1. Introduction

Welcome to the series of E-learning modules on Central Limit Theorem. In this module we are going cover the concept of Central Limit Theorem, implication of the theorem, difference between Central Limit Theorem and Normal Theorem and Importance of the Central Limit theorem.

By the end of this session, you will be able to know:

- Understand the Central Limit Theorem
- Describe Assumptions and Implication of the theorem
- Discuss the difference between Central Limit Theorem and Normal Theorem
- Understand Required sample size to follow the theorem
- Explain the importance and applications of the theorem

It is possible to draw more than one sample from the same population and the value of an estimator will in general vary from sample to sample.
For example: the average value in a sample is an estimator.

The average values in more than one sample, drawn from the same population, will not necessarily be equal. The probability distribution or probability density function of an estimator is known as sampling distribution.

The sampling distribution describes probabilities associated with an estimator when a random sample is drawn from a population.The probability distribution of sample mean drawn from a population can be derived for the following two cases.

Case1:
Suppose we know that the probability distribution of the population is Normal with mean μ and variance sigma square. In such a case each for Xi, i^{th} observation in the sample will have Normal with mean mu and variance sigma square distribution.
Then it can be proved mathematically that the probability distribution of x bar will be a Normal curve with center at mu and spread of sigma by root n, and we call this result as Normal theorem

Normal Theorem:
If the Random Variables X1,X2, etc till Xn are independent and identically distributed as Normal (mu, sigma square) then x bar is equal to summation xi by n is distributed as Normal (mu, sigma square by n)

2. Case 2

Case II :
We may not know the exact probability distribution of the population or the population probability distribution may not be Normal or close to Normal at all. In such a case the probability distribution of x bar is equal to summation xi by n (which depends on n) starts to take a bell shape

when n is very large the probability distribution x bar is equal to summation xi by n by Central Limit Theorem is almost Normal with center at mu and spread (Tail thickness) sigma by root n

The sampling distribution of the sample mean, x bar is approximated by a normal distribution when the sample is a simple random sample and the sample size, *n*, is large.

In this case, the mean of the sampling distribution is the population mean, μ, and the standard deviation of the sampling distribution is the population standard deviation, σ, divided by the square root of the sample size. The latter is referred to as the standard error of the mean.
In symbols, the standard error is sigma by root n

Suppose {Xk } is the sequence of Bernoulli random variables taking values 1 with probability p, and 0 with probability 1 minus p is equal to q, then Expected value of Xk is equal to p and Variance of Xk is equal to pq. If Sn is equal to summation Xk, Expected value of Sn is equal to np and Variance of Sn is equal to npq so that Sn minus np bu root npq is a Standard Normal variate with mean 0 and variance unity.

Theoretical basis of this result is called CLT is given by Laplace. Ever since attempts have been made by probability theorists to weaken the assumptions under which the above holds.

Firstly the condition that {Xk} is a sequence of independent Bernoulli Random variable is replaced by the condition that {Xk} is a sequence of independent identically distributed Random variables with a common mean mu and variance sigma square . Next the condition that the {Xk} has an identical distribution is relaxed.

Now the condition that the {Xk}'s are independent is being relaxed and replaced by milder restrictions. Thus the above property which is proved originally for Bernoulli Random variables holds true for the more general random variables. This is called invariance principle.

3. Size of the Sample

How large must a sample be for the Central Limit theorem to apply?
The sample size varies according to the shape of the population.The CLT states that under rather general conditions sums and means of random samples of measurements drawn from a population tend to have an approximately Normal distribution .

Suppose you toss a balanced die n is equal to one time. The Random variable X is the number observed on the upper face. This familiar Random variable can take six values, each with probability one by six. The shape of the distribution is flat or Uniform and systematic about the mean mu is equal to three point five with a Standard deviation sigma is equal to one point seven one.

Now take a sample of size n is equal to two, from this population that is toss two dice and record the sums of numbers on the two upper faces summation Xi is equal to x one plus x two. There are thirty six possible outcomes each with probability one by thirty six. There is dramatic difference in the shape of the sampling distribution of x bar. It is now roughly mound shaped but still symmetric about the mean mu is equal to three point five.

Similarly for n is equal to three and n is equal to four the sampling distribution clearly shows the mound shape of the Normal Probability distribution still centred at mu is equal to three point five. Notice also that the spread of the distribution is slowly decreasing as the sample size n increases.

