Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology - Kharagpur

Lecture - 17
Sampling Distributions (Part 2)

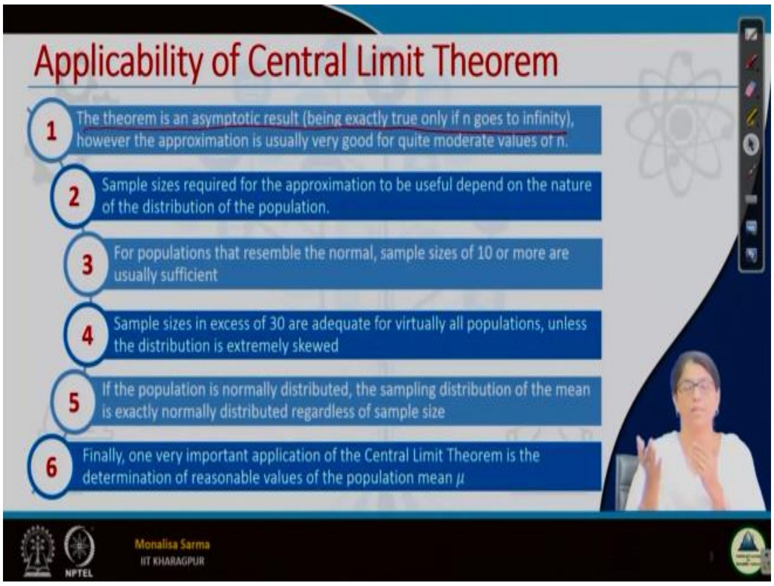
Hi, welcome back again so, last class like we were discussing sampling distribution last class was the first class on sampling distribution.

(Refer Slide Time: 00:35)

And then, so, in the last class, we have covered what we mean by sampling distribution. So, and then we have also seen the uses of sampling distribution, then we have discussed central limit theorem, but why I have given this as a concept covered because this class is basically a continuation of the last class, I could not cover the whole portion because of the time constraint.

So, I have given the same concepts I am showing it here as well. So, it is just basically the extension of the last class.

(Refer Slide Time: 01:06)



So, now, where we have stopped last class is that we were discussing the applicability of central limit theorem, remember, what was some central limit theorem. So, basically, when we have to infer something out of populations, because population is a big number, we cannot infer and infer something specifically we some summary statistics, like say, we want to know about the mean of the population.

We want to know the variance of a population, like say, we want to compare 2 different populations. So, these are the things when we want to know something about a population, since it is a big population is a big unit. So, what we do is that we take a sample out of it from the sample, we calculate the statistics. And from the statistics, we try to infer about the population parameter, I am reminding you again when we talk of sample we call it statistics.

And when we talk of populations, the characteristics of a population we call it parameter. So, statistics is the counterpart of parameter in sample basically. So, we have already seen why sampling distribution first, what is sampling distribution that I think you can remember sampling distribution is nothing but a population distribution only. And why we call it a sampling distribution?

Because, here the random variable that the random always when we have a population, probability distribution, it is always for a random variable, is not it? A probability distribution is always for an random variable, why? Because the; random variable can take any values, and then what is the probability of taking these different values. Now, here in case of a sampling distribution, my random variable is a sample statistics, is not it?

So, my random variable is a sample statistics. So, and this sample statistics, there are lots of variants to the sample statistics; if we take different samples, we will get different statistics value will be different, there is a variation since there is a sampling variation. So, to and address these variations, what we do is that, we find out its distribution. And that is why we call it as a sampling distribution.

We do not call it probability distribution; we call it sampling distribution and since, we have already seen in the last class, how to, what to say, basically, how to have the sampling distribution, how to construct the sampling distribution. To construct a sampling distribution, we need the different values of the random variable we will take that means different variables that this sample statistics will take different value as well as its probability of occurrence.

If I let me not say the probability of occurrence let me say is the relative frequency of occurrence. So, how do I find that? So, that is really not possible to find out the different value that the sample statistics will take as well as the frequency or the relative frequency observation and that is really not possible, because for a sample, we have to take all possible sample for a population will have to take all possible samples from all possible samples only then we can find out its relative occurrences is not it?

So, since that was not possible, so, we took help of some theorem. Now, what is that theorem? The first theorem that we have learned is a sampling theorem, is not it? So, sampling distribution of mean; that was the first theorem. So now, we have also proved that theorem how we prove that theorem for proving the theorem, we have taken a very small populations, our population.

That population if we can remember we are considered a what to say a uniform population from the uniform populations, we tried to take out the sample, sample of size 2, and for least sample, we try to find out what is the mean and then from that, we could basically calculate different statistics of the sample as well as since it was a uniform population, we have calculated the parameter values of the samples that was the mean and the variance.

