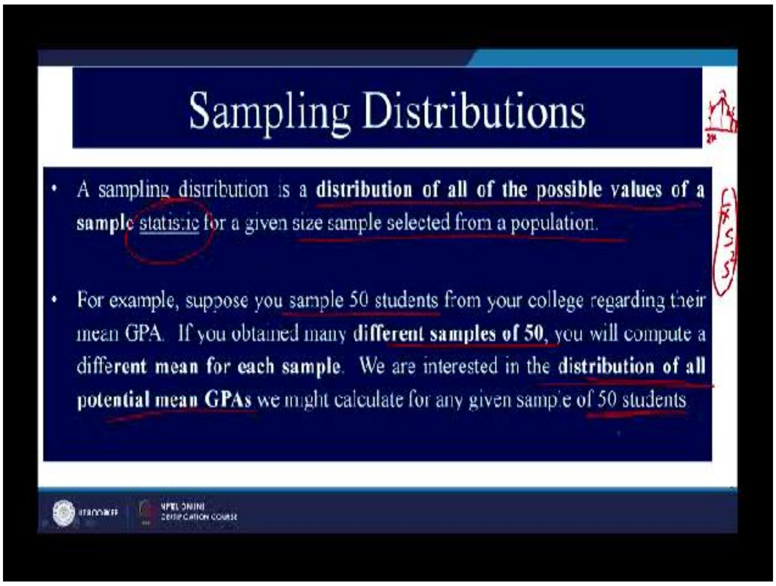
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology-Roorkee

Lecture-23
Sampling Distributions-I

Good afternoon friends, I welcome your all in this session as you are aware in previous session we discussed about sampling and sampling techniques. In today's session we will see what is sampling distribution, we have seen probability distribution, what is probability distribution. Probability distribution is a distribution of outcomes of an experiment. So let us say just to recall what we did in probability distribution.

(Refer Slide Time: 01:10)



Let us say you are contesting an elections and the probability that you will get out of let us say total 10,000 votes, probability that you will get 2000 votes is 0.3, probability that you will get 3000 votes 0.4 and probability that you will get 1000 votes is 0. let us say probability that you will get 1000 votes 0.25 and probability that you will get fall 100 votes is 0.05. So you can have this is 0.05 right.

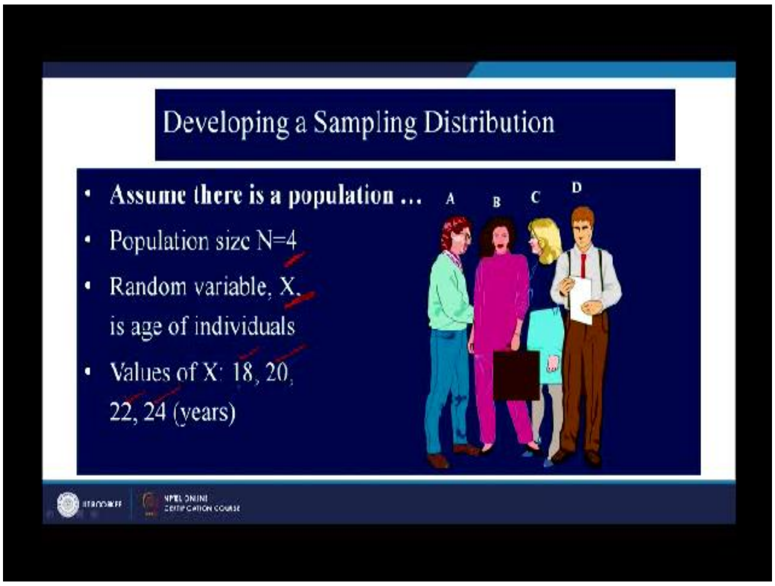
Probability of getting 500, so this distribution is nothing probability distribution right, so similar to that we have what sampling distribution whenever we take a sample from population and if we calculates its mean then for every sample the mean would be different and when you draw the distribution of those sampling means it is call sampling distribution of sample mean right.

So a sampling distribution is distribution all of the possible values of sample statistic right, what is sample statistics it is about it is X bar right is not it, X bar standard deviation is not it variants, it is all about sample right. For a given sample size selected from population. Let us say again let us say you will select 50 students from let say population of 8000, so you can select this 50 students by having different ways is not it.

So let say you have selected 1 sample of 50 students find out you found out mean again second sample of 50 students, then mean would be different is not it, so when you draw the distribution of these sample means it could be called sampling distribution. So we are interested in the distribution of all potential let us take an example here. So suppose you sample 50 students from your college regarding their GPA you obtain many different samples of 50.

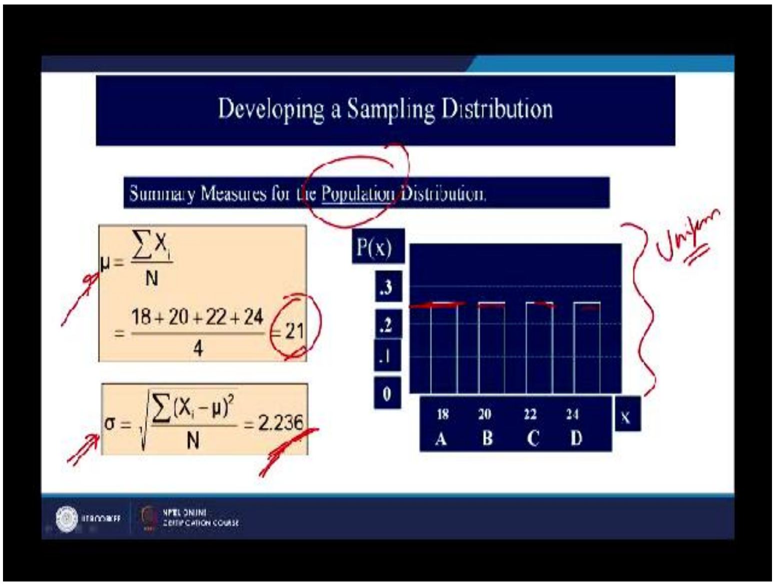
For every sample of 50 you will have different mean right, we are interested in the distribution of all potential mean GPA we might calculate for any given sample of 50 students right. So let us look at how to draw sampling distribution. So before drawing sampling distribution let us look how to draw population distribution. So let us say there is population of size 4.

