Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture - 16
Confidence Intervals (Part A)

Hello, welcome back. In today's lecture, we will be focusing on confidence intervals. We have previously looked at point estimates for the population mean. So, we get a single value as an estimate for the population mean. However, this population mean is an unknown parameter and the estimate is only as the names suggests, a guess value based on the sample you have chosen. You are of course taking the sample mean and projecting it as an estimate of the population parameter.

Suppose I take another sample I will get another value of the sample mean. And since these 2 point estimates are different, we really do not know which of them is better. Both of them are taken as random samples. The elements forming the random sample are independent. However, they may give different values. They usually give different values. So, which one of them is closer to the truth, closer to the population mean. So, this is a question, which we will be addressing in today's lectures.

So, rather than giving a point estimate, it would be nice if we can give an interval estimate for the population mean. Now, we do not really have to do anything beyond the collect the random sample and find the sample mean and the sample variance. Using the sample mean and the sample variance, we can construct the confidence interval. We can construct the interval estimate for the population mean.

So, what is the term confidence or what is the term confidence interval mean? So, this is what we are going to look at today's lecture. So, we have all travelled by trains and sometimes we may have gone to remote locations and the trains are very rare or infrequent in such places and we would really like to be in the station on time. So that we can catch the train and reach our home without any delay. So, the question we may ask the local people would be what time would the train be expected.

And the person depending on his experience or knowledge may give a interval on the arrival of the train. So, the train may have been running through that place for the last 30, 40 years and so there would be a kind of population mean on the arrival of the train to the station. But, nobody has really logged in the exact arrival time in the last 40 years. So, nobody really knows the average or the mean arrival time of the train.

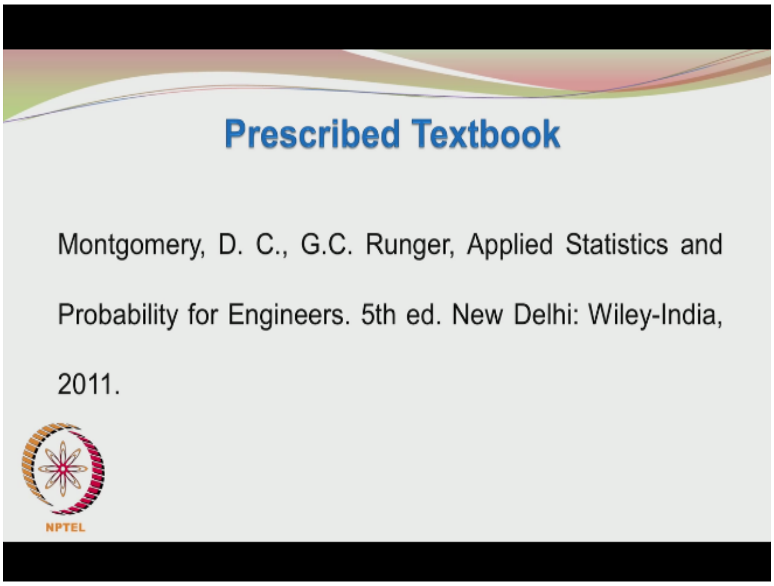
So, any person when being asked what is the arrival time of the train, he may say usually the train comes to the station let us say at 2:30 or he may say it may come between 2:20 and 2:40. Another person may say well, to be at the safe side so that you do not really miss the train, you may assume that the train comes from between 2 to 3. So, the wider the interval given, the more safer you are in actually catching the train.

So, we are thinking that the larger interval of 2 to 3 will capture the mean time of the train arrival to the station. But, nobody really wants to go to the station to much in advance, it will be very boring to wait in the station. So, we would require a precise (()) (05:39). Suppose, somebody says that the train is going to come between 02:20 and 02:40, it is acceptable. Some person may confidently proclaim that it may be coming between 02:25 and 02:35. So, this will help us to plan our journey better.

But, at the same time if the interval becomes very narrow, then there is a danger of us missing the train. For example, the person might have said, the train is going to come between 02:25 to 02:35pm. So, we may reach the station around 02:24 and we may be told that, the train has left okay. So, we have more confidence when somebody says the train is going to come between 2 and 3 but, that is a very wag estimate. So, it is not a precise estimate.

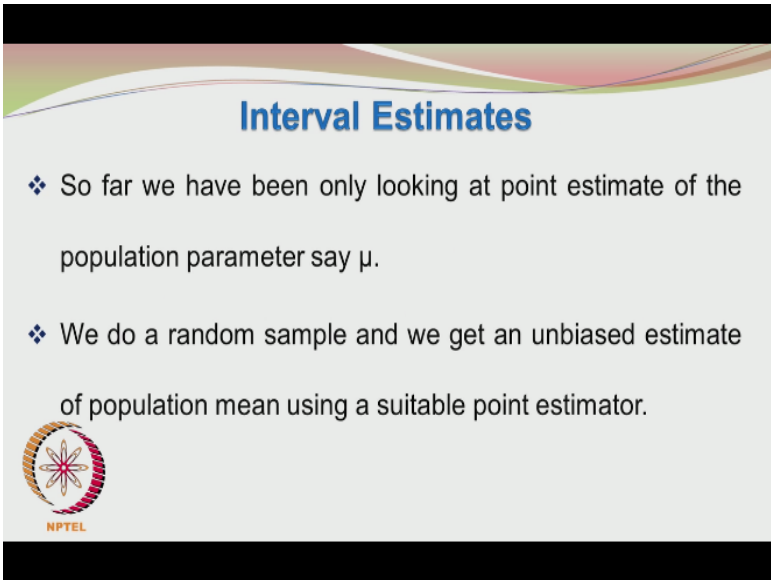
But, if you try to make the interval or the range of arrival very precise, that is also danger that the train might have left. So, how to construct the interval in which we have a high confidence and it is also precise. So, that is what we are going to look in this course.

(Refer Slide Time: 07:02)



The notation and the basic ideas are based on the prescribed text book for the course, the one written by Montgomery and Runger.

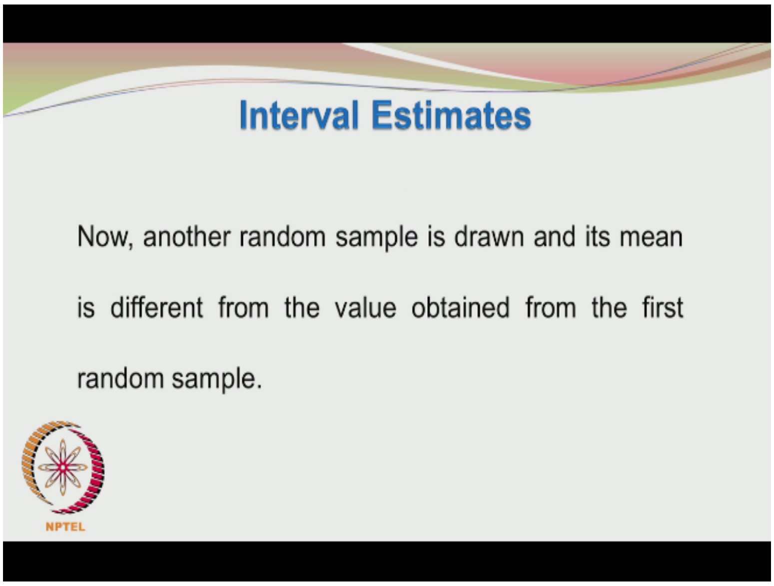
(Refer Slide Time: 07:14)



So, we were looking at only point estimates of the population parameter so far. This idea which I am going to tell you right now is not only meant for the population parameter mu but also for other parameters. But, we have to look at their probability distributions, we have to find out the probability distribution of the variance or the standard deviation. So, we will be applying the concepts of confidence interval to the population parameter mu based on the random sample mean.