And we could prove that, the theorem, which is there the theorem is correct, anyway, we do not need to prove it this big, big researchers they have done it, but for our own curiosity we saw it is really true, but we definitely cannot do it for a bigger population. So, for a smaller population, we have seen that is right, but the sampling distribution of mean that distribution, it did not speak about the shape of the distribution, what distribution is the, it just talks about the mean value as well as the variance value.

Then we infer that we have referred to a more stronger theorem, that is a very well known theorem, that is the central limit theorem, the central limit theorem not only talks about the mean and the variance of the sample statistics of the sampling distribution, but it also talks about the shape of the distribution. So, what was the shape? It is approximately normal. Now, then, we talk about the applicability of the central limit theorem here basically, we stopped.

So now, what was the applicability? Applicability also, I have discussed first of course, this theorem is exactly true when the sample sizes sample size goes to ∞, but then that is also not necessary if it has a moderate value that also it is good and it if our parent population is normal, then a sample size of around 10 is also sufficient a dependent population is it is not normal, it is not very much also skewed then a model size of 30 is enough, but the problem sample is very much skewed, then we need a very bigger sample size.

Now, finally, one very important application of central limit what is the important application of central limit theorem? Can you just tell me with the knowledge of probability distribution? From this probability distribution, where we have used probability distribution till now, I have discussed probability distribution in many numbers of classes. So, one of the main uses of probably distribution was given a probability distribution, if we have the distribution.

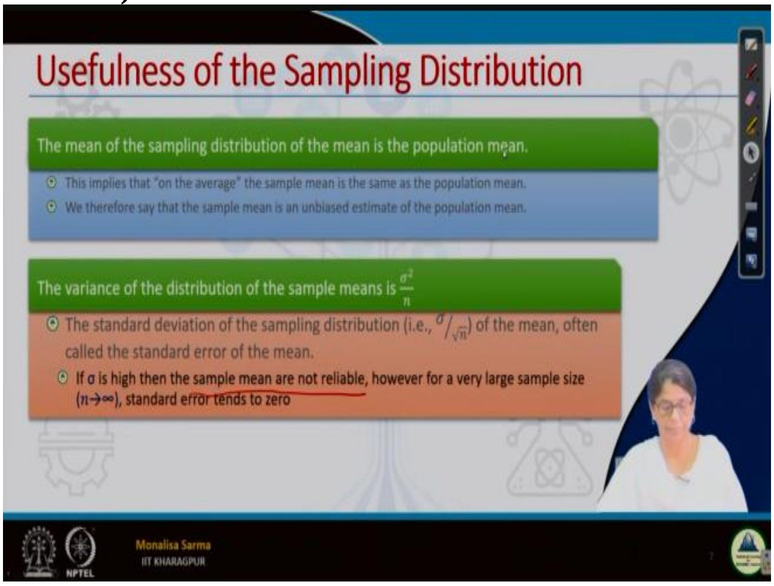
If we have known the mean of the distribution, if we know the variance of the distribution from that distribution, we could find out the probability of occurrence of a particular value, is not it? Like suppose we have a probability distribution of the lifetime of a bulb. So, if lifetime of a bulb follows a particular probability distribution, if we know that if we know the mean of the distribution, if we know the variance of the distribution, it is normal distribution, definitely central limit theorem means we are talking of normal distribution.

So, if we know the mean and the variance of this distribution, then everything is known to us like then from that we can find out the probability of occurrence of any value like if we are interested suppose, if you are interested in finding out what is the probability that the lifetime of the bulb is say around them, say 2 years. So, we have the distribution, we have the mean we are the variance from that we will be able to calculate.

We have done this sort of example, if you cannot remember please I suggest please just to have a quick recap of those probability distribution thing, because we will be needing it very much in this coming lectures as well. So, that is what so, one important application of central limit theorem is the determination of reasonable values of the population means that is the whole story, is not it?

