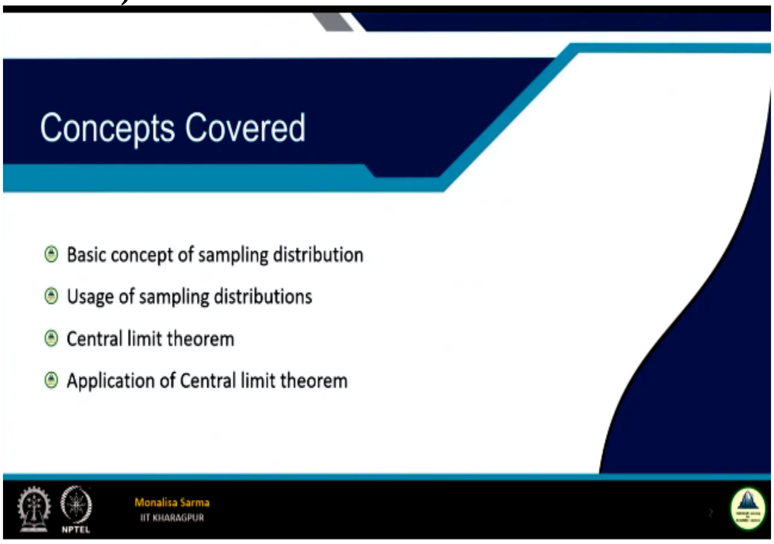
Statistical Learning for Reliability Analysis

Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 16
Sampling Distributions (Part 1)

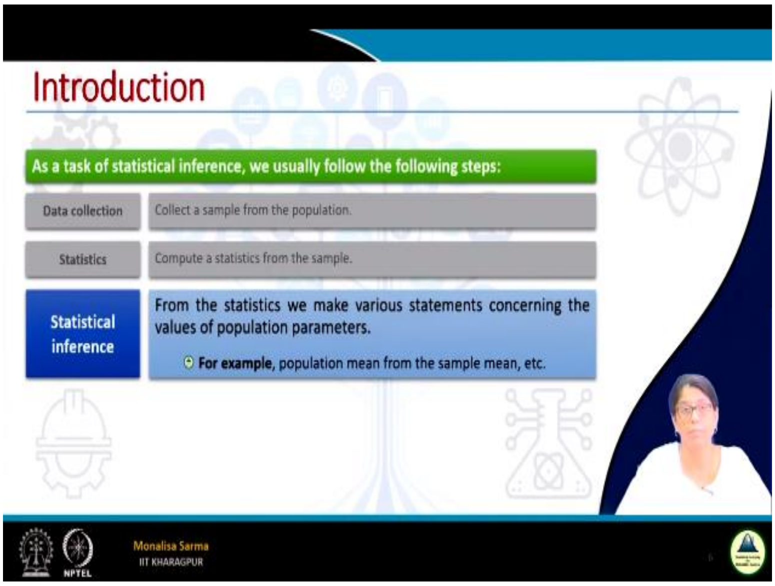
Hello, everyone. So today we are starting a new topic that is sampling distribution, it is a very interesting topic, I am sure all of you will enjoy doing this. So, the sampling distribution I will be taking in 3 different parts. So today I will be covering the first part and conjunctively the other parts.

(Refer Slide Time: 00:46)



So now in this lecture today, I will cover the basic concept of sampling distribution, then the uses why we need a sampling distribution, then I will talk about central limit theorem, a very well known theorem and of course, the application of the central limit theorem.

(Refer Slide Time: 01:02)



Now, as you remember, in our first class of statistics, latest from that is the second lecture, I have talked of statistical inference what is a statistical inference? If we can remember statistical inference if we talk about statistical inference, the main objective of statistical inference that what we use in most of our application is like given if we want some information about the population, we have definitely population is a big thing.

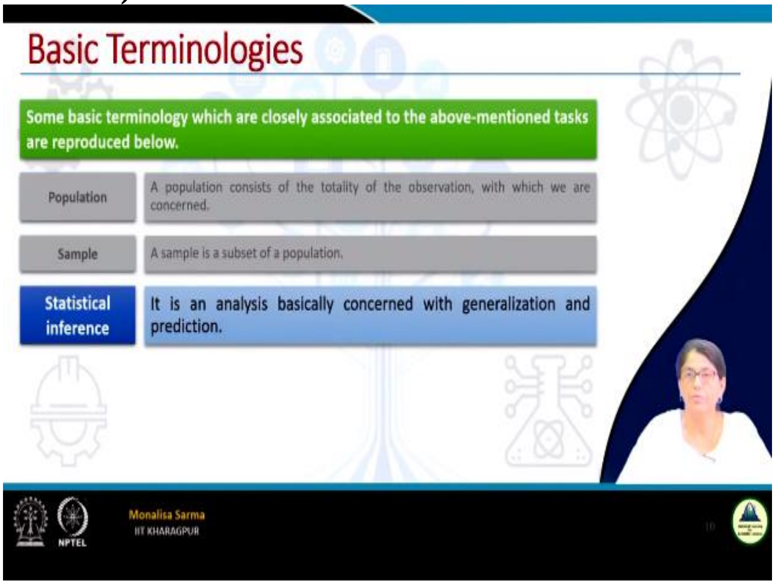
So, if we want some information from population, what we do is that we get a sample from the population of course, the sample has to be an unbiased sample and from the sample we try to extract the different information, which based on this information, we try to predict the information about the population that is recall the statistical inference in a very broad sense. So now, in the when we talk of statistical inference, what we need to do?