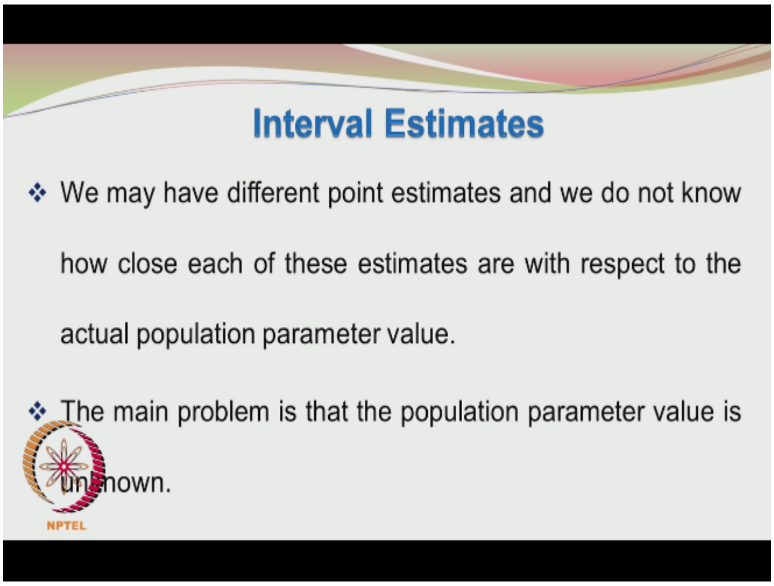
Even if we do a random sample and we get an unbiased estimate of the population mean okay, it is only an estimate.

(Refer Slide Time: 08:12)



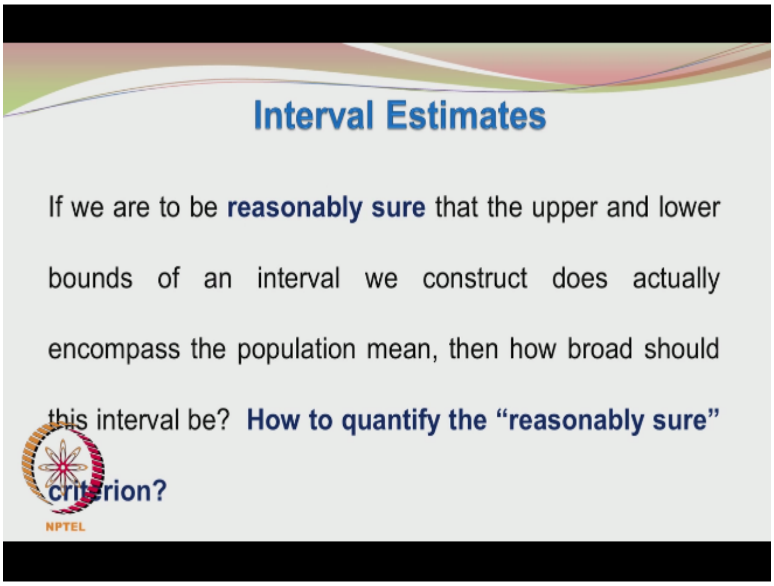
Another random sample may give a completely different estimate of the population mean, which of them is correct? In other words, which of them is more closer to the population mean mu.

(Refer Slide Time: 08:31)



Since the population parameter mu is unknown, we really do not know which of the random samples gave the sample mean that was closer to mu okay. So, we are really in the dark on which of the x bar values to believe.

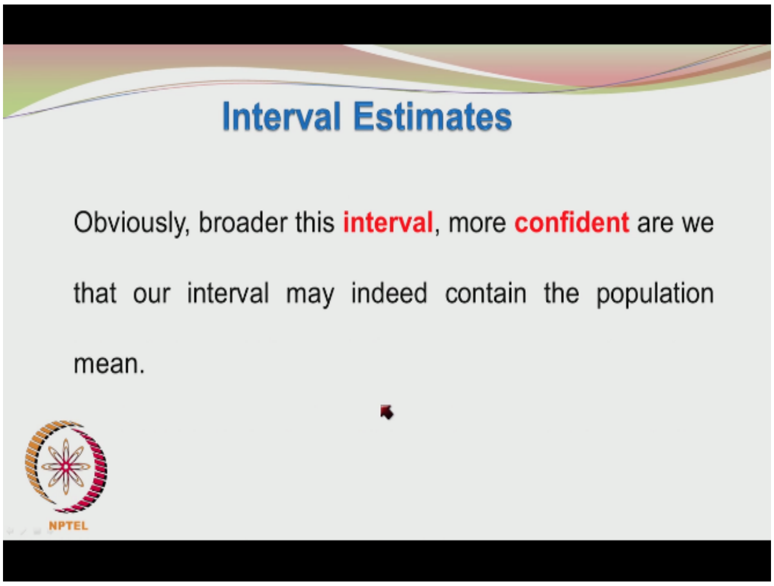
(Refer Slide Time: 08:58)



So, other than giving a point estimate, it makes more sense to give an interval estimate. It makes little practical sense to take many random samples. So, the nice thing about this concept is we will be basing or constructing a suitable interval around the population parameter mu, based on a single sample. Based on the information provided by a single sample. So, we want to be reasonably sure that the upper and lower bounds of an interval we construct does actually encompass the population mean.

So, we want to know how wide this interval must be. How to also quantify the reasonably sure criterion. One person is reasonably sure may differ from another person's reasonably sure criterion. So, we want to quantify this reasonably sure criterion.

(Refer Slide Time: 10:08)

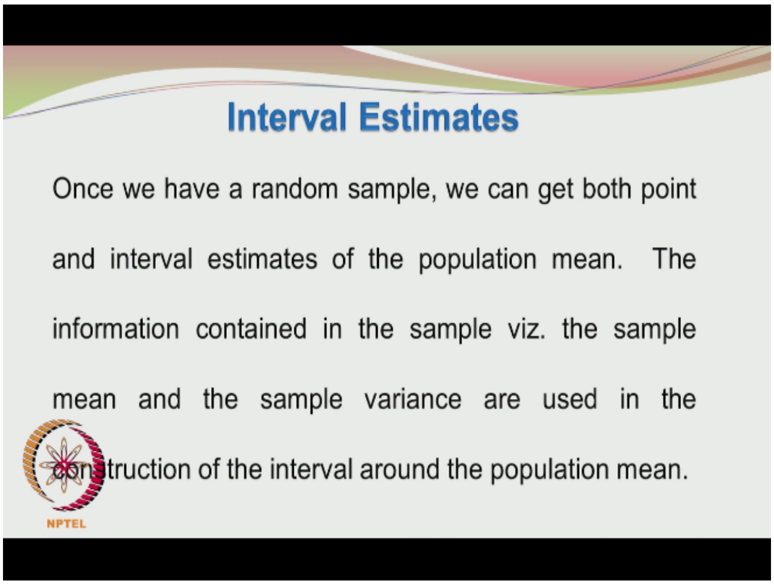


When the interval is wider, we are more confident that we have encompassed the population mean through our interval okay. So, we have a sample mean and we are constructing a confidence interval using that sample mean and if we choose to make this particular interval quite wide then, we are reasonably sure again I am using the term reasonably sure, we are reasonably sure that we have encompassed the population mean.

As I told earlier, if we say the train is going to come between 2pm and 3pm, we are making sure that the person or the passenger is not going to miss the train. So, we tell that the train is going to come between 2pm and 3pm. So, the broader the interval becomes, we become increasingly sure that this interval will encompass the population mean. However, there is no sense in making this interval very wide.

We cannot be very safe okay, we cannot say that right the train is going to come sometime tomorrow afternoon and you are better of waiting there from 12 noon. So, as the interval becomes wider and wider, the practical utility of the interval reduces.

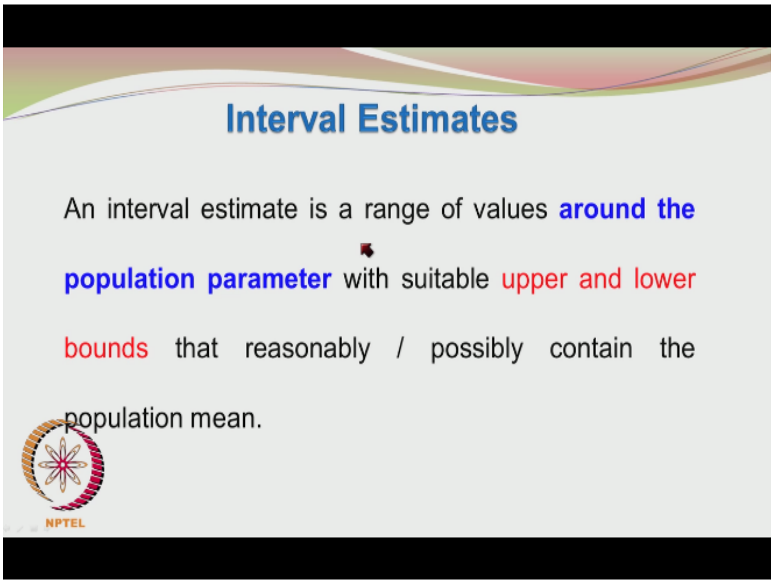
(Refer Slide Time: 11:56)



So, this is the key point. We are having only one random sample with us. But, that random sample comprises of n entities and we can use these n entities to get the sample mean and the sample variance. We may be taking the marks scored in a particular exam or the height of the entities in the sample or the weights of the people whom we have queried. So, we are going to have a collection of n attributes or data points and we can use this to find the random sample mean and the random sample variance.