The distribution of x bar is approximately Normally distributed based on a sample as small as n is equal to four. This phenomenon is the result of an important statistical theorem called the Central limit theorem.

If random samples of n observations are drawn from a Nonnormal population with finite mean mu and Standard deviation sigma, then when n is large the sampling distribution of the sample mean x bar is approximately Normally distributed with mean mu and standard deviation sigma by root n
The approximation becomes more accurate as n becomes large.

Regardless of its shape the sampling distribution of x bar always has a mean identical to the mean of the sampled population and a standard deviation equal to the population and standard deviation sigma by root n.
Consequently the spread of the distribution of the sample means is considerably less than a spread of the sampled population.

The CLT can be restated to apply to the sum of the sample measurements summation xi, which, as n becomes large also has an approximately Normal distribution with mean n mu and standard deviation sigma into root n
As you reread the CLT, you may notice that the approximation is valid as long as the sample size n is large'.
But now how large is 'large'?
Unfortunately there is no clear answer to this question. The appropriate value of n depends

on the shape of the population from which you sample as well as on how you want to use the approximation.

However these guidelines will help to decide when the sample size is large enough?

- If the sampled population is normal , then the sampling distribution of x bar will also be Normal. No matter what sample size you choose. The result can be proven theoretically but it should not be too difficult for you to accept without proof
- When the sampled population is approximately systematic the sampling distribution of x bar becomes approximately Normal for relatively small values of n . Remember how rapidly the flat distribution in the dice example become mound shaped (n is equal to three)
- When the sampled population is skewed, the sample size n must be larger, with n at least thirty , before the sampling distribution of X bar becomes approximately Normal

These guidelines suggest that for many populations before the sampling distribution of X bar will be approximately Normal for moderate sample sizes, an exception to this rule is sampling a Binomial population when either p or q is equal to one minus p is very small.
The sample size varies according to the shape of the population.

However, for our use, a sample size of 30 or larger will suffice. A sample size of hundred or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed, n greater than thirty may be sufficient to use the CLT.

Note what the CLT says: If you have independent Random variables X one, X two Etc till X n, each ,with the same distribution which has mean and variance , then X n bar is standardized by subtracting its mean and then dividing by its standard deviation has a distribution that approaches a standard Normal distribution.

The key thing to note that it does not make any difference what common distribution X one, X two. Etc till X n; have as long as they have a mean and variance.
Irrespective of the shape of the underlying distribution of the population, by increasing the sample size, sample means & proportions will approximate normal distributions if the sample sizes are sufficiently large.

4. Assumptions & Importance of CLT

Statement of Central Limit Theorem
Suppose X one ,X two, etc X n, be n independent random variables having the same probability density function each with Expected value of Xi is equal to mu and Variance of Xi is equal to sigma square , for I equal to one, two , etc. till n then Sn is equal to X one plus X two plus etc till plus X n is approximately Normal with mean n mu and variance n sigma square.
Also, Z is equal to x bar minus n mu by sigma into root n is asymptotically Normal with mean zero and variance one.

General assumptions of Central Limit Theorem
1)The variables are independent
2) All the variables follow same distribution
3) The mean and variance of all the variables exists
4) The mean and variance of all the variables are same
5) Number of variables are very large (that is n tends to infinity)

The approximations can be garnered from the Central Limit Theorem and can be listed as a corollary
Corollary:
If X one, X two, etc till Xn, are independent identically distributed Random variables with common mean mu and variance sigma square then,

Probability of ‘a’ less than x bar minus mu by sigma by root ‘n’ less than ‘b’ is equal to Phi of ‘b’ minus Phi of ‘a’
This equation gives an approximate value of the probabilities of certain events described in terms of averages and sums.

Importance of Central Limit Theorem
The practical utility of the Central Limit Theorem is inherent in its approximation.
The important contribution of Central Limit Theorem is in Statistical Inference. Many estimations that are used to make inferences about population parameters are sums or averages of the sample measurements. When the sample size is sufficiently large, you can expect these estimators to have sampling distribution that are approximately Normal.