First we need to collect the data that is, as I told you, first we will have to collect a sample from the population what is sample, what is population this we have already discussed in my earlier lectures. Then from the sample, we will have to compute a statistics now, what is the statistics? In a population remember, we talked about the characteristics of the population when you talk about populations there, there are certain characteristics which define a population like it may be the mean it may be the variance.

So the mean and variance these are this we basically called a parameters of a population, please remember this we called a parameters of the population. Now the counterpart of parameters and sample we call it statistics we know longer call it parameter, we call it statistics. So mean of a sample variance of a sample, we call it the statistics of the sample, rather than parameter of the sample but in case of population, we call it the parameter of the population.

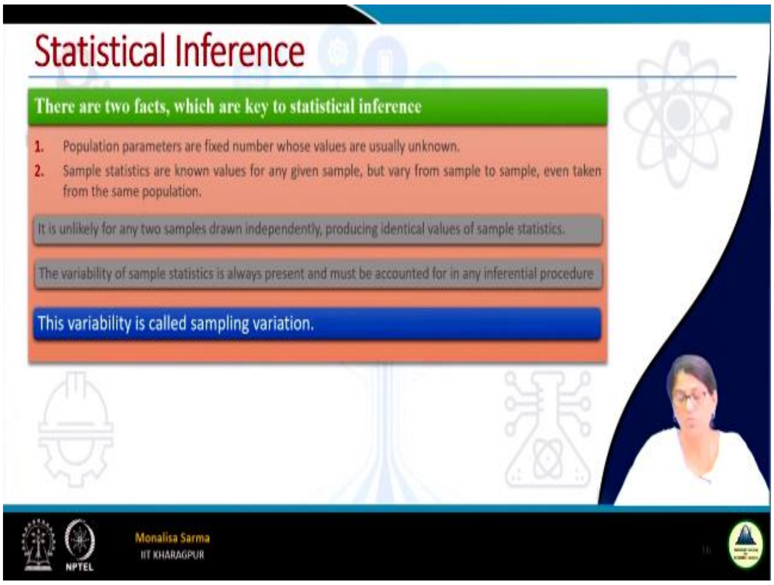
So once we collect the sample, the next job is we will compute the statistics from the sample. So, from the statistics, then we will make various statements concerning the different values of the population parameter. For example, we may talk about the population mean from the sample mean, if we get some we find out the mean of the sample from there, we will try to infer something about the population this is one sort of statistical inference. Likewise, there can be in many other statistical inferences, which you do for a population from a sample.

(Refer Slide Time: 03:23)



So, some of the basic terminology which are closely associated with the statistical inference basically, our population which I have already discussed and not be repeating here than sample then for it is statistical inferences already we know all this.

(Refer Slide Time: 03:40)



So, now, coming straight to this, there are 2 factors which are keys to statistical inferences, what are the first is population parameters are fixed numbers whose values are usually unknown. So population parameters, when you talk about suppose a mean of a population is a fixed parameter, but then it is really not possible to know the mean of a population like let me take a very small example suppose I am considering the some LED bulbs a company is producing a bulk of LED bulbs that light.

So now, if we consider it a whole lot that lot that is the population. Now, if I want to find out the mean what to say mean lifetime of those bulbs, then it is really not possible to find out the mean lifetime of those bulbs, by considering the whole lot, because there may be say in 1 lot blub, 1 lot bulb is produced in a blub. So if I want to find out the mean lifetime. So I need to know the lifetime of all the bulbs.

Then using whatever formula we have for mean you already known the formula for mean, using the formula for mean will have to tell the mean lifetime of the bulb, but that is nearly not possible. We cannot take the mean lifetime of all the bulbs. So, now so what is the remedy done sample statistics are known values for any given sample, but vary from sample to sample even taken from a very small population. So, now, we are interested in mean lifetime available.

So, we really not possible for us to calculate the lifetime of all the bulbs and then find out the mean. So, what we do is that the normal procedure is that we take a sample from the population a small sample maybe and from that sample we try to find out the mean lifetime of the bulbs we take the lifetime of all the bulbs how do we take the lifetime of all the bulbs do a different experiment basically, maybe what is say we use the bulbs for quite a long till it goes bad.

So, that definitely we cannot do of all the population we will take a smaller sample and from the smaller sample we keep the light on till it goes back and then from that we can find out a mean lifetime of the bulbs this is just an example of course, this is not the way to find out the mean lifetime of the bulbs there are different other techniques also to find out these values these statistics, so, I will not valid into those.

So, now, if we take a sample from 1 sample suppose, I have taken a sample of 10 bulbs from 10 bulbs if I take the lifetime and then I find out the mean I get a particular value say x. Now, suppose I take a different sample and then again I take the lifetime of those value when I take the mean I get maybe that is x 1, this x is not equal to x 1 there are very less probability that x will be exactly equal to x 1 there will be some variation between this x and x 1.

So, this sample statistics so the second point that means, sample statistics we can know it we take a sample from the sample we can find out the sample size because this is a small number small amount, but then this small sample statistics vary from sample to sample even we have taken it from the same population. So, it is very unlikely that 2 samples drawn independently producing identical values of the sample statistics.