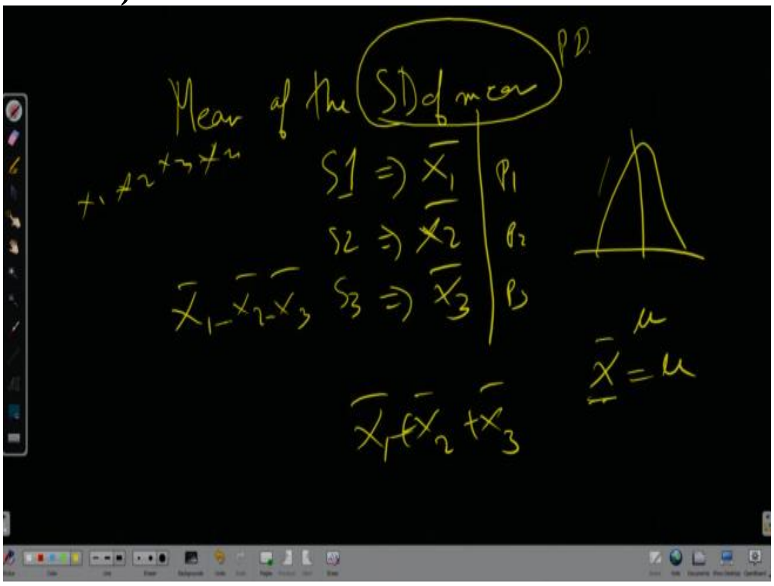
That is why only we are doing that, why are we can I am developing the sampling distribution, we have to talk we want to talk something about the population. So, what we want to talk? We want to tell about the population we are mind it here, we are talking about sampling distribution of mean we are finding out the sampling distribution of mean that means why we are doing it, we want to talk something about the population mean. So, this is one of the important applications of central limit theorem.

(Refer Slide Time: 08:50)



So, the sampling distribution of the mean what we have seen it is the population mean, is not it? Sampling distribution so, when we have drawn a sampling distribution, how we have drawn different means, different samples have different means, and it is different probability of occurrence based on that we have drawn the sampling distribution, is not it? Now, what is this mean? The mean of the sampling distribution we found it is equals to the population mean.

(Refer Slide Time: 09:23)



So that means see so, 1 sample suppose I have taken sample 1 and my this population is say X 1 bar another sample my population is say X 2 bar another sample say my population is say X 3 bar. So, along with that, I have also probability of this occurrence, say some probability p 1 p 2 p 3, whatever it is from that I found my sampling distribution that is not an issue. Now, my issue here is that, what is the mean of the sampling distribution.

I already from the central limit theorem, I have known that my distribution shape is normal, and I have known my mean is equals 2 population means now the population mean is μ that means my sampling distribution mean is also it is μ. Now, what is this X bar? X bar is basically mean of all this is not it? Mean of X 1 bar + X 2 bar + X 3 bar, this is nothing this mean of all this is my X bar.

That is, that is when I talk about mean of the sampling distribution, mean of the sampling distribution of mean see the statement properly. What is sampling distribution of means? Sampling distribution means, means I am talking about the probability distribution. So that means the mean of the call the distribution, so what is that that is my X bar. So what is X bar? X bar = μ, that means what?

On an average this implies that, the sample mean, X bar is what mean means on an average, the sample mean is the same as the population mean. This is one so suppose I do not know the population mean, I took a sample from the sample, whatever mean I get, then I can say on an average population mean will be the sample mean only, is not it? So we therefore say that sample mean is an unbiased estimate of the population mean.

We will be seeing many applications from this application, the concept will be more and more clear. Now, even if you have some doubt, I am sure things will get clearer when we do the problem. So we therefore said a sample mean is an unbiased estimate of the population mean. Now, what is the variance of the distribution? See here, here, when I am talking about a variance; what I am talking about a variance?

Let me take this again, say this here, when I am talking about the variance, what is this? Variance among this X 1 bar X 2 bar X 3 bar variance among these how these values differ? When I am talking the variance of the sampling, distribution mind it, I am talking about how when you talk a variance of any number means what there are a few numbers X 1 X 2 X 3 X 4 when I talk of variance means how these numbers are varied, is not it?

So this is when I am talking about the variance of the sampling distribution, I am basically talking about a variance among the mean among the different means of the sample. So, the variance of the distribution of the sample mean is σ^2 / n that we have seen, so now that means, if the variance is σ^2 / n, so, what is the standard deviation? Standard deviation is $\sigma / \sqrt{n}$, this standard division is also called a standard error of the mean why it is called the standard error of the mean?

Means, this is the difference among different means of different samples, we suppose, we took 10 means, so, we found a 10 samples from 10 samples we got 10 means, so, the difference in this means, this is nothing we are calling it as a standard error of the mean. Now, if the population in the population σ is very high if my population is a very varied population variance is very high.

Variance is very high means standard deviation also be very high σ^2 / n σ will also be up in my population. In my population and variance is very high then what happens my when my sampling distribution what is the variance of the sampling distribution? Variance of the sampling distribution is σ^2 / n variance of the population is high means my variance of the sampling mean will also be high is not it?