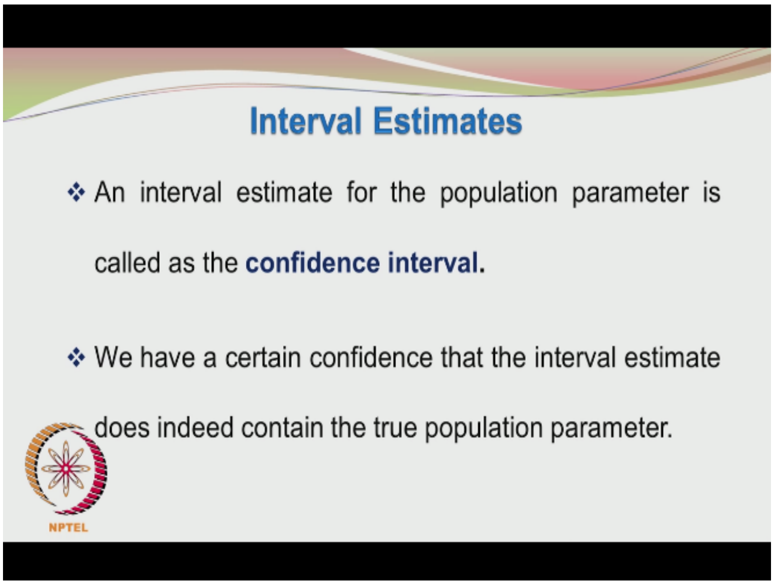
Using this information or using the random sample mean and the random sample variance, we can have both the point estimate as well as an interval estimate. We can construct the so called confidence interval using the information contained in a single sample.

(Refer Slide Time: 13:10)



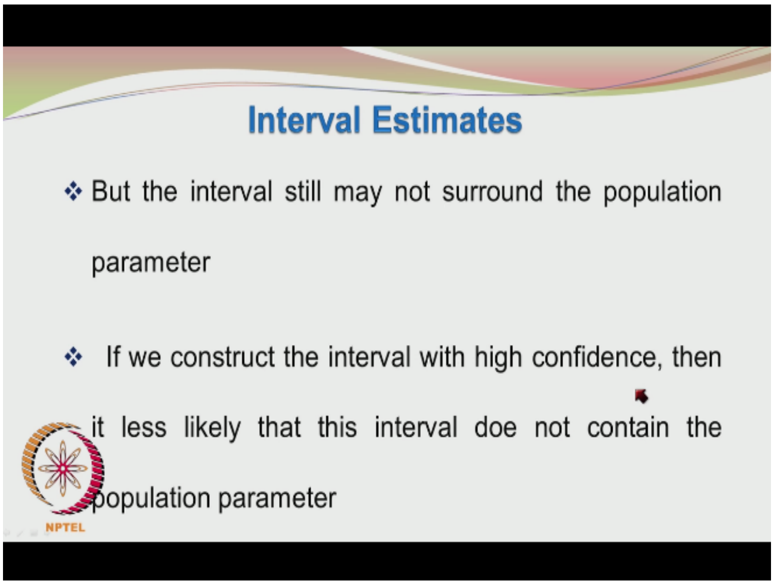
So, we define the interval estimate as a range of values around the population parameter with suitable upper and lower bounds that reasonably possibly contain the population mean. We do not claim that the interval estimate we have constructed with suitable upper and lower bounds will certainly contain the population parameter mu. No, we have not made that statement. We only say that reasonably or possibly contain the population mean.

(Refer Slide Time: 13:54)



So, an interval estimate for the population parameter is termed as the confidence interval. We develop a certain confidence that the interval estimate does indeed contain the true population parameter.

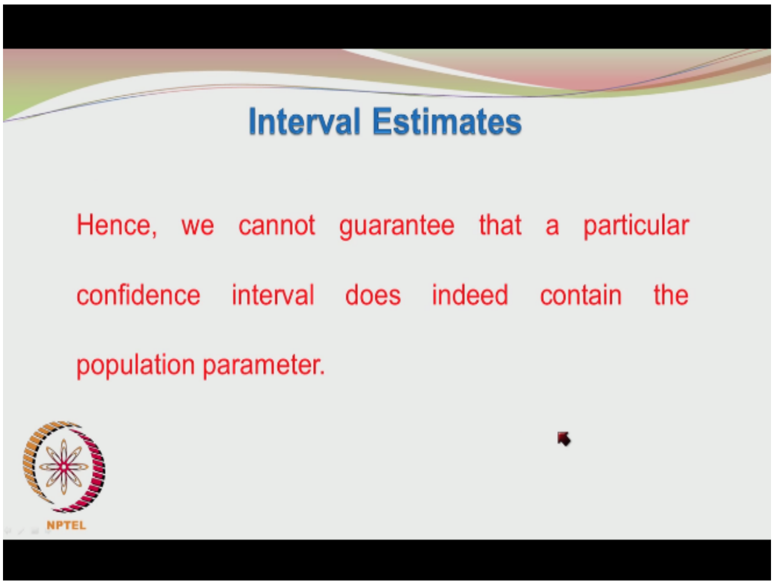
(Refer Slide Time: 14:14)



It is likely that the interval may still not surround the population parameter. The moment I start using the word likely, we are introducing the element of uncertainty or we are implying that there is a probability associated with this interval. If we construct the interval in such a manner that we can do it with high confidence then, it is less likely or the probability is small that this interval does not contain the population parameter.

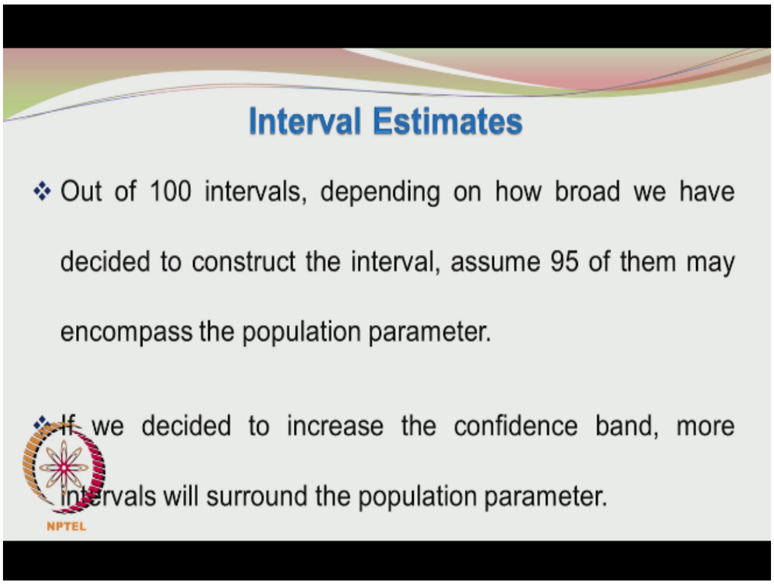
So, it is less likely that this interval does not contain the population parameter, there is a typo I will just correct it here. Right.

(Refer Slide Time: 15:12)



We cannot guarantee that a particular confidence interval does indeed contain the population parameter.

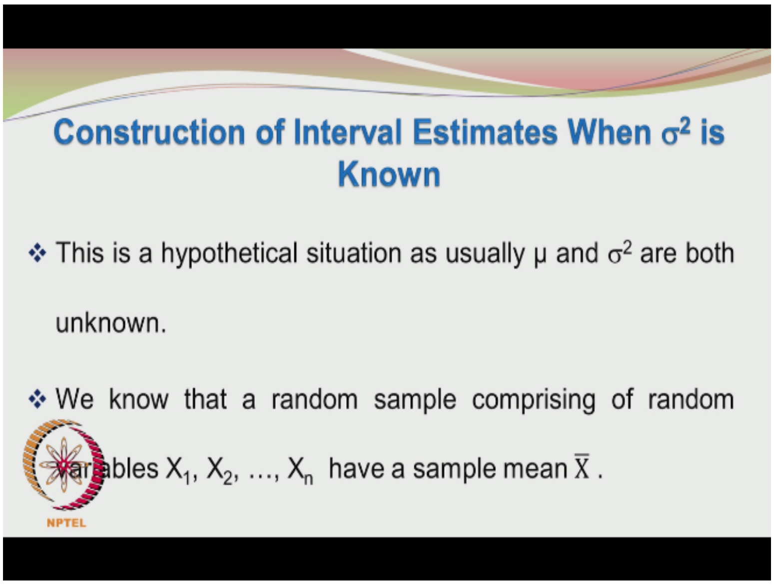
(Refer Slide Time: 15:24)



Let us take infinite intervals which is not possible but, let us assume that 100 intervals are good enough and let us say that we construct these intervals in such a manner that 95 of them may encompass the population parameter. So, this 100 may be 1000 if you want. But, to show that percentage confidence I have used 100.