So, the variability of sample statistics is that means always there. So as I told you in the beginning when I talk of statistical inference. Statistical inference means we will take a sample from the sample we will try to find out the statistics based on the statistics we will try to infer about the parameters of the population. Now, see the dilemma here the samples that is when we talk about the sample statistics, the sample statistics vary from sample to sample, I have taken a sample of 10 numbers.

I have taken another sample of 10 numbers, the statistics we say get from this first sample, which I get from the second sample, it is very unlikely that I will get the same value. So, there are variations. So, if they are a variation then how do I do the statistical inference. So, that is the thing. So, this variation we must take into account this variation in this inferential procedure. So, how do we do that? So, this variability is called sampling variation, because variation around the samples so it is called sampling variations.

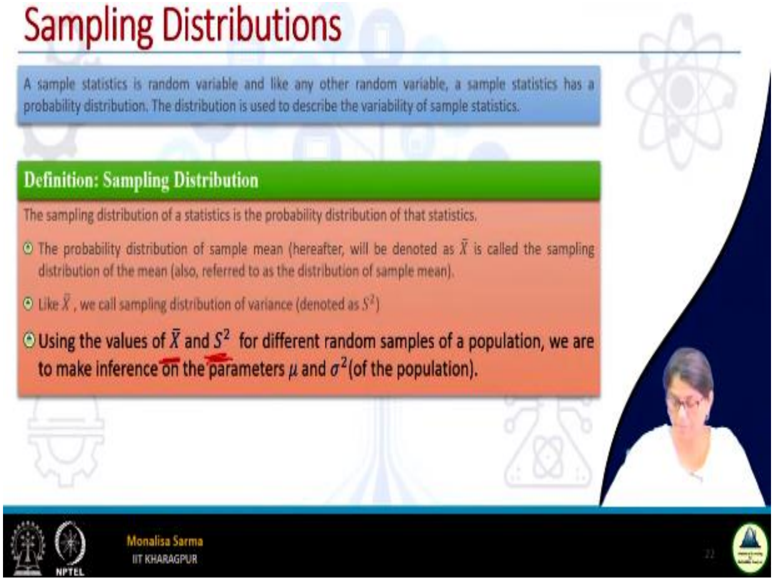
Now, see when we talk about this say population mean only sample mean so, we take a sample we get a mean another sample we get another mean say first sample mean I got a x 1, we take another sample, we got the same mean x 2, we take another sample say, we got mean x 3. So, this sample mean will be different and the probability of getting this value is different there it is also different.

So, now remember what is probably a distribution? Probability distribution is nothing but the value that a random variable will take. And the probability of those random variable these 2 things, if we put together it is the only problem the distribution Now, here what our sample mean or variance or whatever statistics we consider. So, this sample statistic is a random variable, because it may take different values, it is not that it will take a fixed value, so it will take different value. So, our sample mean is a random variable.

And since this is a random variable, we can also have its probability of occurrence. So, these 2 things put together this sample mean, it is a random variable. And since it is a random variable, it has different probability of occurrence. So, considering these 2 things together, we can have a probability distribution of that. So, this probability distribution is nothing but it is called a sampling distribution. So that is it that is the distribution sampling distribution is nothing but a probability distribution only.

But it is a probability distribution of the sample statistics that is also a random variable. Now what is this random variable? This random variable is a sample statistic that is why we call it a sampling distribution.

(Refer Slide Time: 10:09)



Now, this distribution so, when we have a distribution what is the use of probability distribution I have talked already remember why we need to study probability distribution one of them important objective of studying probability distribution is that we can find out the different values that are random variability and with this what we can find out? We can find out the probability of occurrence of a particular random value.

So, because now we have for the sample mean or the sample statistics, let us not take mean because there can be any other statistic for the sample statistics now, since when we have a distribution. So that means we from this distribution, we can use to describe the variability of the sample statistics what are the different values that sample may take as well as this probability of occurrence.

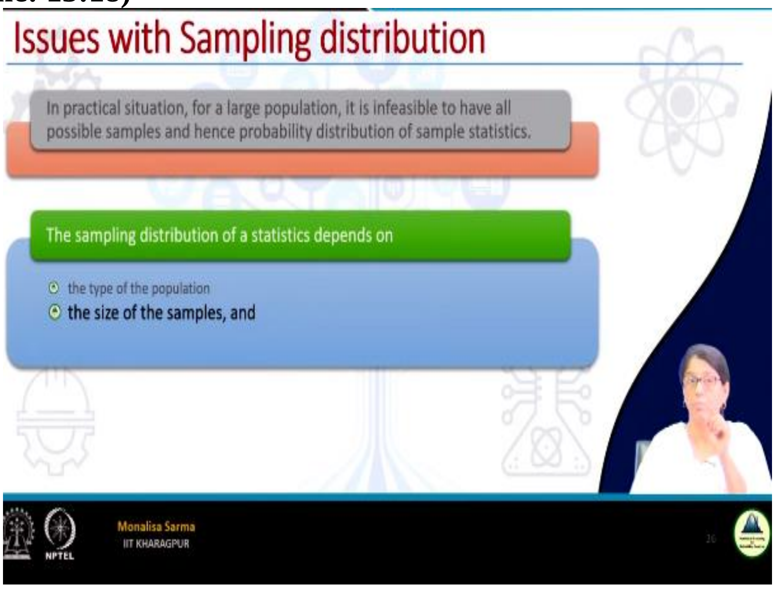
So, the definition of sampling distribution, the sampling distribution of a statistic is the probability distribution of that statistic. So, probability distribution of sample mean, we remember for the population mean, how do we designate the population mean we denote by mu you remember and population variance removed we denote by σ^2. So, similarly for sample mean we will no longer use mu rather we will use either X bar or Y bar general convention is using X bar in some books you will also find using Y bar.