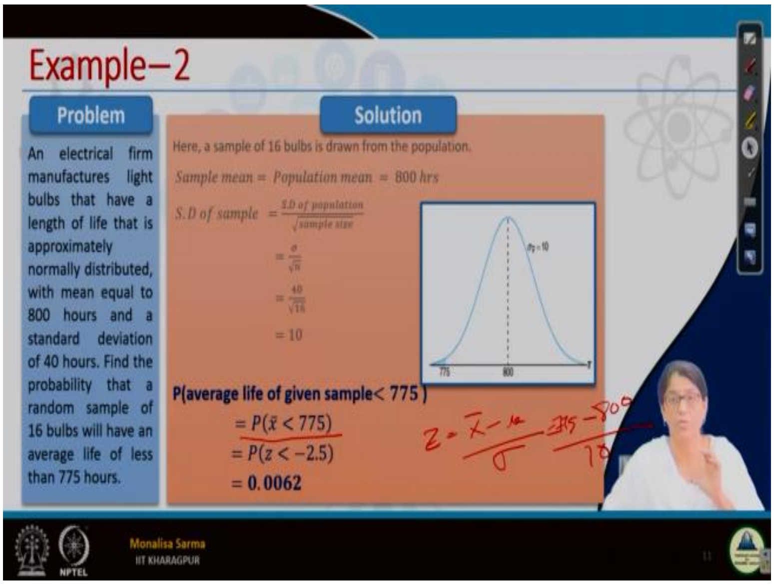
So, that means, 1 population sample mean suppose I maybe I got 5 another I got a 20 and other I got a 25 another I got a 41 I got an 80 is not it? If my population variance is high, my the variance of the distribution of sample means will be high. In such case, when my variance among the different sample mean is high, that means, this statement on an average the sample mean is same as the population mean this statement is not a very reliable statement.

That means, if my σ is higher than the sample mean are not a reliable estimate of the population mean. If the population variance is very high, suppose I am talking about the lifetime of the bulb, 1 bulb say may run 10 hours another bulb say more than 200 hours, another bulb say may run 100 hours, another bulb say run 500 hours. So, it is very much varied figure. In that case, when I take it from sample, from the sample, whatever I take the mean, that sample mean is not a reliable estimate of the population mean.

That is why I am telling when σ is higher than the sample mean are not reliable. However, for a very large sample size, because see and σ^2 / n, if n is the denominator, it is more than my sampling variance will also be less, is not it? So for the large sample size my standard error tends to 0, but if I take n is very large, then what happens, my difference becomes very, very less difference is what?

Variance of the distribution means the difference among the means that becomes that comes very less. So that is why my first statement on this central limit theorem being exactly true if n goes to ∞, that this theorem is being exactly true only if n goes to ∞. If n goes to ∞, that what happens? Sampling the variance among the sample mean is it tends to 0 then on an average sample mean is equals to population means that is a very reliable estimate.

(Refer Slide Time: 16:19)



So now we will just do 1 problem, definitely, we will be doing more problems in our tutorial class also, just 1 problem so as you understand the things a bit so see the problem. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equals to 800 hours and a standard deviation of 40 hours. So what it is given an electrical firm, it manufactures light bulbs.

So and its life, it is normally distributed. It is this firm, it is claiming that the life of the bulbs is normally distributed. And it is also claiming that its mean is 800. That means my μ is 800 and standard deviation that is my σ is 40 hours. This farm is claiming mind it. So suppose I am the owner of the farm, I am claiming that the mean life is 800. And the standard deviation is 40 hours, what I am claiming may not be true.

Because I am just it may be an educated guess, basically, I am predicting something out of some past experience or whatever it is, but it is not possible for me to tell that is exactly true. Why? Because how can I find out the life of all the bulbs? I did not check off all the balls, how can I tell? So this is just a manufacturer that is claiming. Now find the probability if that is true assuming whatever he is telling is true.

Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. So look the question other way around. I have gone to purchase some electric bulk from a manufacturer and the manufacturer has told me this my life is normally distributed. I am educated person I know about distribution and all. So, that is why he talked in terms of technical terms, he told me this, life for the bulb is normally distributed.

And its mean is 800 and standard deviation is 40. They told me that, but I am not such a gullible person that I will believe everything I wanted to check what did I do? I took a sample of 16 bulbs and I have used some technique to find out the life of those bulbs, immediately I could not find definitely took time for me before purchasing a whole lot. So, I took a 16 bulb and from that I found try to find out the life of those bulbs the life of those bulbs, what I got the average life of those bulbs I got 775 hours.

Now, from this, this is what from the sample what a good that is correct value, is not it? I have checked it I have checked it and I found that it is 775 hours. So now assuming that what the manufacturer have claimed that mean is 800 and standard deviation is 40. Is it really possible for me to get an average level of 775 hours that I will see if it is possible? If assuming that is true, if it is possible from a sample to get an average level of 775 hour if it is true.

Then what the manufacturer is claiming I can tell that is true. If it is not possible, then I can tell that what the manufacturer is claiming that is not true. H is trying to fool me. Now how it is true or not true how can I tell? When I will, find out the probability of this having a life of less than 775 hours. If this probability is very, very small, then that means the manufacturer is going to fool me that is not true.