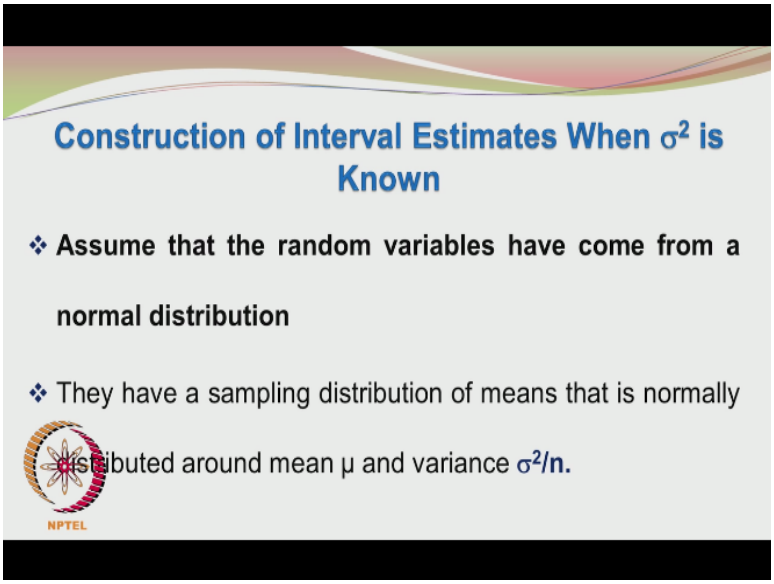
If you want to construct a large number of intervals, then, you have to take infinite intervals and then out of the number of intervals chosen, if we assume that 95% of those intervals should encompass the population parameter. 95 is a usual number we use. So, if we decide to increase the confidence band, more intervals will surround the population parameter.

(Refer Slide Time: 16:50)



So, we will be doing the discussion with the assumption that the population variance sigma squared is known. This is an assumption okay. Sigma squared is also a population parameter and just as mu, we usually do not know sigma square. But, for the purpose of discussion, let us take that sigma squared is somehow known. Later on we will see how to handle situations, when sigma squared is also not known. So, a random sample comprising of random variables X1, X2 so on to Xn has a sample mean X bar.

(Refer Slide Time: 17:50)

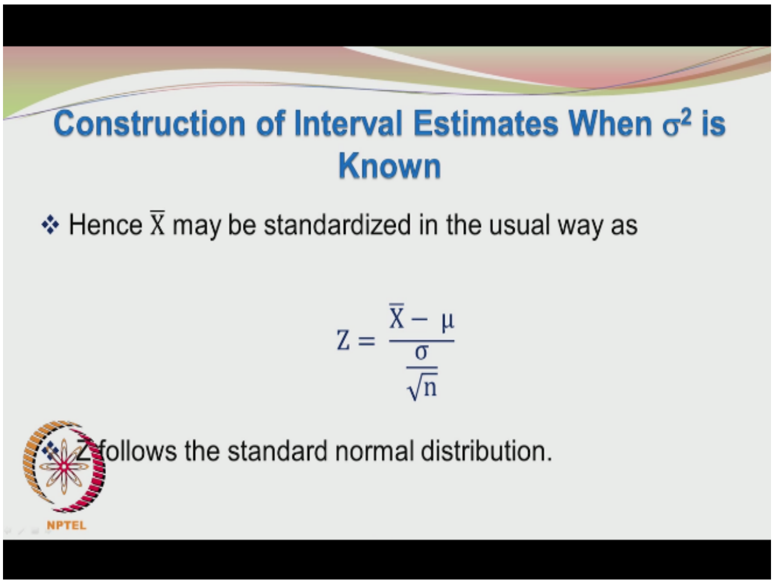


Assume that the random variables have come from a normal distribution. So, we are making 2 assumptions here. The first assumption is the sigma squared is known, the second assumption is the random variables have come from a normal distribution. We know that a linear combination of independent random variables is also a random variable. The random sample has been chosen such that the elements are independent of each other.

If they are taken from a normal distribution, then the resulting linear combination of the random variables will also result in a normal distribution. It is important to note that the random variables X1, X2 and so on are independent of one another. A sample mean is based on adding the random variable attributes and dividing by the total number. So, it is a linear combination. So, the sampling distribution of the means is also normally distributed and we know the properties of the sampling distribution of the means.

We know that the sampling distribution of the means is centered around the population parameter mu and the variance of the sampling distributional mean is sigma squared/n. Here n is the sample size. It is also an important parameter. Even though n is not present in the population probability distribution function, it plays an important role. The sample size n plays an important role in influencing the confidence level of the interval as well as the precision of the interval we are constructing. We will see more on this shortly.

(Refer Slide Time: 20:24)

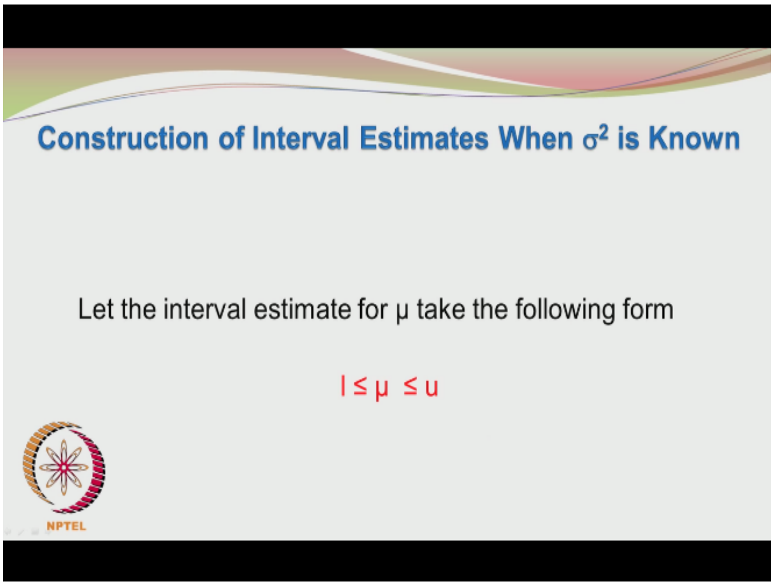


Now, X is coming from a normal population, a linear combination of all the random variables gives the sample mean and this sample mean is also having a normal distribution with mean mu and variance sigma squared/n. So, different sample means will have different X bar values. We are now interested in looking up at the probabilities. So, rather than working with different sample means, it will be helpful if we normalize them somehow.

We use the standard normal variable Z, we know that the standard normal variable Z belongs to a normal distribution of mean 0 and variance 1. So, to normalize X bar, we convert it into Z using the transformation Z=X bar - mu whole divided by sigma / root n. Sigma is square root of sigma squared, sigma is the standard deviation of the population. Since sigma squared is known, sigma is also known. This Z follows the standard normal distribution.

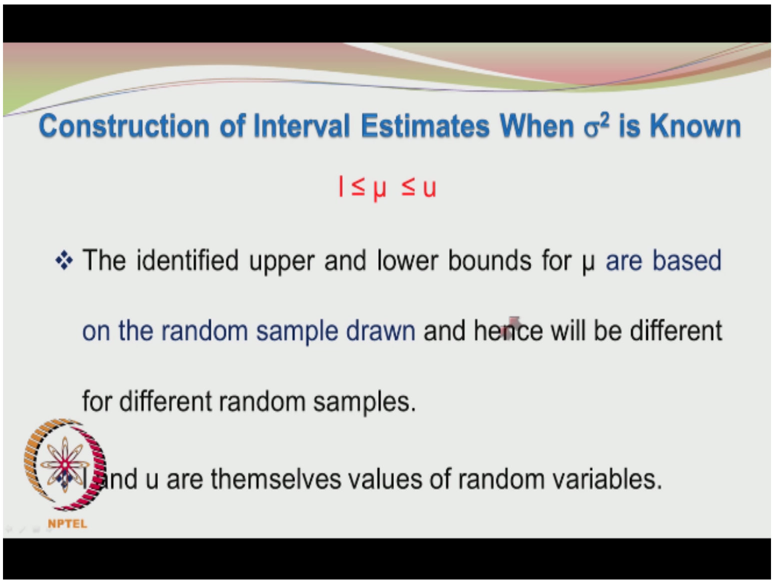
The standard normal distribution is one, which has a mean of 0 and variance of 1.