(Refer Slide Time: 04:14)



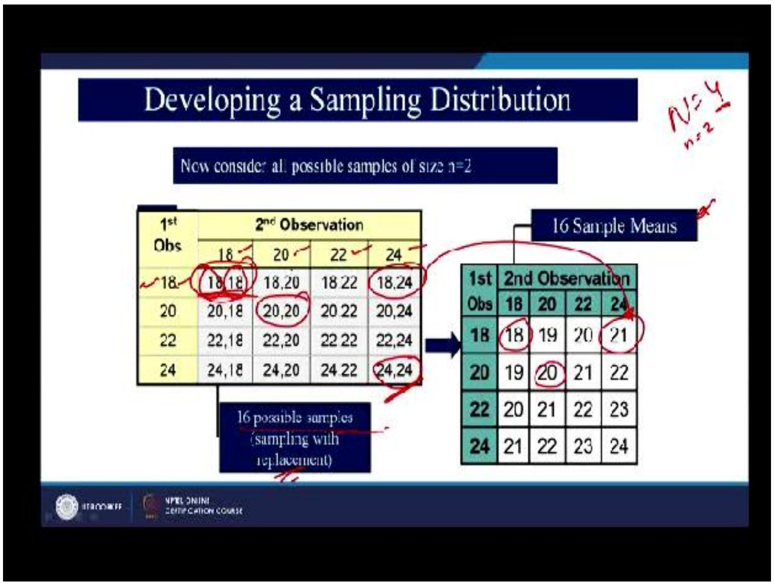
So let us say there are 4 students A, B, C, and D right and let say X is random variable and it is representing age of these individuals. So the age of let say 4 student 18, 20, 22 and 24 of year’s right. Now if you look at the mean of this population, so how would you calculate mean uses take addition of all these 4 values/4, right.

(Refer Slide Time: 04:47)



So this is mean, this μ right population mean right is called parameter ok, then standard deviation of population is this, X – μ whole square divided by n. So this 2.23. So mean is 21, standard deviation is 2.23. So you get this is population distribution. So if you take a student out of 4 students and then the probability of A getting selected is 0.25 right is not it. Similarly for B C and D remains constant right.

(Refer Slide Time: 05:55)



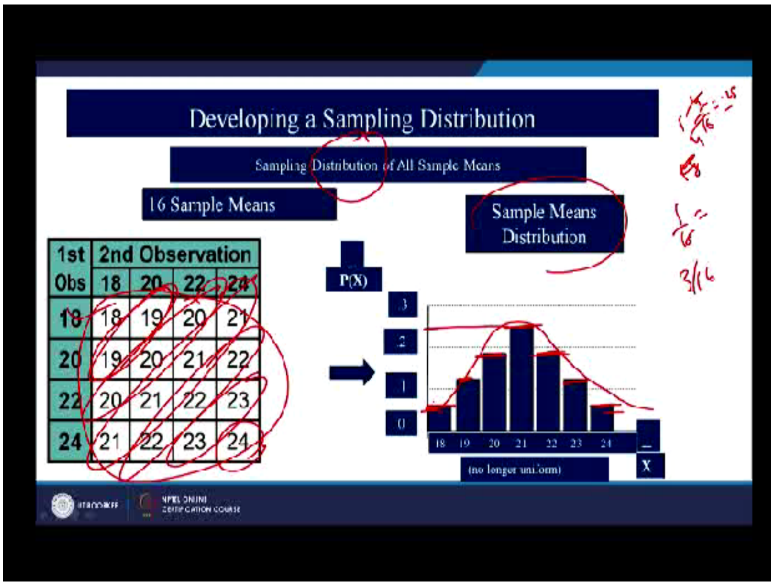
So this is nothing but a population distribution and what kind of distribution is this, this one is a uniform distribution ok, let us move on towards sampling distribution, now let us say your population is 4 and you are selecting 2 samples right, sample size is 2, so what you are doing let us say you have taken a simple 1 sample from 4 and noted down the age of that sample, let us say it is 18.

And you just again put it in the in the container or box whatever it is right, then one more sample, so again you are getting 18, so this is first time you are doing sampling right, so you have taken 2 samples and this is with replacement right. So what are the in first observation let us say you are getting 18, second observation again 18 right, so these are 2 samples. Similarly what are different possible combinations.

So you can have 16 such combinations, so first observation 18, second 18, first 18, second is 20, first 18, second is 22, first 18, second is 24 let we set 4 observations and similarly other observation, so total 16 possible samples with replacement right. And if you take mean of all these pairs then mean of first pair is 18 is not it, mean of this is 20 is not it, mean of this is you can calculate right.

So this is 21 right mean of this is 21 is not it, and so on right, so you have got 16 sample means earlier how many what was the mean earlier it was 21 right just look at this just 21 right.