So that is the general convention of using denoting sample mean. So, now when we do the distribution of X bar that is a sample mean then we call it as a sampling distribution of the mean. So, the probability distribution so say this line the probability distribution of the sample mean is called the sampling distribution of the mean. So this if we find out the probability distribution of sample mean that we call it a sampling distribution of the mean similarly, the sampling distribution of the variance also we will discuss later.

So, sampling distribution of variance, we will be discussing in the next class, this class we will continue our discussion on the under sampling distribution of mean. So, using the values of this X unlike for population variance, we use the term σ^2. Similarly, for sample variance, we use the term S^2 this is the population sample mean it is X bar Y bar for sample variance we use S^2 and all the books you will be finding it is denoted by S^2.

So, using the values of X bar and S^2 for different random sample of the population, we are to make inference on the parameters of population that is parameters μ and σ^2 so that is statistical inference.

(Refer Slide Time: 13:18)



Now, how we make the sampling distribution because infer about a population we need to have the sampling distribution because just taking 1 sample will not do because, by them as we have already seen the sample statistics differs from different samples. So, that is why we need the sampling distributions once we have the sampling distribution then only we can infer about the population. Now how to have the sampling distribution?

So to like be, we have learned about probability distribution remember, we have learned in to have a probability distribution what we need? We needed different values that a random variability as well as the probability of occurrence of those values, what is the probability of occurrence? Probability of occurrence is nothing but the relative frequency is not it? A relative frequency of occurrence. So now similarly so how we will construct a sampling distribution so given a population.

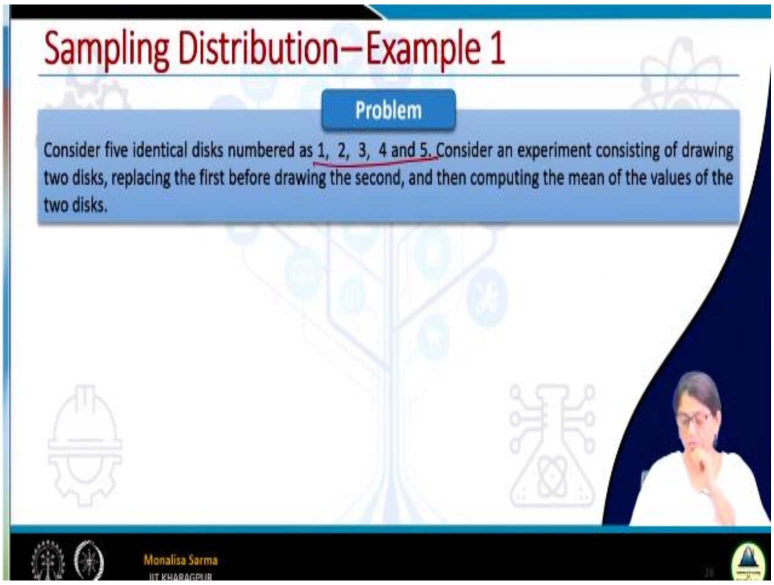
So, what we will do? We will take as many sample as possible given a say big population from there suppose, we take a population initially suppose, we took a population say first 1 sample from there suppose, we got a mean of X 1 bar then again X again we took another 1 sample and different set of sample say we took sample mean X 2 bar another again we take a third sample for the suppose we take with likewise we suppose took different samples from different samples we found different means.

So, it may be that again there may be some samples which made mean value maybe same So, based on this we can find out what is the relative occurrence of this sample mean relative occurrence of X 1 bar, relative occurrence of X 2 bar, relative occurrence X 3 bar. So, relative occurrence of that particular variable what is that that is nothing but the probability that so, with this we can calculate that we can, what to say develop the sampling distribution.

Now for a huge population, is it really feasible to take out all sub samples and then find out the sample statistics? It is really not possible we may miss out on some sample. So, we do not have to do that many researchers have already done that and based on that, they have already come up with some based on different empirical study they have already come up with some theorem. Now, I go to the theorem later, but firstly just to see how the researchers have come up with different theorems for that, I will just take a small example.

First of all, the sampling distribution of the statistics depends on what it depends on the type of the population, whether the population is normal distribution or totally not normal distribution or near normal distribution or some other distribution, why I am stressing on normal I will come to that again. It also depends on the sample values it also depends on the size of the sample, it also depends on a method of choosing the samples whether the sample is a bias sample or a non bias 1.