You can then use the Normal distribution to describe the behaviour of these estimators in repeated sampling and evaluate the probability of observing certain sample results. These probabilities are calculated using the Standard Normal Random Variable
Z is equal to Estimator minus mean by Standard Deviation

The CLT plays an important role in Statistical Theory. It is one of the usual assumptions in Statistics to assume that the underlying observations to follow Normal distribution, at least, approximately. The theory of error used by physicists, or astronomers can be justified on the

basis of Central Limit Theorem. Thus investigating the most general set of conditions under which the Central Limit Theorem can hold is of theoretical and practical interest.

The significance of the central limit theorem lies in the fact that it permits us to use sample estimators to make inference about the population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

Similarities and dissimilarities between Central Limit Theorem and Normal Theorem
Normal Theorem

- Population Mean equal to mu ,Standard Deviation equal to sigma
- Shape of the population is known to be a Normal curve(mu, sigma square)
- Sample average x bar is said to have Normal (mu, sigma square by n) for any n
- Central Limit Theorem
- Population Mean equal to mu, Standard deviation equal to sigma
- Shape of the population is either unknown or not Normal
- Sample average x bar is said to have Normal (mu, sigma square by n) approximately for only large n

5. Check for Normality

CHECK FOR NORMALITY:

- Use descriptive statistics.
- Construct stem-and-leaf plots for small or moderate-sized data sets and frequency distributions and histograms for large data sets.
- Compute measures of central tendency (mean and median) and compare with the theoretical and practical properties of the normal distribution.
- Compute the inter quartile range. Does it approximate the one point three three times the standard deviation?
- How are the observations in the data set distributed?
- Do approximately two thirds of the observations lie between the mean and plus or minus one standard deviation?
- Do approximately four-fifths of the observations lie between the mean and plus or minus one point two eight standard deviations?
- Do approximately nineteen out of every twenty observations lie between the mean and plus or minus 2 standard deviations?

Why do I care if X bar, the sample mean, is normally distributed?
Because I want to use Z scores to analyze sample means.
But to use Z scores, the data must be normally distributed.
That's where the Central Limit Theorem steps in.

Recall that the Central Limit Theorem states that sample means are normally distributed regardless of the shape of the underlying population if the sample size is sufficiently large.

- Z is equal to X minus mu by sigma

- If sample means are normally distributed, the Z score formula applied to sample means would be:

Z is equal to [X-bar minus mu X-bar] by sigma X-bar

Background
To determine mu X-bar, we would need to randomly draw out all possible samples of the given size from the population, compute the sample means, and average them. This task is unrealistic. Fortunately, mu X bar equals the population mean mu, which is easier to access.

Likewise, while computing the value of sigma X-bar, we would have to take all possible samples of a given size from a population, compute the sample means, and determine the standard deviation of sample means. This task is also unrealistic. Fortunately, sigma X bar can be computed by using the population standard deviation divided by the square root of the sample size.

Note:
As the sample size increases, the standard deviation of the sample means becomes smaller and smaller because the population standard deviation is being divided by larger and larger values of the square root of n.

The ultimate benefit of the central limit theorem is a useful version of the Z formula for sample means. Z formula for sample means is given by:
Z = [X bar minus mu] by sigma by root n

Example :
The mean expenditure per customer at a tire store is d eighty five dollars, with a standard deviation of nine dollars. If a random sample of forty customers is taken, what is the probability that the sample average expenditure per customer for this sample will be eighty seven dollars or more?

Because the sample size is greater than thirty, the central limit theorem says the sample means are normally distributed.
Z is equal to [X bar minus mu] by sigma by root n
Z is equal to eighty seven minus eighty five by nine by root forty
Z is equal to two by one point four two is equal to one point four one
For Z is equal to one point four one in the Z distribution table, the probability is point four two zero seven.

This represents the probability of getting a mean between eighty seven dollars and the population mean eighty five dollars.
Solving for the tail of the distribution yields
Zero point five minus zero point four two zero seven is equal to zero point zero seven nine three.
This is the probability of X-bar greater than or equal to eighty seven dollars.

Interpretations:
Therefore, seven point nine three percent of the time, a random sample of forty customers from this population will yield a mean expenditure of eighty seven dollars or more or

From any random sample of forty customers, seven point nine three percent of them will spend an average eighty seven dollars or more.

Here's a summary of our learning in this session where we have :

- Understood the concept of Central Limit Theorem
- Understood the Assumptions and Implication of the theorem
- Explained the required sample size to follow the theorem
- Explained the importance of the theorem
- Understood the difference between Central Limit Theorem and Normal Theorem
- Understood the applications of Central Limit Theorem