(Refer Slide Time: 22:03)



So, what is the form of the interval we want? We want a lower limit for the population parameter mu, we want a higher limit or an upper limit for the population parameter mu. So, let us express this as l<=mu<=u.

(Refer Slide Time: 22:27)

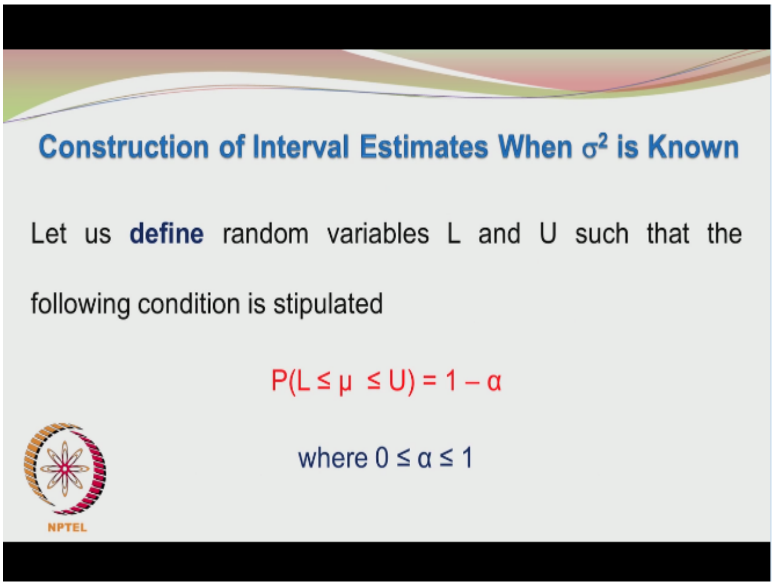


This l and u will be different for different random samples okay. So, depending upon the random sample you are drawing l and u will get identified or estimated. So, we cannot predict a priory what is the value going to be taken by l and u, it depends upon the sample which is being drawn. The random samples are based on random variables and we know that any combination of random variables is also a random variable.

Hence, the sample mean is a random variable and if we are going to construct the bounds for mu based on random samples, we are going to then construct intervals based on the random samples. So, l and u will then represent random variables corresponding to the lower limit and upper limit respectively. What I am trying to say is the intervals may take different bounds depending upon the random sample chosen. The random sample is a random variable.

So, the interval we are constructing based on the random sample is also random okay. So, different intervals may take different values. Different samples may have different sample means. Different random variables X1, X2 so on to Xn may take different values. So, what I am trying to say is the intervals we are going to construct also behave in a random fashion. And since these intervals are bounded between land u, l and u may take different values and hence l and u are themselves representatives of random variables capital L and capital U.

(Refer Slide Time: 25:09)

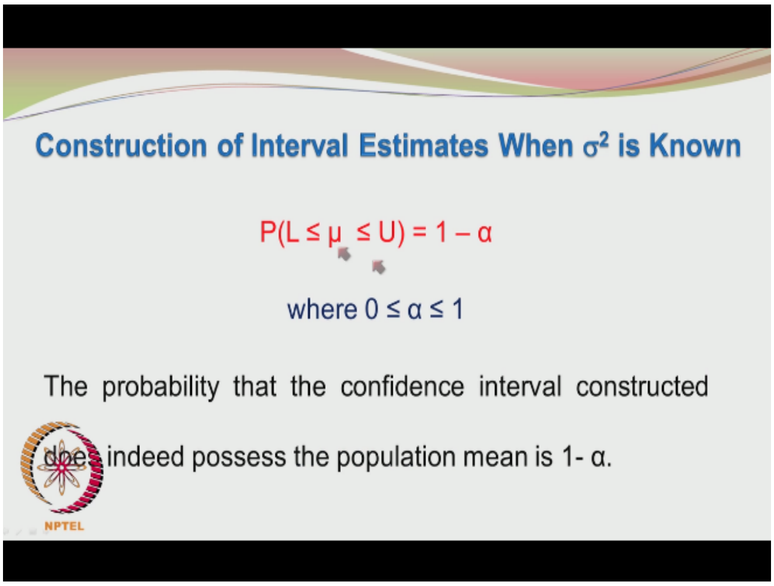


So, with this background, let us define the random variables L and U such that the following condition is obeyed. Probability of capital L<=mu<=U=1-alpha. We know that the sampling distributions of the means have a probability distribution. It has a probability distribution function. So, in this particular case, we are assuming that the sampling distribution of the means is a normal distribution okay. So, we have a normal distribution curve available with us and that gives the distribution of X bar values.

So, using that curve, we define that mu has a lower bound and an upper bound such that probability of L<=mu<=U=1-alpha. You may ask where is this mu coming from? Please remember and recollect that the sampling distribution of the means will have a mean value of mu, which is the population parameter. So, the random samples X bar are spread around the population parameter mu. So, that is why the sampling distribution of the means will have mu at the center.

And using the normal distribution curve associated with this probability distribution of the sample means, we can define probability of L<=mu<=U such that it is = 1-alpha. Alpha is a fractional value. It is bounded between 0 to 1.

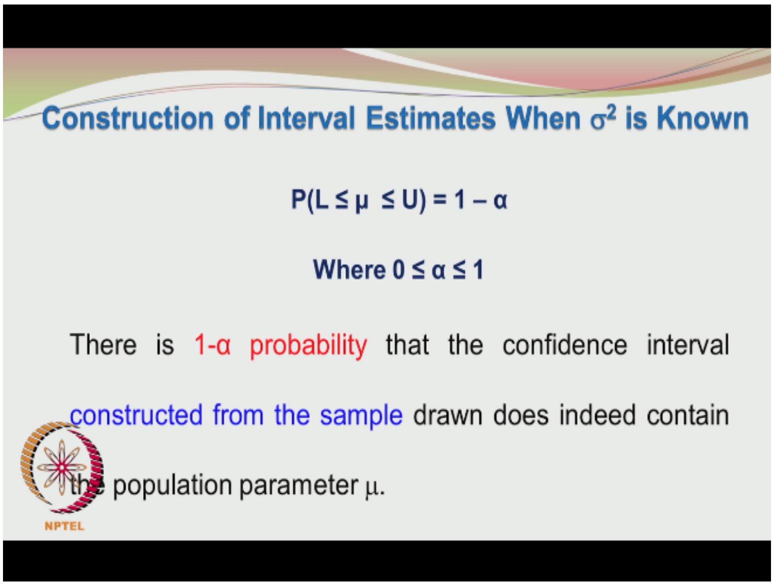
(Refer Slide Time: 27:42)



So, we are constructing the interval estimate around the population parameter mu, we assume that sigma squared is known and the sampling distribution of the means is normal. Based on this information, we define L and U such that probability of L<=mu<=U=1-alpha, where alpha is bounded between 0 and 1. This means that the confidence interval constructed does indeed possess the population mean with the probability of 1-alpha.

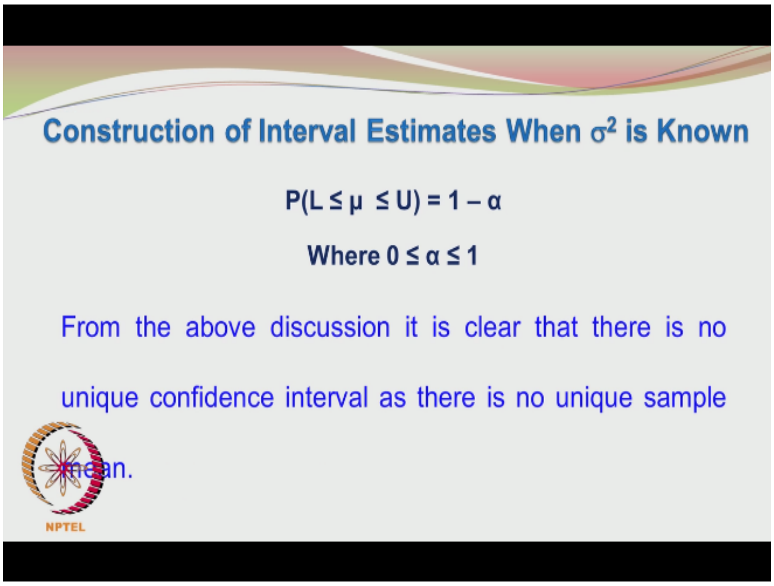
So, if alpha is 0.1, 1-alpha would be 0.9. So, the probability that the confidence interval constructed having the population mean is 0.9. We were talking about the 95% confidence intervals or confidence bands, so in order to get the 95% confidence, we have to put alpha as 0.05. So, 1-alpha will then become 0.95.