(Refer Slide Time: 16:35)

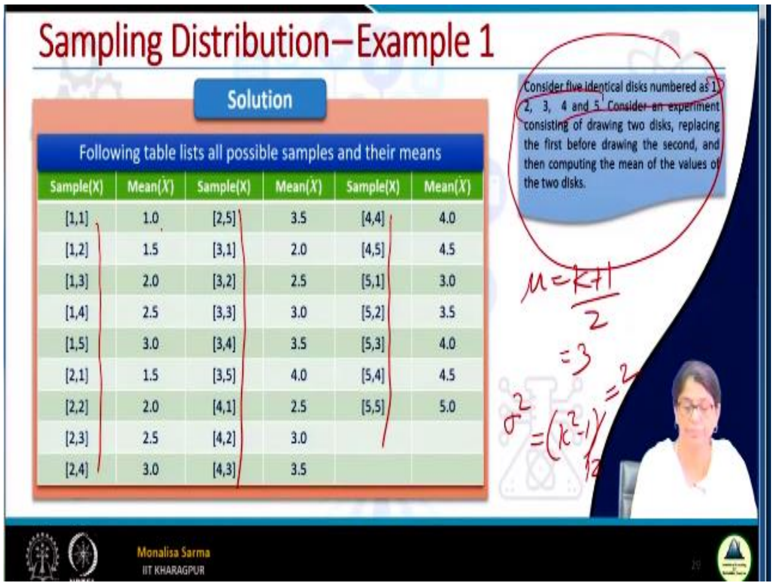


So, what is a bias sample, what is a non bias sample? I have already discussed. Now, coming to the example how we will find out the sampling distribution let us take a very small population suppose and we are considering a population where there are only 5 identical disks, this disks are numbered say 1, 2, 3 and 4 and 5 identity this is my population, that is very small population. So consider an experiment consisting of drawing 2 disks, replacing first before drawing the second what?

I will draw 1 disk I am replacing that and I will draw it again second one the maximum is what to say drawing 2 disks and then computing the mean of the values of the 2 disks. My interest is I want to compute the mean of the values of the 2 disks. Why I am doing that, my main objective is to find out the mean of the population. Of course, since this is just a 5 value, so we can I can very easily find out I mean, my if there is a population of huge number, it is really not possible to know the mean.

So here, let us not take it that way let us take this is a population, my intention is to find out the mean of the population somehow this mean of this population, I will try to find out by taking a sample. So what is my sample? Sample is 2 disks. So, how I am picking 2 disks, I am picking 1 disk and I am replacing it back again, I am picking the second disk, so I am replacing it back that is the way I am picking 2 disks.

(Refer Slide Time: 18:08)



So, this way, I am picking the disk see what these are the different samples I may get the total 5 samples the different possible samples, these are the samples you can see these are the different possible samples all possible here I have because this was a small population, it was possible for me to find out all possible values. So, I found that the all possible values of the sample these are my different way how I can pick 2 values 2 disks?

Now, what I have done is that so, if I pick knows see this, first of all consider does this 5 identical this picking 1 and replacing it. And again, I am picking 1, what is the probability of this picking, what distribution is first? This is definitely a uniform distribution is not it? It is same as rolling a die and coming which value is coming on the top it is the same thing. So it is very much a uniform distribution. So, uniform distribution, what is the probability f(x) is 1 / K where K is the total number of digits. So here what is the probability 1 / 5.

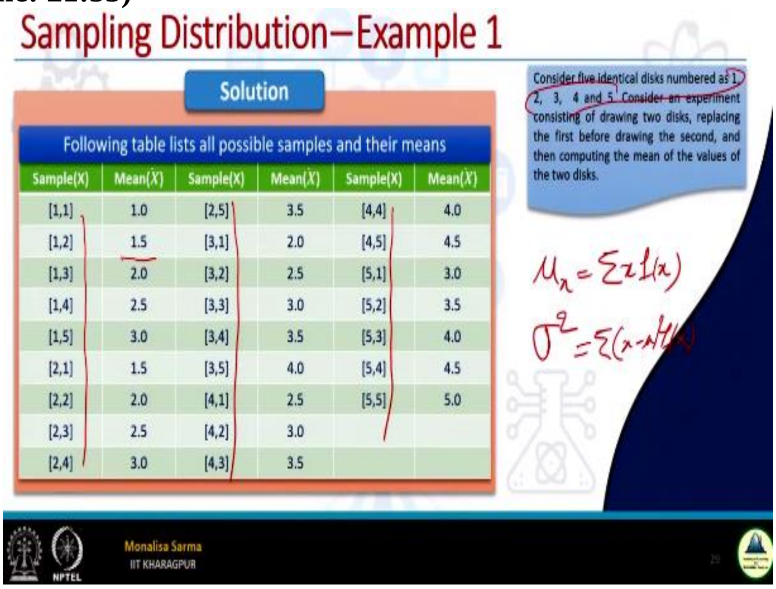
Since this is a uniform distribution, so now in a uniform distribution, can you remember what the how do we calculate the mean and the variance? So if considering this as the population, so if I need to calculate the mean, what is the uniform distribution calculation of mean if you can remember, μ= K + 1 / 2. Is not it? μ= K + 1 by 2. So what is here K? K is 5, 5 + 1 / 2 that is μ= 3 for this population that means my μ= 3 since a small populations I knew all about so I could find.

So, what is my variance of this population variance formula if you can remember K^2 - 1 / 12 variance formula we have discussed into when we have discussed any uniform distribution if we put the value of K as 5, so, what do I get? So, it is this I will be getting equals to 2. So, my μ= 2 for this population my μ= 3 and my σ^2 is equal variances = 2, that is for this population μ= 3 σ^2, because this is a uniform probability distribution that is now come to the sample what sample we have taken?

We have picked out all possible values of the sample all possible samples taking 2 together, these are, there is no other way by which we can pick 2 samples, there is all the possible scenarios are jotted down here then from all of these 2 samples, I tried to find out the mean. So, 1 and 1 if I pick 1 and 1 that is the mean. Mean is definitely 1 if I pick 1 and 2 mean is definitely 1.5 1 + 2 divided by 2 is not it? It is definitely 1.5 definitely I found that the mean.