(Refer Slide Time: 08:15)

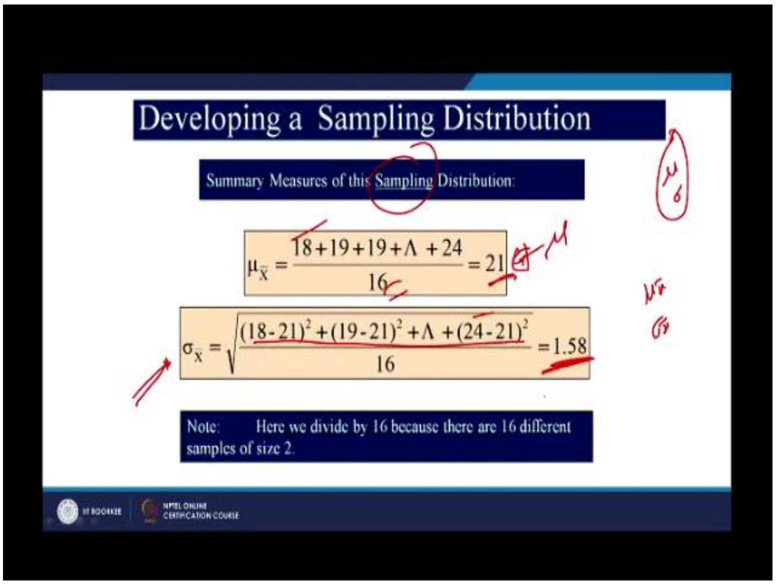


So let us draw sampling distribution, so this is sampling distribution of this particular table or these values. So how did we get this, what is this value and how to get this sample mean is 18 here, this one time 18 is appearing so 1/16, you will get this value 19, 2 times right so 2/16, this is 2/16.

So 18 and 24 just 1/16, so the for 18 and 24 it is the same value right. Similarly 19 and 23, 2 times right; so this 23, 2 times; 19, 2 times then 20 3 times, 22 3 times, so 3/16 right. So this is 20, this is 18 and 24, 19 and 23, 20 and 22, 3 times and this is just 4 times 21, so 21 by this 4/16, this is 0.25, this is 0.25, this ¼ is not it. So 0.25, so this how now if you look at same this is sample distribution.

And if you compare sample distribution with population distribution here this is not a uniform distribution is not it, it looks more like normal distribution is not it, so what is the mean and standard deviation of sampling distribution.

(Refer Slide Time: 10:19)

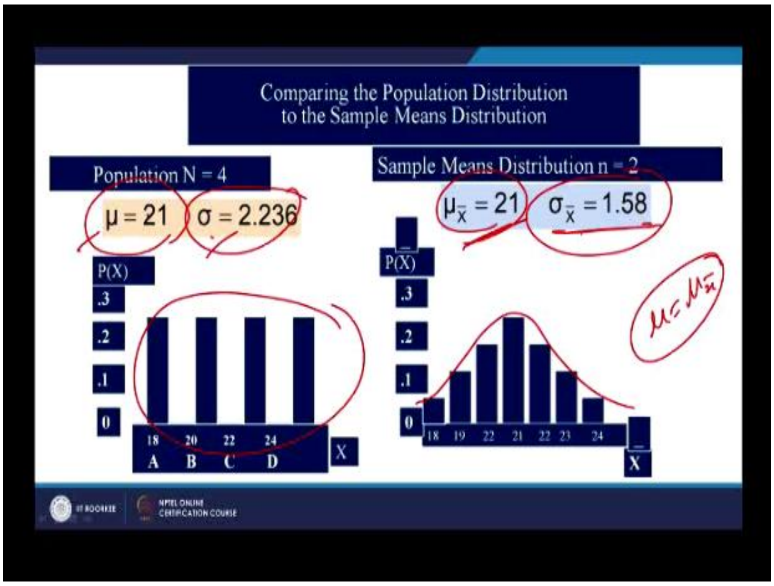


Earlier it was this was mean and standard deviation of population distribution right, here it is mu X bar and sigma x bar. So 18 you just take the average of all these number 18+19+19+20 20, 20 and so on right, all these be 16 values divided by 16, so you will get mean of the sampling distribution which is 21 and if you note down carefully then this is same value as of mu is not it. So we can say that the mean of population distribution and mean of sampling distribution are one and the same thing.

But if you look at let us say standard deviation of sampling distribution, then it is 18-21, 21 is the mean whole square, so we just keep on doing all this calculation for 16 observations right. So the last one was 24 is not it, it is 24, so 18-21 whole square, 19-21 whole square, 19-21 of square, and so on right. So finally you will get 1.58, so what was the standard deviation earlier just 2.23.

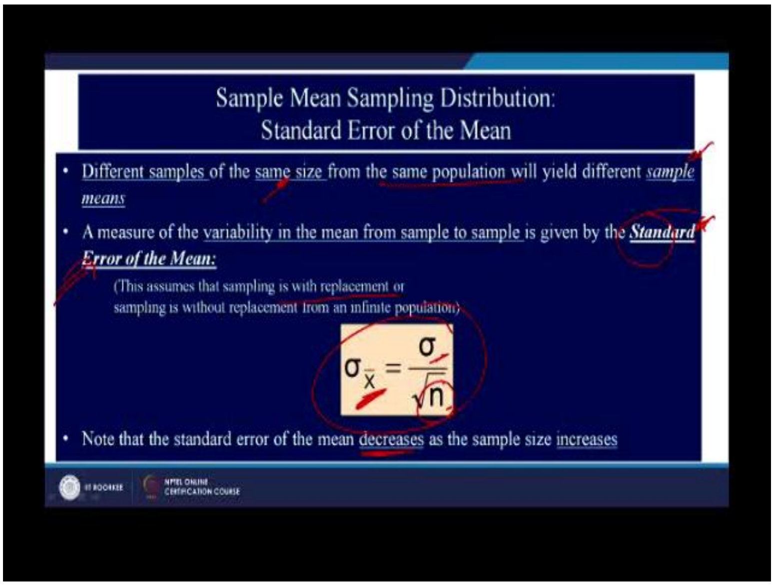
Now it is it is come down is not it, it is come down to 1.58. ok. So what is the inference will say that the weather population distribution or sampling distribution. The mean remains same but the standard deviation decreases in case of sampling distribution. So this is comparison of these 2.