If this probability is quite high, because while doing the calculus I am assuming that whatever he told me is correct with that assumption when I found that this sample having an average life of 775 hours, if it is probability is quite good as a good quality then. The manufacturers his claim is correct. So, how we will do that? So, that means, we will be using the sampling distribution of mean, is not it?

So, sampling distribution of mean for that using sampling distribution I mean, what will be the mean of the sample distribution population mean that is 800 what will be the standard deviation of the population mean because I need to find out the probability of less than 775 hours that is our sample is not it? I need to find a probability of less than 700 from a sample. So, for to find a probability; I need to probability distribution what probability distribution?

Sampling distribution of mean. So, now, when we talk about normal distribution, when I discuss normal distribution, I think you can remember there are 2 parameters which can describe a normal distribution if I know those 2 parameters, I can easily find out the normal distribution what are those 2 parameters? Mean and standard deviation or mean and variance whatever it is now, for me in the sampling distribution.

I know my mean of the sampling distribution is population mean 800 standard variance of the sampling distribution is σ^2 / n what is my sample size? Sample size is 16 right see here. So, what is my sample mean? Sample mean is 800 hours then what is my standard deviation or variance where basically whatever whichever you find out then a standard deviation of the sample standard deviation of the population by square root of sample size if I find out sample variance.

So, variance by sample size is not it? So, the $\sigma / \sqrt{n}$ what is σ? Σ is 40, 16 is my sample size. So, I got my standard division is 10 so, this is normally distributed this is basically 800 you can see the figure this is 800. This is the mean and if this is correct, I want to find out what is the probability that it is less than 775 it is definitely it will be to the left of the mean because mean is 800, 775 which is less than.

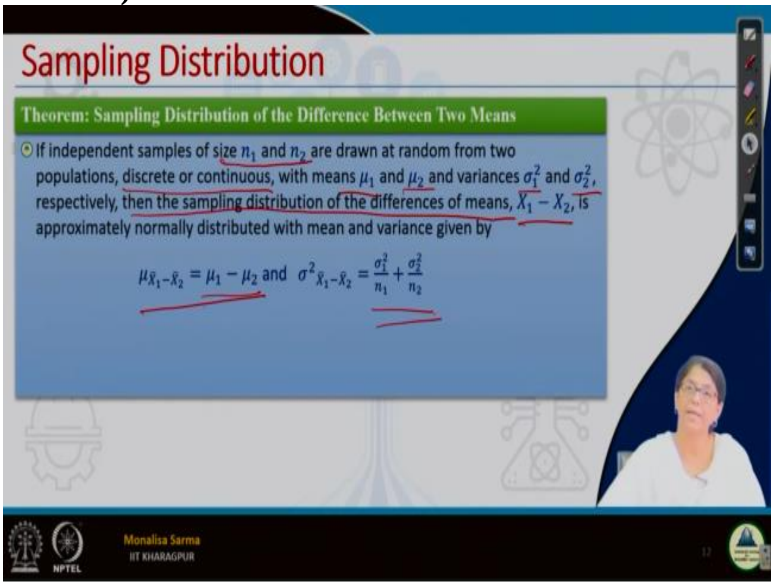
So, I need to find out the probability of having this value 775. How do I find out? So, I need to find out this probability we have seen normal distribution while we have solved problem using normal distribution first, we will calculate the take the value to the standard normal value that is the z value from z value we can construct the table I have done many problems on this the I do not think you will forget that because we have done many number of problems.

So, now, first thing is I have to find out the z value of this. So, how do I find out the z value of this? How does that remember z equals to? What is the z equals to? $x - \mu / \sigma$ is not it? So, now, what is my x? x is 775 μ is 800 the σ this is σ. So, that means, I need to find a probability x bar less than 775 this I have converted to z value. So, what is z? z is equals to here it is my x is x bar, x bar - μ / σ.

So, what is x bar? x bar is 775, 775 - μ is 800 / 10. So, z value is less than probability that z value is less than - 2.5. Here, I am not showing the table I have shown in many classes, how to consult the table. So, from the table we can find out so, what is the z value? What is the probability corresponding to - 2.5? It is 0.0062. So, this is a very, very less probably not even 1% probability is not it?

However the question is not asking that, but we can consider that we can tell that we can claim that the manufacturer whatever manufacturer is claiming is not correct, the life of the bulb is not 800 if the life of the bulb is 800 with the standard division of 40 hours, I would not have from the sample result I would not have got a sample life of 775 hours, because probably this is true, probably of getting this I found it to be very small that means this is not true. Understood the use of sampling distribution at least we will do many more examples.