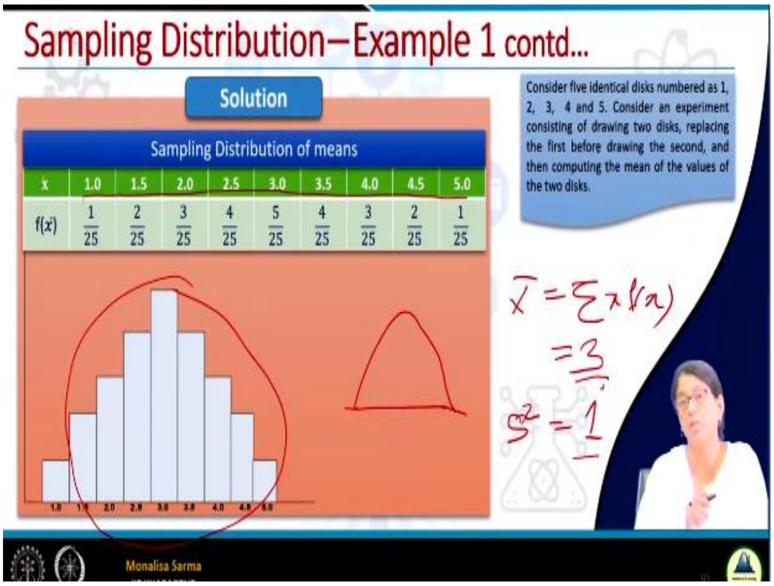
Now, if this is the case, if this is the sample this is the mean we know what is the formula for mean? In a probability distribution given a probability distribution when the distribution is specified, whether it is uniform distribution, binomial distribution or whatever distribution whenever distribution is specified, then we know for each specific distribution, what is the value of the formula, but actually what is the main formula? Main formula μ is x f(x).

(Refer Slide Time: 21:53)



What was mu of X? μ of x is summation of x f(x) is not it? This is how we find out that mu and mu have a distribution. And by putting values in this x and f(x) that how we have simplified and I found out different formula for different distribution, but this is the generalized formula for μ x is this and similarly, we have the general formula for sigma square as well remember is equal to summation of $(x - \mu)^2$ f(x).

(Refer Slide Time: 22:46)



So, now, this whole thing this whole table, if we want to put it in a form of a thing, what is that histogram that means if you want to put it in a form of a probability table, see how we can put it say this is the different values of what we got here the different values the mean takes the different values, we found out the different values mean can take this all these are the different values that mean can take there are no other values.

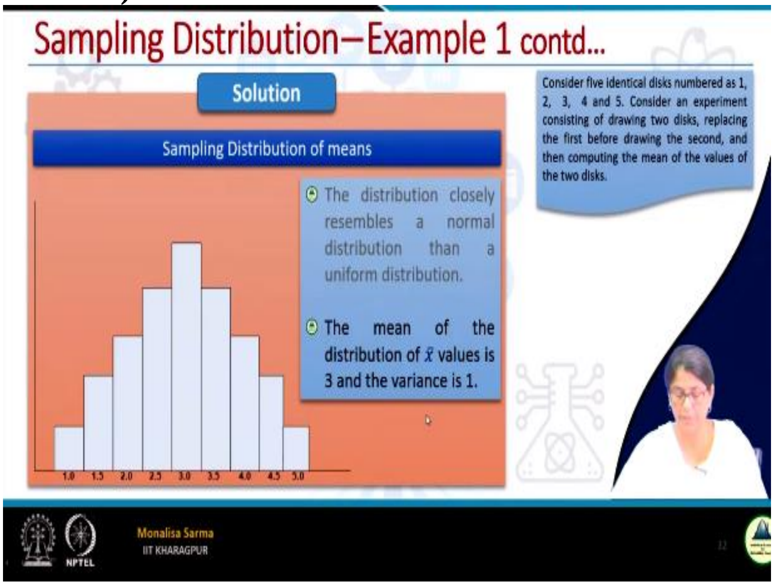
So, what is the probability that mean takes value 1 we have taken the relative frequency. So, relative frequency how much is the time that a mean takes 1 see here mean takes 1 that is only 1 value is there only once and how many total combinations there are total 25 combination. So, what is the relative frequency distribution is 1 / 25. Similarly, for 1.5 we found what is 2 / 25 this is how we found out this is nothing but the probability distribution this is in a tabular form, is not it?

This is nothing but the probability distribution in a tabular form, this is the X bar is a random variable f(x) bar is nothing but the probability of X bar. This we can see it in the form of a histogram also, this is the histogram of this corresponding data. Now, given this distribution, if we find out the mean, because I do not know what distribution it is, I do not know whether it is uniform distribution, normal distribution, this distribution I do not know what distribution if I want to find out the mean.

Then what is the formula for mean the general formula? General formula as I told is mu is equal to summation of x f(x). So, if I do x f(x) $\mu = \sum x f(x)$, x = x goes from 1 to n. So, x of f(x) means 1 x 1 / 25 + 1.5 x 2 / 25 + 2 x 3 / 25 + 5 x 1 / 25. You can calculate this and you will find that you will get 3. If you calculate this you will get 3 similarly, if you calculate the variance formula what this summation of $(x - \mu)^2 f(x)$.