(Refer Slide Time: 12:14)



This is uniform distribution, mean this, standard deviation this, this is normal distribution with mean same as of this and standard deviation is reduced to 1.58 ok. So, is there any relationship between this and this. We know that there is no relationship in fact they are equal only is not it. So we can say that μ=mix bar is not it, this is the relationship right.

(Refer Slide Time: 12:53)



But what about sigma and sigma x bar, so there is a relationship and this is the relationship and this is called standard error of the mean. So as we have said that different samples of same sample size from same population will lead would give you different sample means, so this the variability amongst these sample means would be measured by standard error of mean, it is not called standard deviation but we call it the standard error of mean.

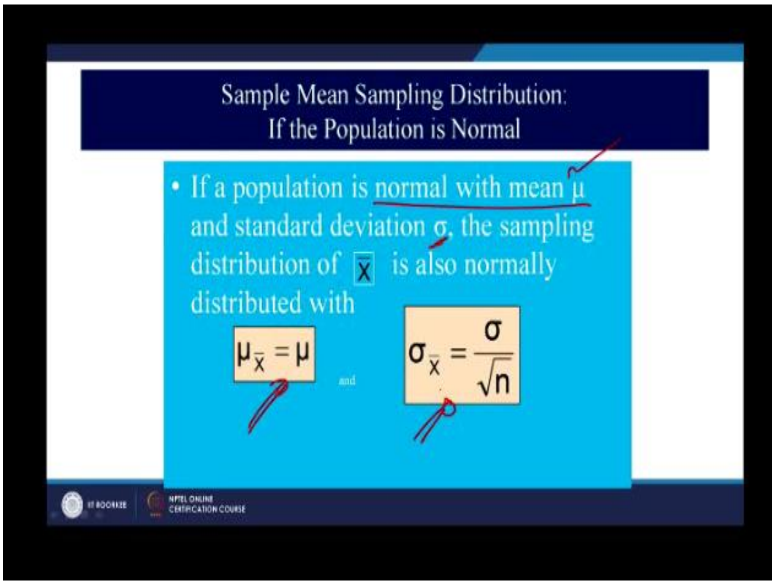
Because we are calculating how these different means are varying or what is the variability among these different sample means. So that we calculated by standard error of mean right and of course this assume that sampling is with replacement or without replacement from an infinite population. So this is a relationship,

So, this is sampling or let us call is standard error of mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

is equal to deviation divided by under root of mean.

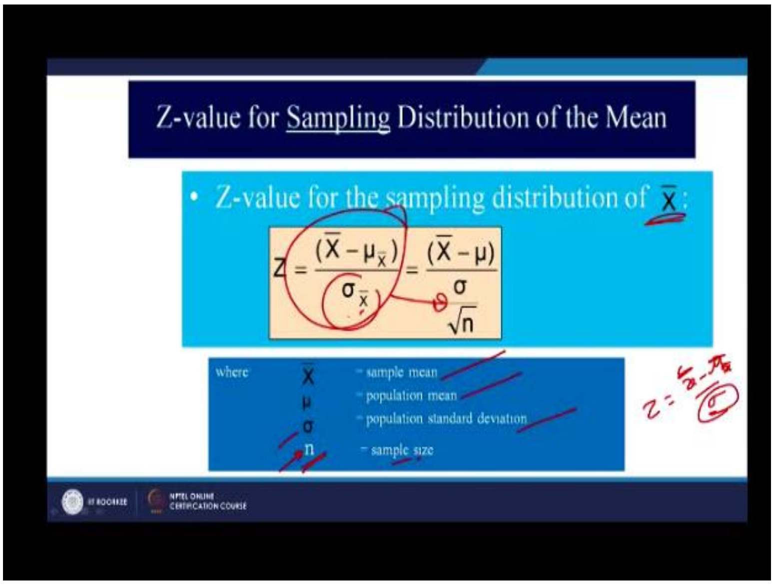
So the standard error of mean decreases as the sample size increases is not it, because this is denomination, so if you increase this value will decrease, in fact this can be proved by taking an example as where which we have taken in next couple of slides.

(Refer Slide Time: 14:35)



So if a population is normal with mean this and standard deviation this, sampling distribution of mean is also normally distributed, with mean this and standard deviation this, this what we have proved is not it.

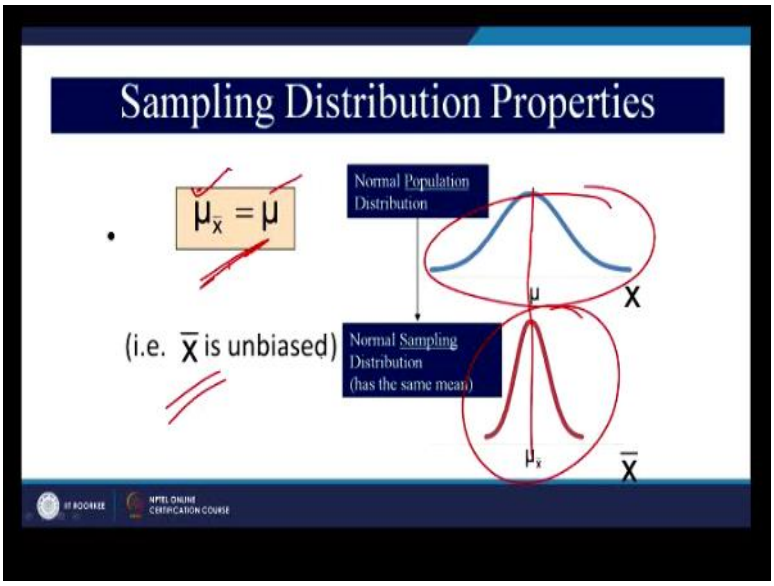
(Refer Slide Time: 14:58)



In fact we have seen ah Z value in case of probability distributions we will see the Z value for sampling distribution as well. So Z value for sampling distribution of sample mean is this, it is very similar to what we have seen, it was x-mu/sigma is not it, the same thing instead of X you get X bar instead of mean you got mean of the sampling distribution and instead of this we have got standard error of mean is not it.