(Refer Slide Time: 24:27)



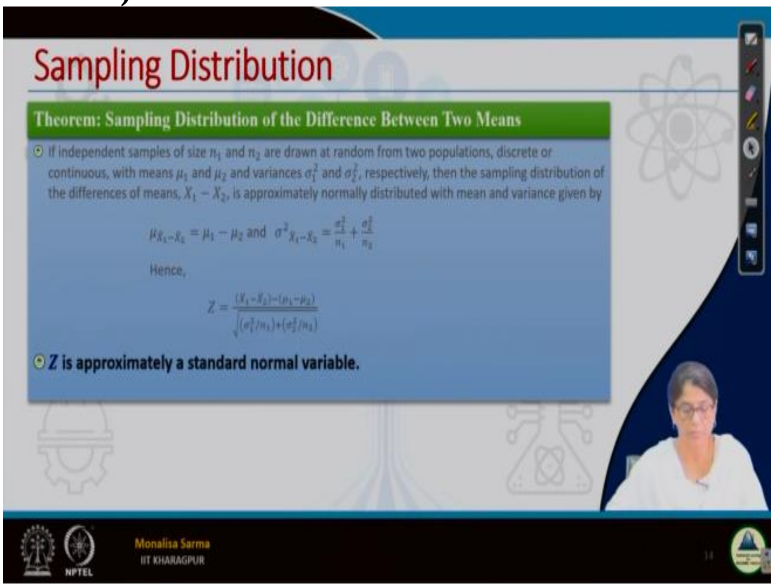
So, there is one more theorem. This is again an important theorem used in many like when I try to compare 2 different populations we will be; needing it again and again. See, go through the theorem if the independent sample of size n 1 and n 2 are drawn, size n 1 and 2 are drawn at random from 2 populations, n1 I am drawing from population A n 2 I am drawing from population B to different population, whether discrete or continuous.

And say first population means μ 1 second population means μ 2 and first population variance is $\sigma\ 1^{2}$ second population the variance is $\sigma\ 2^{2}$s. Now, if I want to find out the if I want to compare these 2 populations, whether I am interested in finding out the difference of to support difference in means, suppose there are 2 different companies are producing this LED bulbs 2 different company A is producing LED bulbs.

Company B is producing LED bulbs, suppose, I want to find out the mean lifetime of these 2 bulbs the mean lifetime of this how much is a difference in the mean lifetime on these 2 bulbs? Then what I will do I will find out it will be substrate. So, then the sampling distribution of the difference of means, that is x 1 bar - x 2 is approximately normally distributed with mean and variance given by mean will be μ 1 - μ 2.

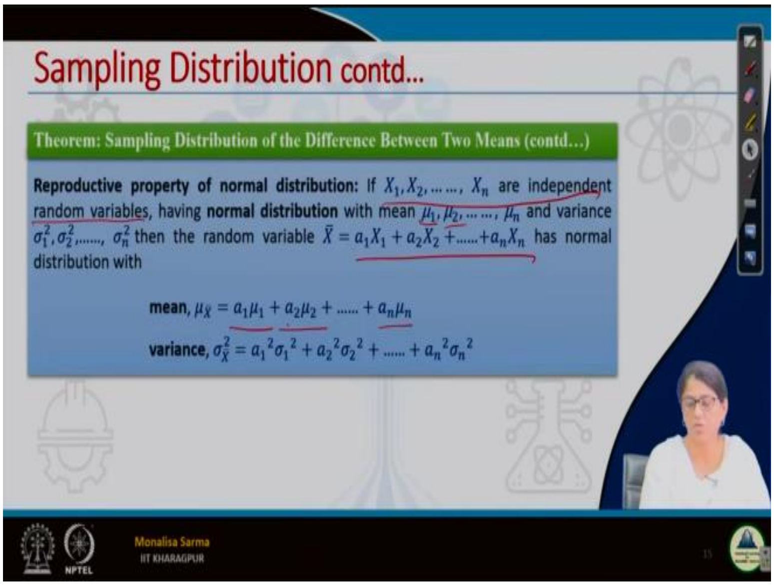
And variance I already talked variance, we never subtract variance always get added up we have also done that variance results on that one we have solved this is not it? Variance, if you find out the difference of 2 variances, it is always added up. So, this is the if we are interested in finding out the sampling distribution of the difference of mean this sampling distribution will have a mean of μ 1 - μ 2 and variance will be an addition of both the 2 variances.

(Refer Slide Time: 26:43)



So, now, to find out the problem, we definitely we will have to find out the z value. So, what will be my Z value? Z value will be x 1 bar - x 2 bar what is the basically that means the x value and this is my μ is μ 1 - μ 2 and this is minus this is my standard deviation. Now, this z is approximately a standard normal variable.