So, if you calculate that σ^2 you will get that 1 I have done a mistake here. And since this is a we are talking of in terms of sample, so, I will not use the term μ and σ^2 what I will use remember and use X bar and here I will use S^2 because this is in terms of a sample, this is what I got in terms of for this probability distribution.

(Refer Slide Time: 25:24)

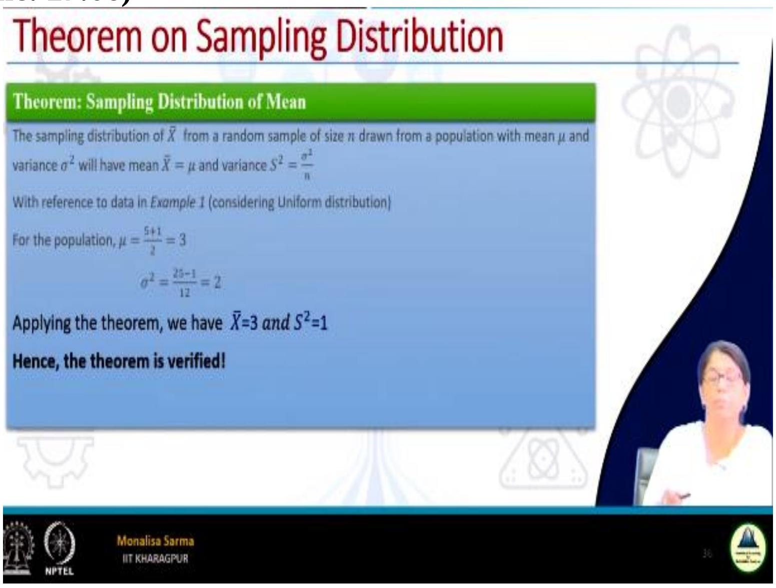


And this distribution you can also see from the table this is the table when we have drawn the histogram, but we have seen this distribution almost resembles a normal distribution, what is the normal distribution how is the normal distribution curve is this and this distribution also this figure also resembles a normal distribution. So, what we got from this calculation when we took all possible samples our sample of size 2 even if you take sample of size 3 whatever it is.

We know, for this explain suppose, we took all possible sample when we took all possible sample and we found out the probability distribution and this probability distribution very much resembles a normal distribution and thus the mean of this distribution we got 3 remember this population mean was also 3 K + 1 / 2 this is a uniform distribution that also is equals to 3 we found and this variance we found 1 and what was from this population, what was the variance we found variance 2 here we found variance 1.

So how is this link? Let us see here. So, this distribution closely resembles a normal distribution, our parent population has uniform distribution, is not it? Our parents population was very much a uniform distribution from this uniform distribution, we have taken a sample of 2 items and the resulting distribution that we got is our, this is a sampling distribution is not it? This resulting distribution is very much a normal distribution. And the mean of the distribution of X bar value is 3 and the variance is 1 if we calculate using the formula x of x, $(x - \mu)^2 f(x)$.

(Refer Slide Time: 27:06)



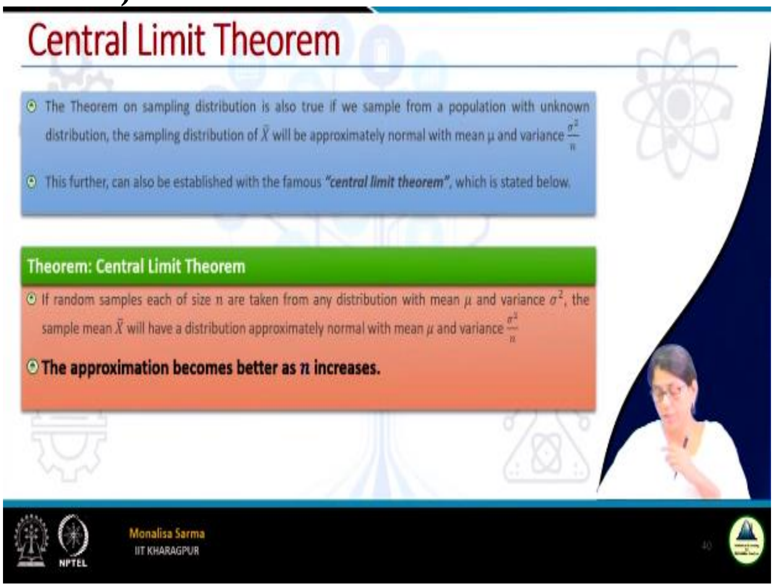
So, what we got? The sampling distribution of X bar from a random sample of size n drawn from a population with mean mu and variance sigma square will have mean X bar = μ and $S^2 = \sigma^2$ by n is not it? That is what we got for these experiment what we got? We got our X bar = μ, our populations mean was also 3 our sample mean also we got 3 that means our X bar = $\mu$and variance S^2 variance of the sample we got 1, what is the variance of the population of 2?

So, how we will get that means sigma square that is variance of the population divided by the size of the sample, what is the size of the sample? Size of the sample is 2 so 2 / 2 = 1. So, this is a very important theorem, this is sampling distribution of mean this is this we got for this experiment. Similarly, the researchers then experimenters they have done empirical study on many subpopulations.

Of course, infinite population, they cannot do they have done on very small population, medium population, large population they have done and from all those empirical experiments, they could come up with this theorem this very important theorem, I am reading the theorem again the sampling distribution of X bar what is X bar? X bar is a sample mean X bar from a random sample of size n drawn from a population with mean mu and variance sigma square will have mean X bar = μ.