(Refer Slide Time: 29:18)



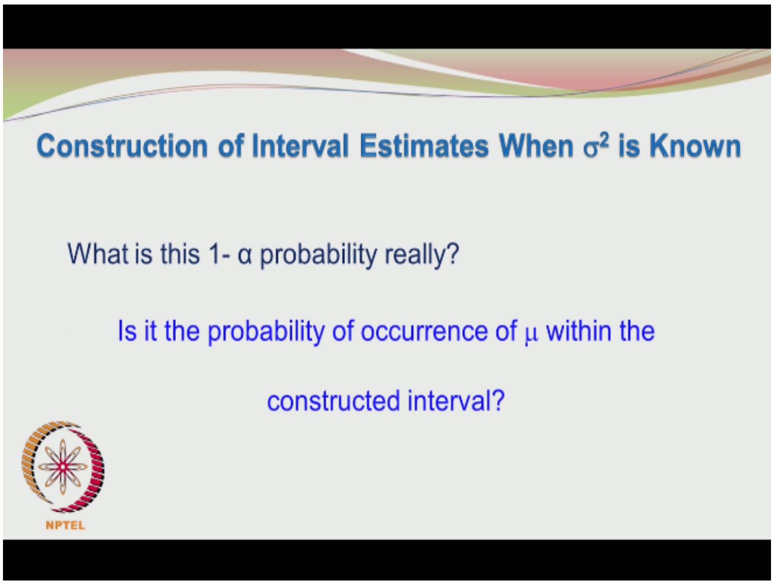
So, reiterating there is a 1-alpha probability that the confidence interval constructed from the sample drawn does indeed contain the population parameter mu.

(Refer Slide Time: 29:33)



Since there is no unique sample mean, there is no unique confidence interval.

(Refer Slide Time: 29:46)

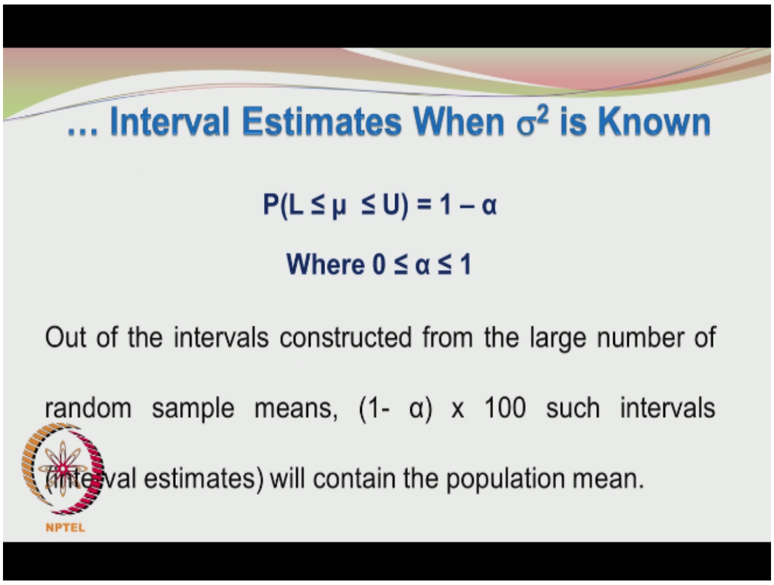


What is 1-alpha really? We just saw that it is a probability that the sample drawn and the confidence interval constructed has a 1-alpha probability of encompassing the population mean. Montgomery and Runger have an interesting discussion regarding this. They say that the probability here is more of a frequency type. So, we are now looking a bit more closely at the 1-alpha probability.

We have already defined this 1-alpha as the probability that the sample drawn and the confidence interval hence constructed will have 1-alpha probability of encompassing the population mean. Montgomery and Runger have an interesting discussion on this. They say that, once we have an interval with us, it may have the population mean mu or it may not have the population mean mu. When it is present, the population mu is present within the confidence interval, it is certainly present.

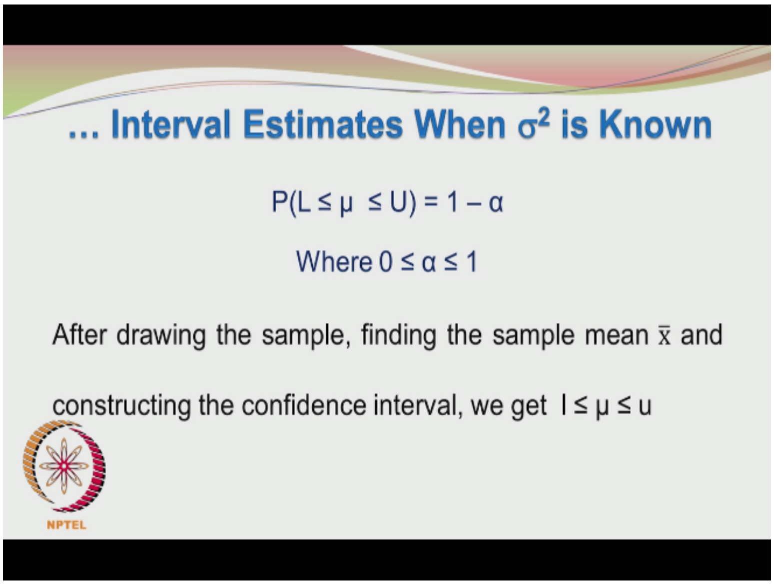
If it is not present in the confidence interval constructed, it is certainly not present or it is certainly absent. So, what is this 1-alpha probability really? So, 1-alpha is a fraction of the confidence intervals we may draw from the population, that will actually contain the population parameter mu. It is quite simple.

(Refer Slide Time: 32:05)



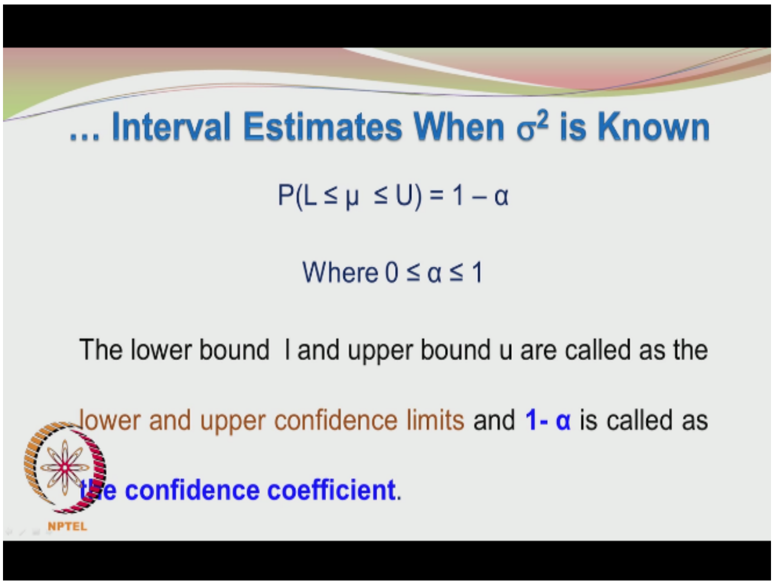
So summarizing, we have probability of L<=mu<=U=1-alpha, where 0<=alpha<=1. Out of the intervals constructed from the large number of random sample means 1-alpha*100 such interval estimates will contain the population mean.

