample variances S_1^2 & S_2^2 of the samples taken from the first and second population respectively. And for large sample sizes n_1 and n_2 (> 30), the sampling distribution of $(\bar{X} - \bar{Y})$ is very closely normally distributed with

mean $(\mu_1 - \mu_2)$ and variance $\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)$.

If population variances σ_1^2 & σ_2^2 are unknown and $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then σ^2 is estimated by pooled sample variance S_p^2 where,

$$S_p^2 = \frac{1}{n_1 + n_2 - 2} [n_1 S_1^2 + n_2 S_2^2]$$

and variate

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1 + n_2 - 2)}$$

follows t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. As similar as the sampling distribution of mean, for large sample sizes n_1 and n_2 (> 30) the sampling distribution of $(\bar{X} - \bar{Y})$ is very closely normally distributed with mean

$(\mu_1 - \mu_2)$ and variance $S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$.

The sampling distributions are also defined for proportions also, but they are beyond the scope of this Unit. You may refer to references for more details on that topic.

Check Your Progress 3

- 1) The average height of all the workers in a hospital is known to be 68 inches with a standard deviation of 2.3 inches, whereas the average height of all the workers in a company is known to be 65 inches with a standard deviation of 2.5 inches. If a sample of 35 hospital workers and a sample of 50 company workers are selected at random, what is the probability that the sample mean of the height of hospital workers is at least 2 inch greater than that of company workers?
-
-

8.9 SUMMARY

In this unit, we have covered the following points:

1. The statistical procedure which is used for drawing conclusions about the population parameter on the basis of the sample data is called “**statistical inference**”.
2. The group of units or items under study is known as “**Population**” whereas a part or a fraction of population is known as “**sample**”.
3. A “**parameter**” is a function of population values which is used to represent the certain characteristic of the population and any quantity calculated from sample values and does not contain any unknown population parameter is known as “**statistic**”.
4. Any statistic used to estimate an unknown population parameter is known as “**estimator**” and the particular value of the estimator is known as “**estimate**” of the parameter.
5. The probability distribution of all possible values of a sample statistic that would be obtained by drawing all possible samples of the same size from the population is called “**sampling distribution**” of that statistic.
6. The standard deviation of the sampling distribution of a statistic is known as “**standard error**”.
7. The most important theorem of Statistics is “**central limit theorem**” which state that the sampling distribution of the sample means tends to a normal distribution as the sample size tends to large ($n > 30$).
8. According to the law of large numbers, a positive integer can be determined such that if a random sample of size n or larger is taken from a population with a parameter, say, population mean (μ), the probability that the sample mean will be very closed to the population mean, can be made as close to 1 as desired.
9. The sampling distribution of sample mean and the sampling distribution of difference between two sample means are normally distributed.

8.10 SOLUTIONS / ANSWERS

Check Your Progress 1

- 1) Here, we are given that
 $N = 3, n = 2$

Since we have to estimate the average life of Televisions on the basis of samples of size $n = 2$ therefore, all possible samples (with replacement) are $N^n = 3^2 = 9$ and for each sample, we calculate the sample mean as shown in Table 8.4 given below:

Table 8.4: Calculation of Sample Mean

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

Since the arrangement of all possible values of the sample mean with their corresponding probabilities is called the sampling distribution of mean thus, we arrange every possible value of the sample mean with their respective probabilities in the following Table 8.5 given below:

Table 8.5: Sampling distribution of sample means

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

2) Here, we are given that

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

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

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

3) Here, we are given that

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

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

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

Check Your Progress 2

1) Here, we are given that

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

By the law of large numbers, we have

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

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

That is, $n \geq 160$

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

- 2) Since parent population is normal so sampling distribution of sample means is also normal with mean μ and variance σ^2/n .
- 3) Here, we are given that

$$\mu = 200, \sigma = 4, n = 10$$

Since parent population is normal so the sampling distribution of sample means is also normal. Therefore, the mean of this distribution is given by

$$E(\bar{X}) = \mu = 200$$

and variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{(4)^2}{10} = \frac{16}{10} = 1.6$$