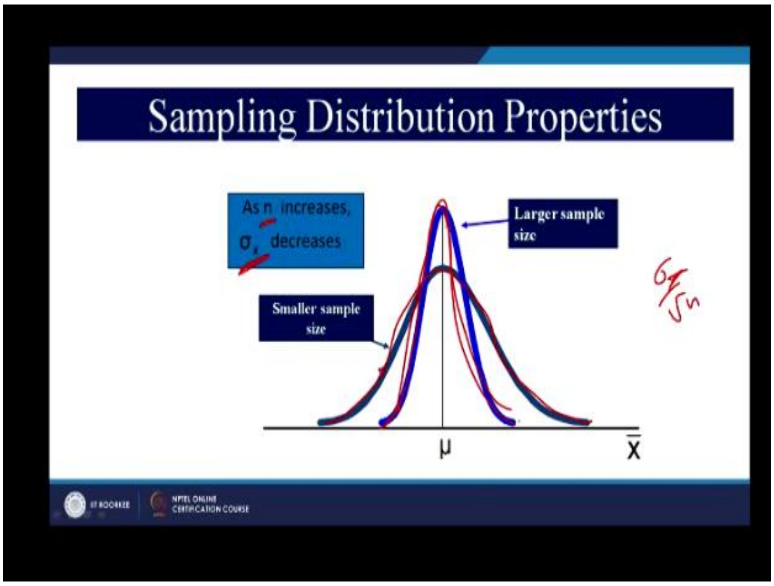
So you can put sigma X bar is equal to this here in this equation, so this sample mean population mean, population standard deviation and small in sample size. So we can calculate value of Z for given value of X bar, µ and n, because if we know sigma and n, we can easily calculate standard error of mean.

(Refer Slide Time: 16:11)



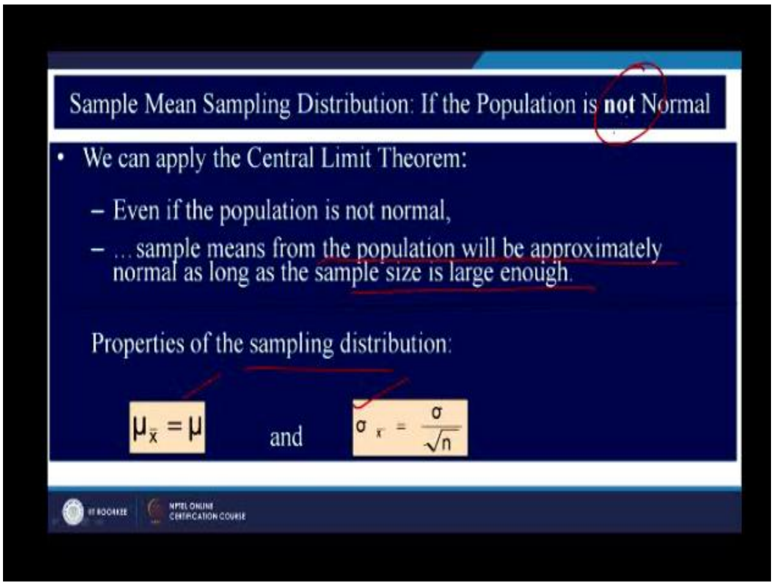
So let us look at this slide, so what is the relationship between these 2, so this normal population distribution and mean is this which is equal to the mean of sampling distribution and this is normal sampling distribution and both has got the same mean is not it, this what we have written here. So mean same in both cases is not it both the cases ok. So what we want we this mean is unbiased calculator of population mean is not it. So you can use or you can use sample mean to know the population mean or sample proportion to know the population proportion.

(Refer Slide Time: 17:11)



So the properties of sampling distribution is that as n increases a sample size increases standard error of mean decreases. So this is the standard error of mean by smaller sample size and for other sample size. So if you look at these 2 curves carefully, as you increase sample size the standard error of mean decreased is not it, because in the formula it is mention like this is not it, so smaller sample size and larger sample size.

(Refer Slide Time: 18:00)