(Refer Slide Time: 27:12)

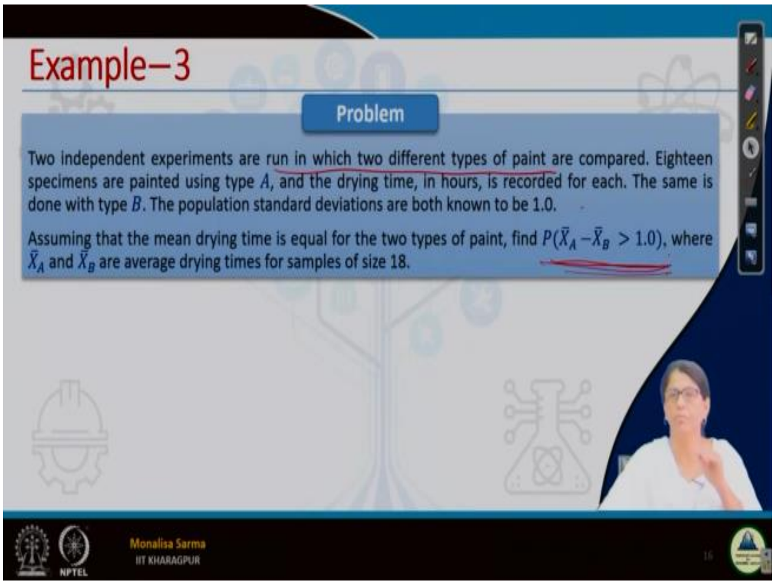


So, this is basically a corollary of the previous theorem we can say see if X 1 X 2 X n are independent random variables having normal distribution if X 1 is a normal distribution X 2 is a random variable having normal distribution X 3 is a random variable having normal distribution all independent random variables and all have normal distribution. And each has different means, suppose X 1 has mean μ 1 X 2 has mean μ 2 like likewise X n has mean μ n and variance.

Similarly, σ 1 ^{2}, σ 2 ^{2}, up to σ n ^{2} then if we are interested in finding out an X bar. What is X bar? X bar if I am telling it this is a relation of X bar to all the other random variable is some linear we are using some constant a 1 a 2 a 2 these are all there is a linear relationship between among the random variables. So, if my X bar is a 1 X 1 a 2 X 2 a n X n then this definitely will because all X 1 X 2 X n are normal distribution then definitely X bar will also be normal distribution and what will be the mean of X bar?

Mean of X bar will be a 1 μ 2 a 1 μ 1 a2 μ 2 up to a n μ n. So, like if we consider the previous theorem it is the same here a 1 and a 2 is basically a 1 is 1 a 2 is - 1 is not it? 1 so, what it will be μ x? μ 1 - μ 2 variances variance all gets added up.

(Refer Slide Time: 29:04)



So, to use this theorem a small example, just go to this problem 2 independent experiment are run in which 2 different types of paint are compared we want to compare 2 different types of paints we have taken 18 specimen from type A how we are comparing we are comparing based on the drying times which drying time which paint takes for drying times we are comparing trying to compare these 2 population.

Based on the drying times and the drying times in hours is recorded for each the same is done with type B the population standard deviation are both known to be 1 we know the population standard deviation of both the population is 1 A as well as B assuming that the mean drying time is equal for the 2 types of paint. Suppose what to say producer has claimed that mean drying times.

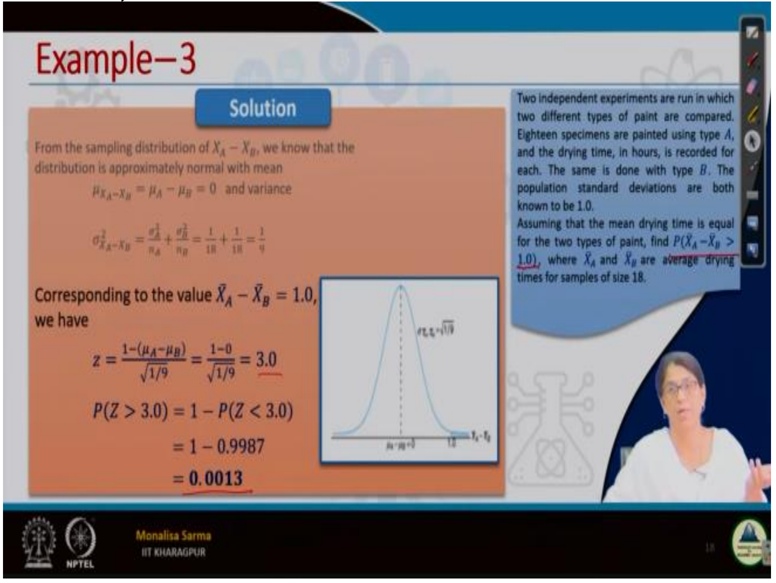
Producer of type A has claimed a particular mean drying time same claimed that the drying time is say y 1 the producer of paint B it has also claimed that mean drying time is y 1 both that means both the mean drying time is same. So, now, we want to compare whether this is really true than what we have taken we have taken a specimen of both types of specimen of size 8 and we found that the mean drying time of both on an average it is greater than 1.