(Refer Slide Time: 32:36)



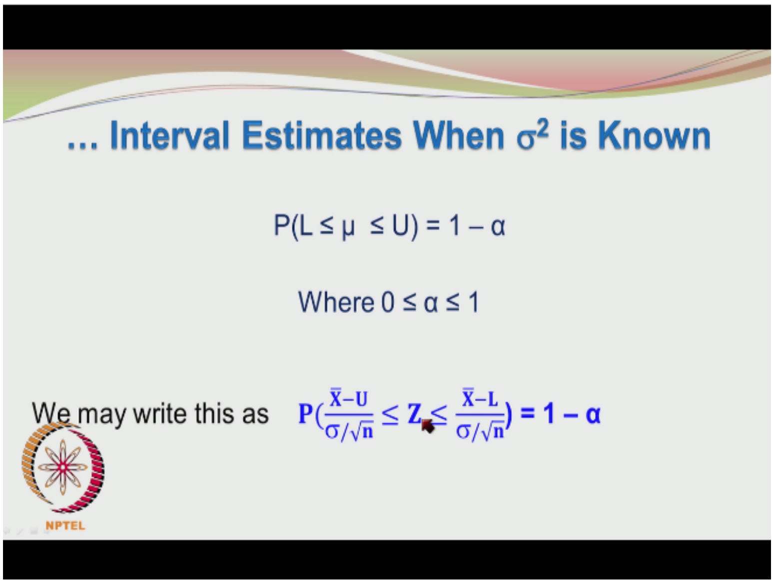
Right, so after we draw the sample, we get the sample mean small x bar and we get the sample standard deviation small s. Using this information, we can easily construct l<=mu<=u. We can identify the value of l and value of u.

(Refer Slide Time: 33:07)



The lower bound l and upper bound u are called as the lower and upper confidence limits and 1-alpha is called as the confidence coefficient. These terminologies are important because when you want to communicate your findings in papers or in conferences or even in group meetings, it is important that you use the standard terminology. So, l and u are called as the lower and upper confidence limits and 1-alpha is called as the confidence coefficient.

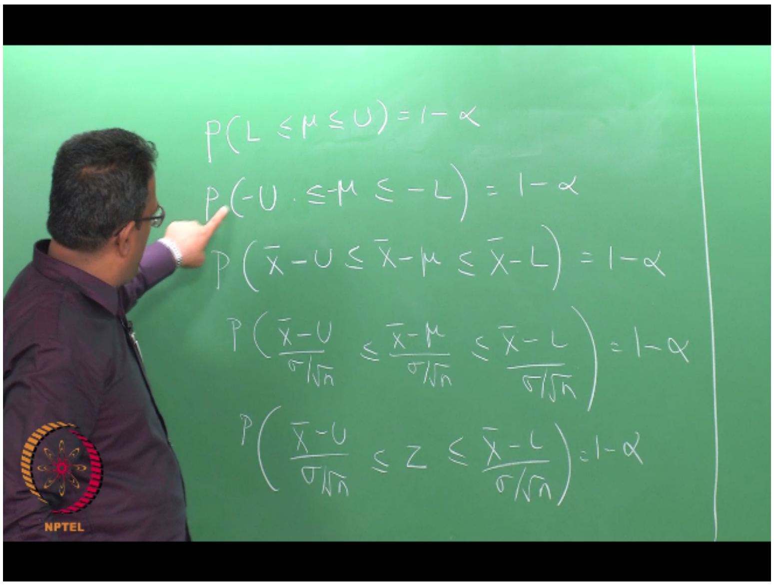
(Refer Slide Time: 33:56)



We do a bit of mathematical manipulations here. We have probability of L<=mu<=U=1-alpha, where 0<=alpha<=1. We may write this as probability of X bar-U/sigma/root n<=Z<=X bar-L/sigma/root n=1-alpha. From this step to this step, it looks a bit difficult but in reality, it is quite simple. Let us put a negative sign here and here and here. Since we are putting a negative sign, the inequality sign gets reversed.

So, we have probability of -U<=-mu<=-L. And then we add X bar so that we get X bar-U<=X bar-mu<=X bar-L. Then we divide by sigma/root n. I will demonstrate this in the board.

(Refer Slide Time: 35:43)

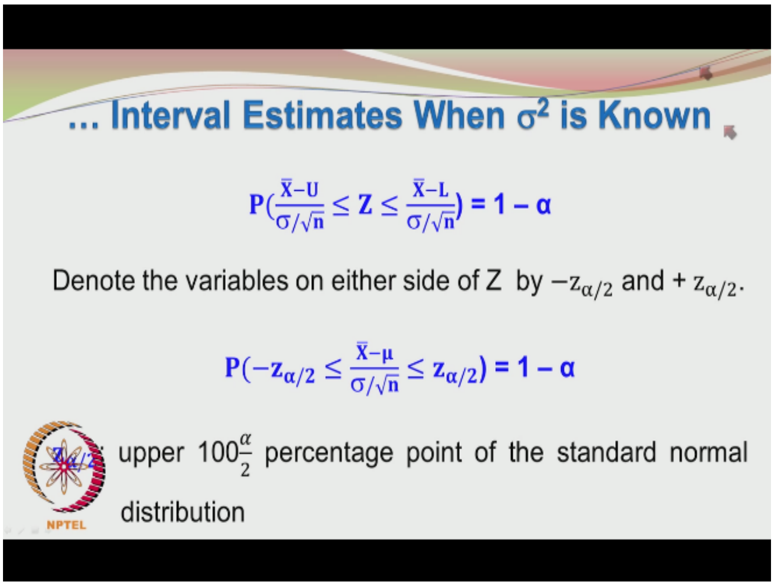


So, we have the standard definition probability of L<=mu<=U=1-alpha. Probability of -U<=-mu<=-L, the moment I put a negative sign what happens is the inequality sign reverses so, U comes here L goes here and then I am adding X bar to all the terms and so I get probability of X bar-U<=X bar-mu<=x bar-L, the probability still remains at 1-alpha, there is no change in that. And then we divide by sigma/root n and we get X bar-mu/sigma/root n<=Z<=X bar-L/sigma/root n.

The purpose of doing this is to get to the standard normal form. Here we have assumed that sigma is known and X bar follows the normal distribution. Since, different X bars will have different normal distributions, we want to normalize them in such a way that we can reduce them to the standard normal form. That is why we are having this Z term here. And since we are dealing with probabilities, we can now use the standard probability tables to get the values.

So, this is an inverse problem in the sense, 1-alpha is given to you. So, what should be the bounds for Z, what should be the lower limit, what should be the upper limit such that the probability will be =1-alpha.

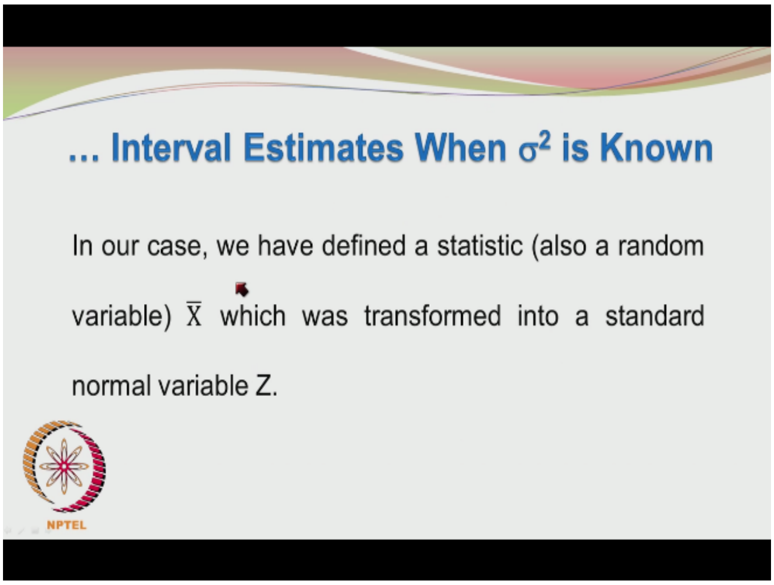
(Refer Slide Time: 39:16)



Right. So, we have probability of X bar-U/sigma/root n<=Z<=X bar-L/sigma/root n and that is given as 1-alpha. We will call the variables on the either side of Z as -Z alpha/2 and +Z alpha/2 respectively. So, this becomes -Z alpha/2 and this becomes +Z alpha/2. So, probability of -Z alpha/2<=X bar-mu/sigma/root n<=Z alpha/2=1-alpha. The terminologies are again important. Z alpha/2 represents the upper 100 alpha/2% point of the standard normal distribution.