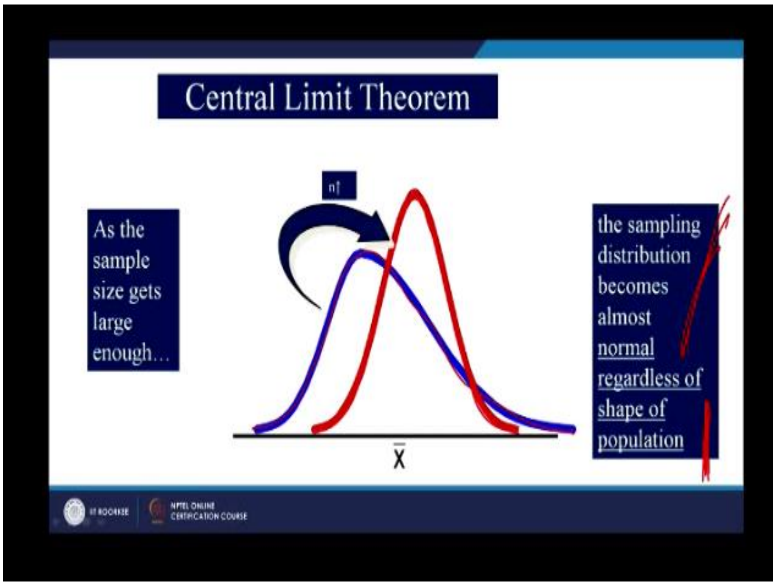


Sample mean sampling distribution if population is not normal, so for what we seen sampling distribution mean is equal to population distribution and sampling distribution standard error of mean is 1 by under root of nth time standard deviation of population. But what if the population is not normal so that relationship which we discussed holds good for normal population. But what would be the shape of sampling distribution if the population is not normal.

Then the theorem is called central limit theorem comes into picture and according to this theorem even if the population is not normal the sample mean from the population will be approximately normal as long as the sample size is large enough. So the sampling distribution of a non-normal population will also be a normal distribution if sample size is large enough and what is that large enough I will see in next slide.

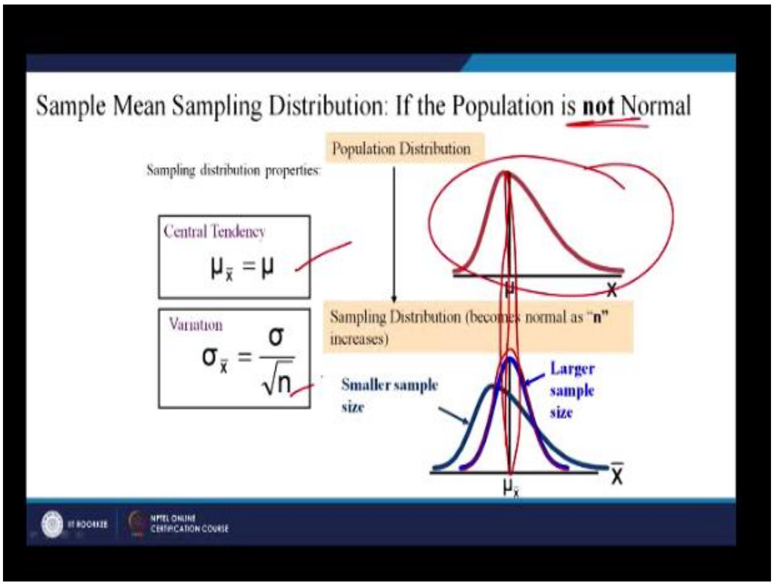
So the properties of sampling distribution will remain same, even if the population is not normal or non-normal ok. So this central limit theorem, what is central limit theorem is the sample size gets large enough the sampling distribution becomes almost normal regardless of shape of initial population.

(Refer Slide Time: 19:45)



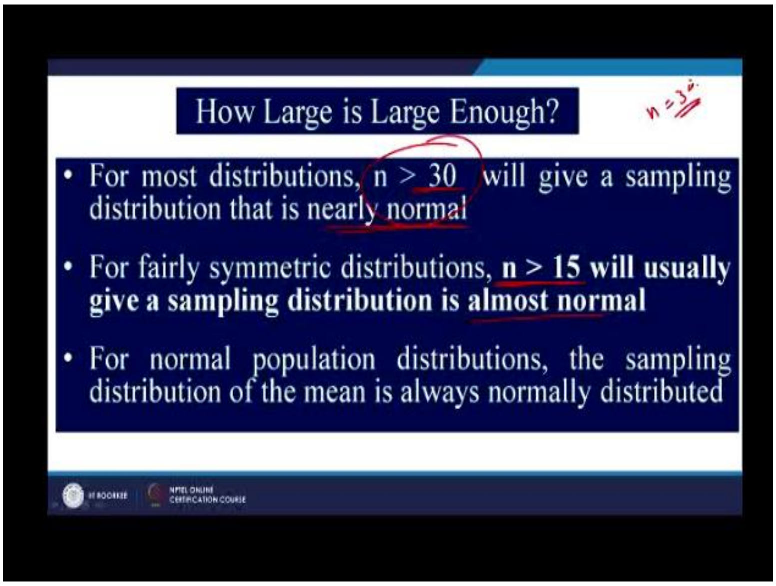
So just see this this is your shape of pollution which is non normal but the shape of sampling distribution is normal ok, because we have selected large numbers of samples from population. So this sample limit theorem quite an important and useful theorem in business applications.

(Refer Slide Time: 20:18)



This what and we have just seen this year population distribution which is not normally distributed but its sampling distribution is normal and their means are same is not it same just one state line is not it, so this central tendency and this is squares ok, so what is central limit theorem irrespective of nature of population the sampling distribution will always be normal as we increase sample size.