- 4) Here, we are given that

$$\mu = 2550, n = 100, S = 54$$

First of all, we find the sampling distribution of sample mean. Since the sample size is large ($n = 100 > 30$) therefore, by the central limit theorem, the sampling distribution of sample mean follows a normal distribution. Therefore, the mean of this distribution is given by

$$E(\bar{X}) = \mu = 2550$$

and variance

$$\text{Var}(\bar{X}) = \frac{S^2}{n} = \frac{(54)^2}{100} = \frac{2916}{100} = 29.16$$

Check Your Progress 3

- 1) Here, we are given that

$$\mu_1 = 68, \sigma_1 = 2.3, n_1 = 35$$

$$\mu_2 = 65, \sigma_2 = 2.5, n_2 = 50$$

To find the required probability, first of all, we find the sampling distribution of difference of two sample means. Let $\bar{X}$ and $\bar{Y}$ denote the mean height of workers taken from hospital and company respectively. Since n_1 and n_2 are large ($n_1, n_2 > 30$) therefore, by the central limit theorem, the sampling distribution of $(\bar{X} - \bar{Y})$ follows a normal distribution with mean

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 = 68 - 65 = 3$$

and variance

$$\text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(2.3)^2}{35} + \frac{(2.5)^2}{50}$$

$$= 0.1511 + 0.1250 = 0.2761$$

The probability that the sample mean of the height of hospital workers is 2 inches less than that of company workers is given by

$$P[(\bar{X} - \bar{Y}) < 2]$$

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

$$Z = \frac{(\bar{X} - \bar{Y}) - E(\bar{X} - \bar{Y})}{\sqrt{\text{Var}(\bar{X} - \bar{Y})}} = \frac{(\bar{X} - \bar{Y}) - 3}{\sqrt{0.2761}} = \frac{(\bar{X} - \bar{Y}) - 3}{0.53}$$

Therefore, by subtracting 3 from each term and then dividing each term by 0.53 in the above inequality, we get the required probability

$$\begin{aligned} P\left[\frac{(\bar{X} - \bar{Y}) - 3}{0.53} < \frac{2 - 3}{0.53}\right] &= P[Z < -1.89] = 0.5 - P[-1.89 \leq Z \leq 0] \\ &= 0.5 - P[0 \leq Z \leq 1.89] \\ &= 0.5 - 0.4706 = 0.0294 \end{aligned}$$

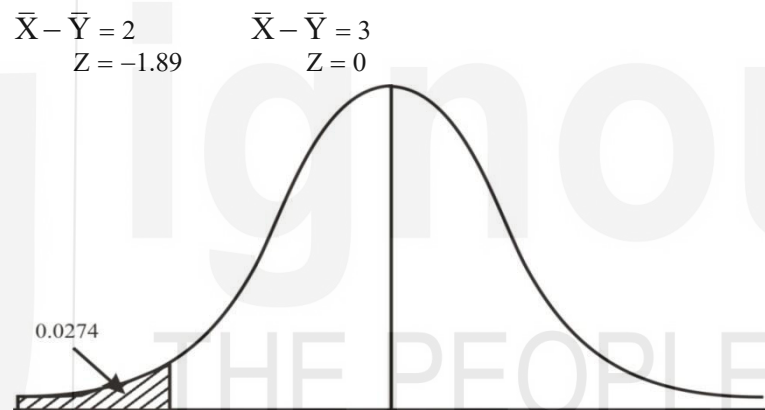


Fig. 8.5: The Distribution of Difference of Two Normal Variables

8.11 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 1: Sampling Distributions, UNIT 1: Introduction to Sampling Distribution (for Sections 8.2 to 8.6 and related portions in Sections 8.0 , 8.1, 8.9, 8.10 and Check Your Progress questions)
2. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 1: Sampling Distributions, UNIT 2: Sampling Distribution(s) of Statistic(s) (for Sections 8.6 to 8.8 and related portions in Sections 8.0 , 8.1, 8.9, 8.10 and Check Your Progress questions)

Reference Book

3. "Introduction to Probability and Statistics for Engineers and Scientists", 3rd Edition, Sheldon M. Ross, Elsevier Academic Press, 2004