And variance sigma square S square = sigma square by n this is called a sampling distribution of mean. So, now, with reference to the data from what we have uniform distribution already I have shown you we got $\mu = 3$ $\sigma^2 = 2$. So, our theorem is basically proved we could prove this theorem basically, we already know that theorem and we have done an example to prove, but actually the researcher has done many experiments together and they could come up at last they could come up with this theorem, this is called sampling distribution of mean.

(Refer Slide Time: 29:22)



So, this theorem on sampling distribution is also true if we sample from a population with unknown distribution, this is what this population what we have considered this is a very much a known population, we know that it is a uniform population, even if we sample from a very much unknown population also the sampling distribution of X bar will be approximately normal we have seen it is approximately normal this size the sampling distribution of X bar will be approximately normal with mean μ and variance σ^2 by n.

So, this further can also be established with the famous central limit theorem this another theorems central limit theorem, this is basically again similar based on same set of experiment one most stronger theorem than the sampling distribution and sampling distribution theorem, we see the sampling distribution theorem here it does not speak about the shape of the sampling distribution say this is the sampling distribution of mean.

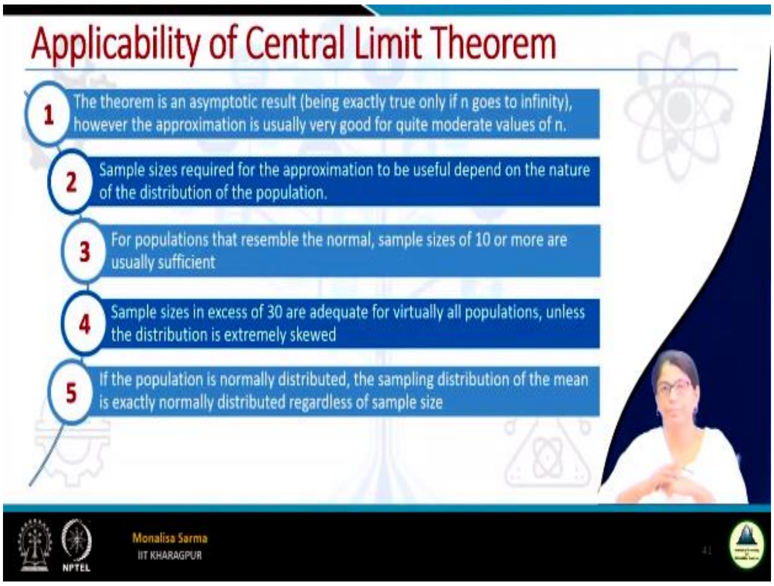
The sampling distribution of X bar from a random sample of size n drawn from a population with mean μ and variance σ^2 will have a mean X bar = μ and variance as far as well as σ^2 / n. It does not speak about the shape of the sampling distribution remains silent on that. So, but already we have seen we have seen the sampling distribution take same normal shape.

So, central limit theorem is a more stronger definition than sampling distribution, it also talks about that. So, what is the central limit theorem say this is a central limit theorem, the very important theorem, if random sample each of size n are taken from any distribution with mean μ and variance σ^2 μ and σ^2 is the parameter of the population the sample mean X bar will have a distribution approximately normal with mean μ and variance σ^2 / n, yes everything same.

But here also it talks that a distribution is approximately normal. So, it is irrespective of the parent population whether the population your normal population, binomial population, normal population whatever population it is, but this theorem it satisfies this theorem, but there is another twist to it that means, if the population this approximation becomes better as n increases this is approximation. This approximation mean is value of μ and value of X bar and S.

But the size or the shape that shape I told that errors distribution is approximately normal that distribution is approximately normal distribution becomes better as n increases for very small value of n, we may not get a normal distribution, when the n becomes more when n is becomes greater than definitely we will get a normal distribution.

(Refer Slide Time: 32:19)



But here, there is another twist to it also see the different points the see all the points are very important, please look carefully that this theorem is an asymptotic result that means being exactly true only if n goes to infinity when our sample size becomes large, this theorem is exactly true. However the approximation is usually very good for quite moderate values of n. Sample sizes required for approximation to be useful depend on the nature of the distribution of the population.

Now, when I told that is when the sample size goes to infinity, this approximation is very good again in fact, it is exactly true when sample size goes to infinity and even for a moderate values of n also this approximation is very good. Now question that arises what should be the sample size? So, this approximation depends on the distribution of their parent populations. If my parent population is normal with my parent population from where, I am taking the sample.

If my parent population is normal, then we do not need a very bigger sample size and a small sample size will also do but in the population is a bit away from normal it is not normal, but a bit totally not normal, but a bit away from normal then we need a moderate value of n, but it a population is very much skewed then we need a very big value of n. So, population that resembles a normal sample size of 10 or more are usually sufficient if it is normal or if it resembles normal not exactly normal.

Sample sizes in excess of 30 are adequate for virtually all population unless the distribution is extremely skewed. It does distribution is extremely skewed then we need a very big sample size maybe to tune of say 100 if the population is normally distributed, the sampling distribution of the mean is exactly normally distributed regardless of the sample size. So, I there are a few more points to be discussed about this sampling distribution about this central limit theorem basically. So I end this lecture here. In the next lecture, I will cover those things. Thank you.