(Refer Slide Time: 21:00)

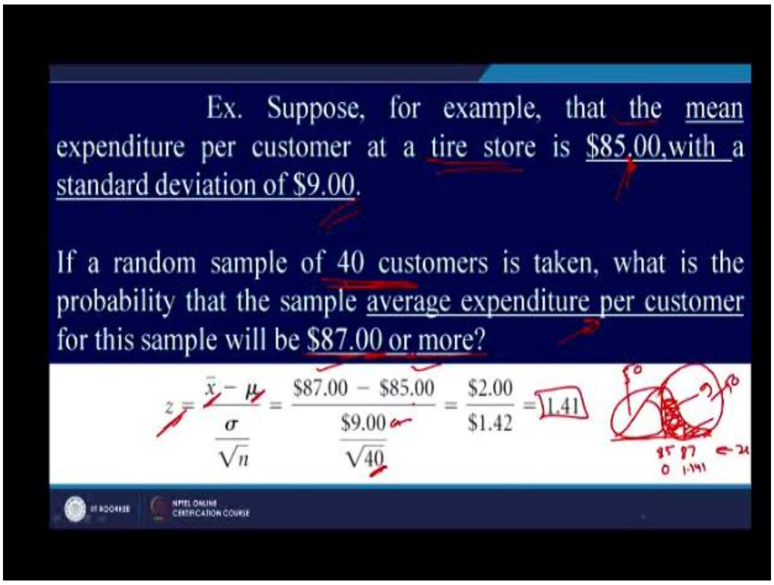


So let us look at this example, so you have a different types of population distribution. So this exponential distribution, so if you take a sample of let say 2 units then the sample distribution will be less. If you increase sample size then would be like this but if it is 30 then it becomes a normal distribution. Initial let say initial population distribution is a uniform distribution this series is uniform distribution. If you take 2 units and draw its draw sample distribution.

Then it would be like this a triangular shape this is when you increase sample size and this becomes normal is not it, when n=30, similarly similar is the case U shaped distribution and normal list of course that has to be there in case of normal this will be normal but in case of non normal distribution as well your sampling distributions are normal right. So this quiet an important slide this will give you an idea about shape of sampling distribution even if the population is not normal.

But we have to take a decision how large sample size should be, so generally it is said that for n = more than 30 or 30 will give sampling distribution nearly normal and for fairly symmetric distributions even if n = more than 15 will give you a sample distribution which is normal. But generally will take n=30 or more than 30 right. So how large is large enough it is 30 or more than 30 ok.

(Refer Slide Time: 22:59)



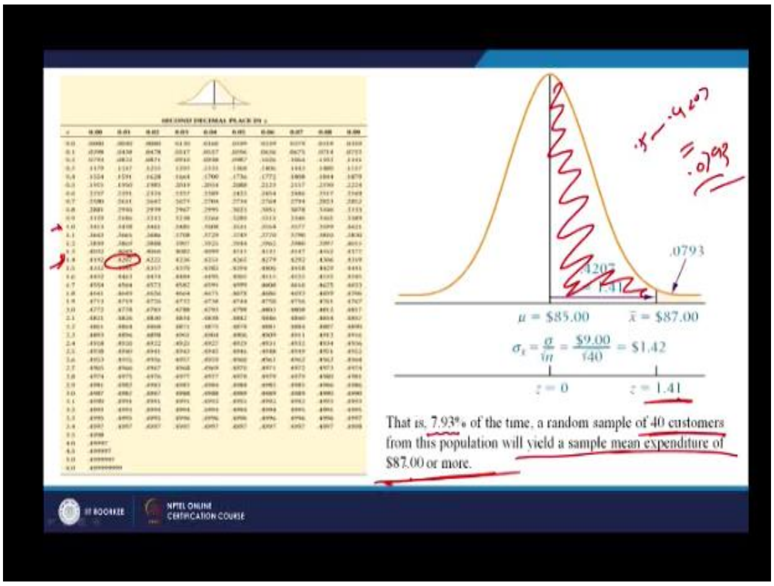
So let us look at couple of examples on sampling distribution so let say at a tire store several customers have purchased tire and the owner of the tire store selected a sample of 40 customers, now put a sample of 40 customers is selected randomly and he knows from past experience that on an average every customer spends 85 dollars to purchase tire with a standard deviation of 9 dollars.

Now if you select randomly 40 customers then what is the probability that the sample average expenditure per customer for this sample will be 87 dollars or more. So we want to know what is the probability or what percentage of time average expenditure of these randomly selected 40 customers would be 87 or more than 87. So how would you draw distribution for this right. So let us say look at this mean is 85, standard deviation is 9 right.

So first of all calculate Z value X bar-mu so this is your sample mean is not it, this your population mean which is given 85. So as X-X bar, X bar-mu this nothing but mean-X value, so 87-85/9 standard deviation is small and is 40 right, this becomes 1.41. So in fact what we want is the average expenditure per customer for the sample will be 87 or more. So 87 is towards right side or left side of this is towards right side right.

So let us write 87 here right, so if you convert this scale x scale*Z scale this is z=0 and this becomes 1.41 is not it, so what our what is our question is the sample will be 87 or more than this right, so what which area we want to calculate 87 or more right, so this area we are interested in is not it. So let us look at Z table because we know that this area is 50% is not it and we know that this area is 50% as well is not it.

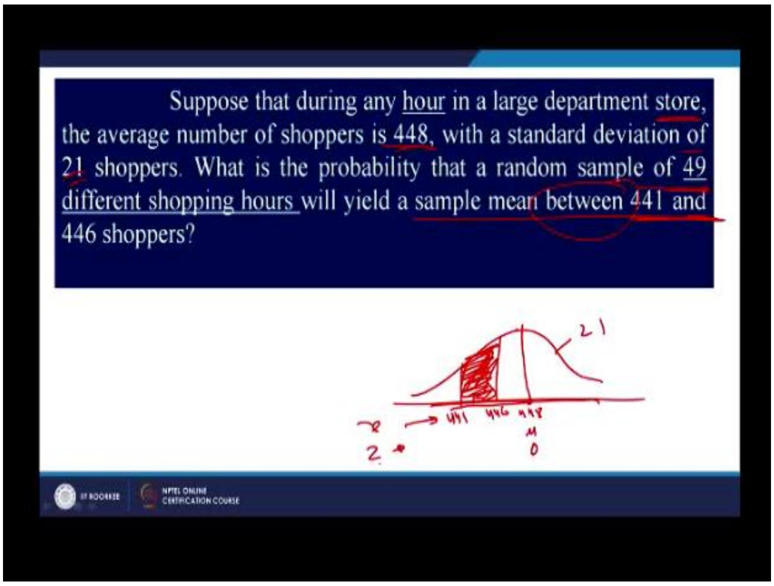
(Refer Slide Time: 26:11)



So if you know this area so we will subtract these area from 50% to get this area is not it, just look at this. So Z values 1.41 is not it, so this is 1 this 1.41 is this, 0.4207, so this area is 0.4207, we are interested in this area so what to do since we know this side area is 50%, so will subtract this area form 0.5, so 0.5-0.4207, this is equal to 0.0793 is not it. So what we will say what is our conclusion.