So, X A bar - X B bar is greater than 1. Now, if this is true, if the mean drying time of population, we consider the population mean drying time of both the population is same and both the population standard deviation is same is 1 how much probability is that we will get distinct probability that X A bar - X B bar is greater than 1 because this is something which we have taken the sample and we got the result this is something which the result which he got.

Now, we will try to find out same as the previous question, we will try to find out what is the probability of this occurrence assuming this is true, what the producer is claiming assuming that is true, same way, if this probability is high, then the producer is whatever the producer is claiming that is true, that means the mean drying time of both the paint is same and it is standard deviation is 1 that is true what they are claiming is that if this probability is very low that means what they are claiming that is not correct.

Anyway, this question is not asking that. This question is just asking what is the probability of this if we calculate this that is done, but try to understand what is the meaning of this problem? That is why I have given all those explanation.

(Refer Slide Time: 31:46)



So, similarly, we have found what is this? As I as you have seen here, how do we find out the z variable this is the z variable X 1 bar - X 2 bar - μ 1 - μ 2. So, μ 1 - μ 2 is 0 because both the population mean is same is not it? And what is the standard deviation of this one 18 + 1 / 10 root over because we have taken a sample of 18 and X 1 bar - X 2 bar it is we are trying to find for greater than 1 so we will find out a z value keeping it 1.

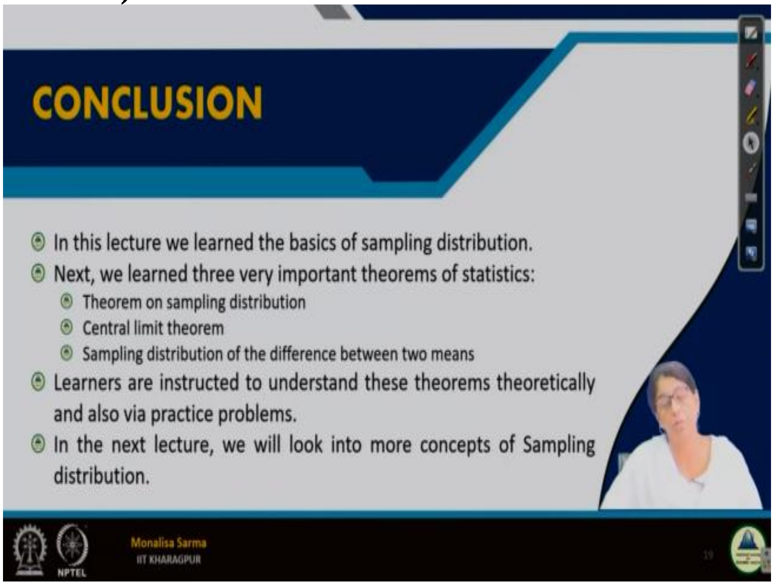
So, this is my this is equals to 0. This is my what to say σ^2 then what will be my Z value? Z value is this, so, I got this a probability of Z greater than 3, because I am interested in finding out P of Z greater than 1. So, probability of Z greater than 3 it will be nothing but 1 minus.

So, because normal distribution remember we always normal distribution is always specified in terms of cumulative distribution, cumulative distribution means it is possible from - ∞ to X that particular value area interest.

So, when we interested in finding out greater than 3, so, it will be 1 - P of Z less than equals to 3. So, from this value, we will get it from the normal table. I am not saying the normal table again and again I have shown it in many classes. So, this is the value we will get. So, this is the probability this is again a very low probability what does this indicate this is a low probability means what does this simple value cannot be wrong.

Because we have taken this we have checked it. So, what can go wrong what the producer is claiming? Because that is based on guesswork that is guesswork, maybe an educated guess maybe based on past experience, but there is no proof to that, but this we have done it and we have got the results.

(Refer Slide Time: 33:48)



So, and I should not say in this lecture in this coming 2 in the last 2 lecture, this lecture and the last 1 lecture, we have learned the basics of the sampling distribution, we have learned 3 important theorems of statistics like theorem of sampling distribution, sampling central limit theorem, and sampling distribution of the difference between 2 means, that also we have seen.

So, as always, I will again ask all the learners to understand this theorem theoretically as well as via practical problems. In the next lecture, we will look into some more concept of sampling distribution.

(Refer Slide Time: 34:25)

So, these are my references and thank you guys.