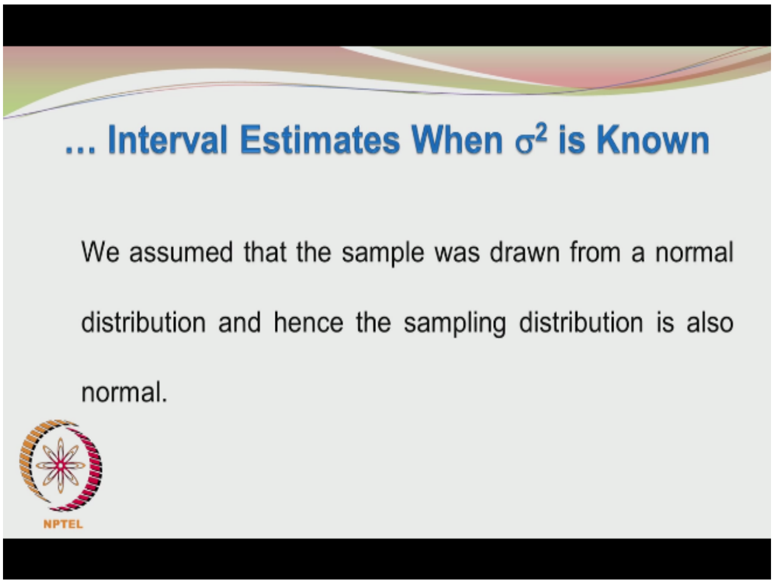
Z alpha/2 is X bar-L/sigma/root n and that value is chosen in a certain manner. How it is chosen, I will soon demonstrate. We are at present defining this group X bar-L/sigma/ root n as Z alpha/2 and we call it or term it as upper 100 alpha/2% point of the standard normal distribution.

(Refer Slide Time: 40:53)



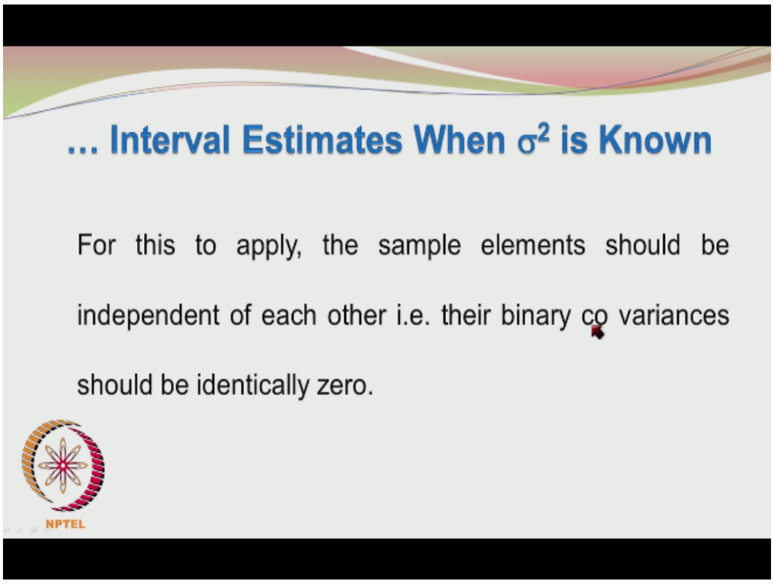
So, if you look back, we have defined a statistic also a random variable X bar, which was transformed into a standard normal variable Z according to the transformation Z=X bar-mu whole divided by sigma/ root n.

(Refer Slide Time: 41:18)



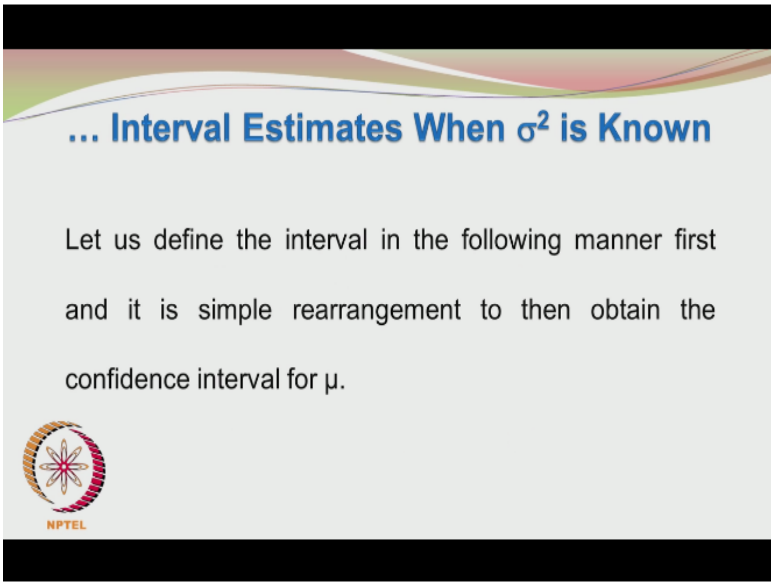
We have assumed that the sample was drawn from a normal population. The X1, X2 so on to Xn were drawn from a normal population. And hence, the sampling distribution is also normal. The samples drawn were independent of each other and hence the sampling distribution is also normal.

(Refer Slide Time: 41:56)



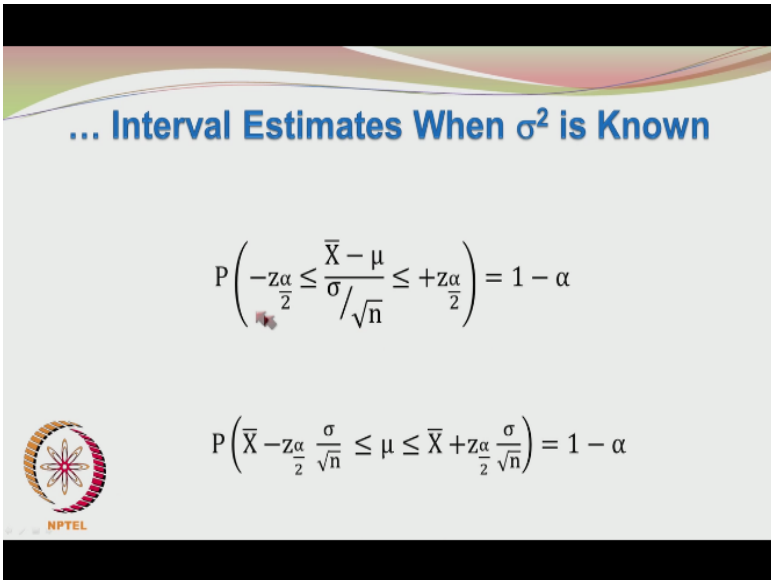
For the sampling distribution to be also normal or for the linear combination to be also normal, it is important that the sample constituents were independent such that their mutual binary co variances will vanish. So, this theoretical derivation and interpretation, we have already seen in our earlier lectures.

(Refer Slide Time: 42:40)



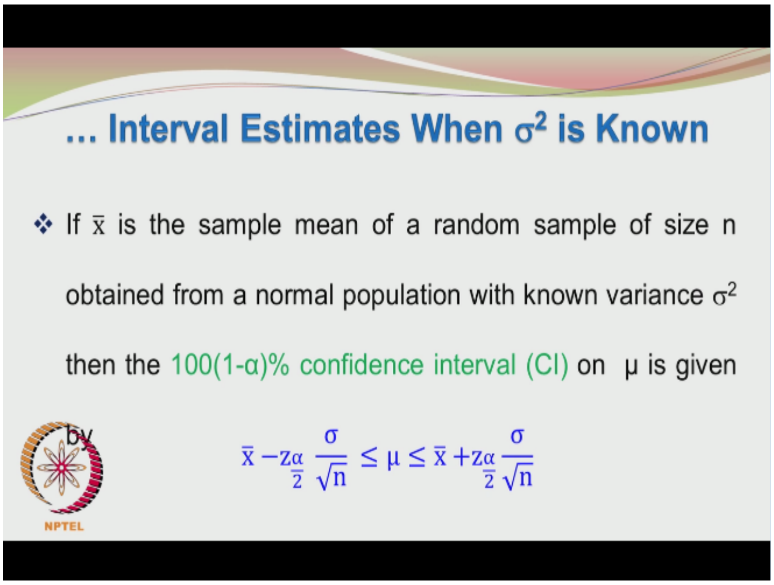
So, we have to define the interval in a suitable manner and then do some rearrangement.

(Refer Slide Time: 42:48)



Here, we have probability of -Z alpha/2<=X bar-mu/sigma/root n<=+z alpha/2=1-alpha. We can then multiply sigma by root n on all sides and then subtract X bar. A bit of simple mathematical manipulations. So, I am just taking sigma/root n here and then I will get probability of X bar-Z alpha /2 sigma/root n<=mu<=X bar +Z alpha/2 sigma/root n. So, I would request you to work out this transformation yourself and see whether you get the final following form.

(Refer Slide Time: 43:57)



So, at this point we will take a small break and once we come back we will see how to present the confidence interval in its final form. We will also see what is to be done in order to make the confidence interval we have developed also be a precise interval. So, we need to look at the issue of the confidence provided by the interval we have constructed. But also the precision of the interval as well. So, we will meet shortly.