Approximately 8% of the time a random sample of 40 customers from this population will yield sample mean expenditure of 87 or more is not it. So whenever you take sample let us say out of 100 samples 8% of the time to take samples of 40 customers from this population then the sample mean expenditure will be 87 more than that right. This quite an interesting example and quite simple as well.

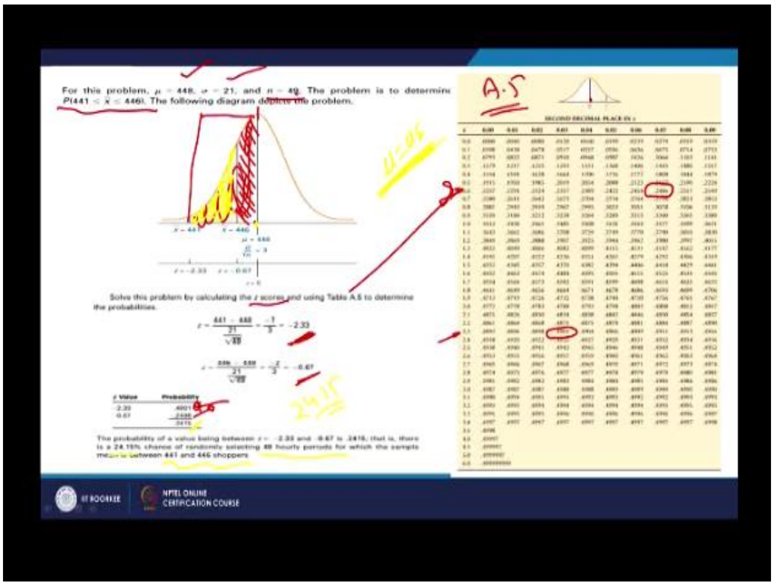
(Refer Slide Time: 27:39)



Let us look at 1 more example, suppose that during any 4 hour in a large department store the average number of shoppers is 448 with a standard deviation of 21 right, so mean is this mean is let us say 448 in standard deviation is 21, what is the probability that a random sample of 49 different shopping hours will yield sample mean between 441 and 446. So 441 and 446, so this is 441, 446, this 448 is which is mean right.

So first of all what we should do so what we have to find we have to find out in this area is not it, because what is the probability that a random sample of 49 different shopping hours will yield sample mean between right 441 and 446 is not it. So this what you want to calculate, so first of all convert all the X values*Z value is not it. So Z value here would be 0 and what would be the Z values towards left side of this and those Z value should be positive or negative.

(Refer Slide Time: 29:27)



All Z values towards right side of mean would be positive and towards left side would be negative ok. So let us calculate Z values so we know mean=448, standard deviation 21, n or simple size 49. So we want to find out p x ranging between 441 and 446. So calculate Z score using this table it is called A5 table, A.5 table, so Z=-2.33, see this is a table which gives you area under the curve from mean right and z=2 this in positive direction and Z=negative direction the area will be same is not it.

So Z=-2.33 right or 441, so we want to know area from here to here right you want this area this area under this area is not it ok. So 2.33, 2.3 and 33 is this 0.4901, so this is 0.4901 from here to here right, what about this area put it in this way right this area, now this area is again from 0 to 0.67, so 0.6 is where this 0.6 is not it and 7 is this 0.2486. So this is from here to here is 0.2486 right. Now we are interested in let me delete this all these things ok**.**

Let me use some other colour for the time being, so we are interested in this area right is not it which is between 441 and 446 we know area from mean to 441 we know idea from mean to 446. So this area will be calculated just by subtracting area from 0.490, 12.2486. So this becomes 24.15. So the probability value of being between this to this is 24.1. So we say that there is 24.15% chance of randomly selected 49 hourly periods for which the sample mean will be between 441 to 446.

So this is another example wherein we have seen how to find out probability between values of X. Now you can solve examples like this and can get the answer so let me summarise what we have done in this session. In this session we have discussed something about sampling distribution, and what is sampling distribution. Sampling distribution is a distribution of sample means drawn from a population of similar sample size and the sample mean would always be equal to population mean.

But the population distribution is let us say a uniform but its sampling distribution would be would not be uniform it would be normal as you increase sample size. So this is true in case of normal population and this is true even for non-normal population as well. So even if your population is let us say a uniformly distributed or U shaped distributed or some other population distribution which is not normal.

Then if you take samples and draw sampling distribution then that sampling distribution would also be a normal distribution and here is the role of centre limit theorem comes into picture and it says that irrespective of shape of normal population as the sample size increases the shape of sampling distribution approaches normal distribution and we have seen relationship between the mean of population distribution and sampling distribution.

And that relationship was what the relationship for was $\mu = \mu_{\bar{x}}$ is not it, so the mean of population distribution mean of sampling distribution were same. But the sampling distribution rather than calling standard deviation of sampling distribution we call it this called standard error of mean right which is this is not it. So with this let me complete this session, in next session we will work out couple of examples related to sampling distribution and will carry out couple of exercises as well, thank you very much.