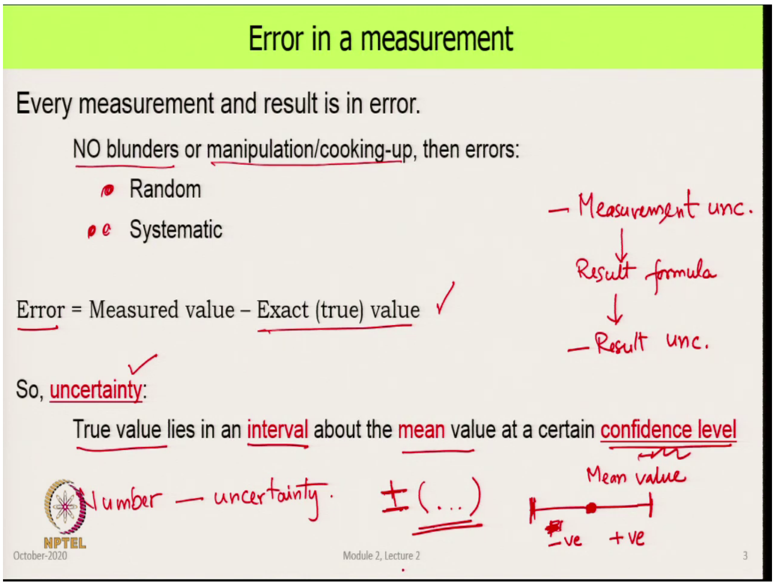
Introduction to Uncertainty Analysis and Experimentation
Prof. Sunil R. Kale
Department of Mechanical Engineering
Indian Institute of Technology, Delhi

Module - 02
Error, Uncertainty
Lecture - 05
Errors to Uncertainty via Statistics

Welcome to this course on Introduction to Uncertainty Analysis and Experimentation. We are looking at the 2nd module on Error and Uncertainty. And in today's lecture, we will see how errors in a measurement can be converted into an uncertainty on basis of statistics principles.

(Refer Slide Time: 00:42)



In the previous lecture, we had seen that every error is in measurement, in that there is no such thing like a exact measurement. And we say now that, if there are no blenders or manipulation or cooking up of data; then the errors that come into a measurement are of two types.

One is an error which is random in nature, the other is an error which is we call a systematic error. The difference has we saw last time that, random errors do not have any fixed pattern, we cannot predict them; systematic errors we know they are errors, we have some basis of putting a value on the magnitude of the error, but we still need some basis to convert that into a statistical estimator. Also we saw that, error is defined as the measured value minus the exact value or the true value.

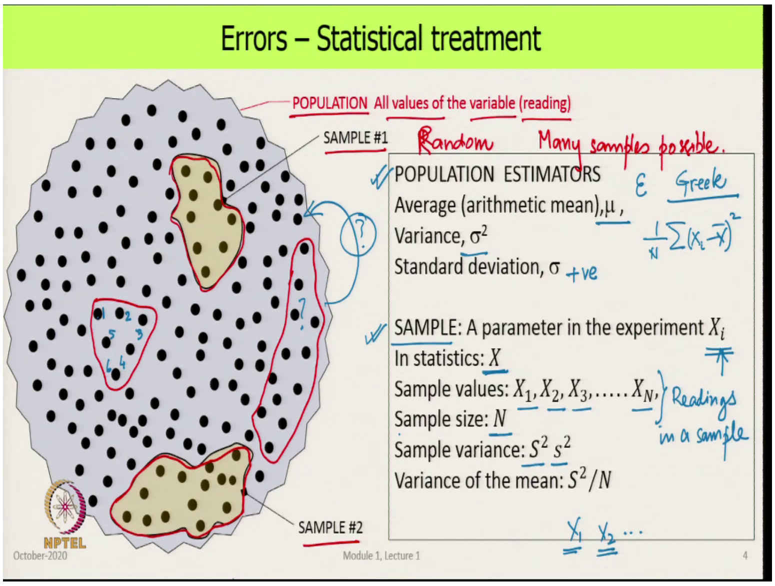
And we said that, since this is never known. So, this is never known and so we do not have an exact number to put on the error of a measurement. What we do instead is to report an uncertainty that, this is an interval about the mean value whose width of that interval is determined by the confidence level. And in this interval we say that, there is such and such a probability here that the true value will lie in this interval.

So, this is the definition of uncertainty and in this entire course and in the entire process of uncertainty analysis, we always give a quantified number on the uncertainty. Generally this uncertainty is expressed as plus minus some value, a number. In under the assumption that, both are equal in magnitude; which means that, the prediction is in the middle of the interval.

So, this is minus uncertainty, this is plus the uncertainty value; this total width, this range, this is the interval estimate and the midpoint of that is our mean value of the estimate. Our job in this course is to find out the methods and the reasons behind them to estimate the value of this.

And we do this entire exercise in two stages, first we look at the uncertainty of a measurement; then using this and the result formula, we get the uncertainty in the result. So, that is a process we will broadly follow and we will start by looking at not at this or at this; but the fact that errors in measurements and in the result, they all will follow a certain pattern and what is the scientific basis on from which we can get values of these things.

(Refer Slide Time: 04:43)



So, we go back to basic statistics and we will very quickly revise what you would learnt in a first course in statistics; that what we have first is a concept of a population. And a population is a collection of all possible values that the variable or in our case the reading that it can take; that means we have been able to measure each and every thing that is possible and all the collection of these numbers is our population.

This is a good concept, practically we cannot realize it. Simply because it takes too much time, it will take too much cost and may in most cases, it is just not possible. So, what we do? We select a small sample out of this, which is what has been shown here by this line. This is a sample number 1, we assume further that we do not have any preference by which we select the numbers that come in the sample; in that what numbers we get, how we get is completely random.

It does not matter what the distribution of the population was; the numbers we pick up are in a completely a random manner. So, like that we could have one sample, then shown here this is another sample; we call this sample number 2. And we can create as many samples as we want say, somebody say that I will take some data points there, this is sample number 3; somebody takes these data points there and this becomes number 4.

And if they do not know each other, they may even pick up each other's points; the values may come in more than one sample that is what the idea is. So, the fact is that, when you make a sample out of a population; there are many ways or many samples that are possible.

So, now what are the standard terms that are used to denote what is the population estimator and what is the sample estimator. For the population, we define the average or the arithmetic mean as mu; this is nothing but the arithmetic mean of all the values, we add all the values divided by the number of values that we have.

Then we define the variance which is sigma square, which is 1 upon N summation of all the values of X i minus X bar square. And the square root of that, the positive value of the square root of that is the standard deviation sigma. So, this something we probably learnt in school. The point to note here is that, in statistics as well as in this course we will follow a same convention that, population estimators are in Greek letters.

So, mu means something related to the population, same thing with others; if it is an error of an particular reading, we will call it epsilon not e. So, that is the population estimators. Now, we ask what does my sample give me and that is where this comes in; first we define that the parameter in both these cases, we call it X subscript i. i denotes the ith parameter, which is in our case in experiments you will have X 1, X 2 like that.

Each one of these is one individual independent measurement. In statistics, it is just called a number; for us it is more specific, it is a measurement or a reading that comes from a particular instrument. And we denote this by X subscript i. Now, there is a slight possibility of confusion, because of the symbols used; if you pick up a book in statistics or any notes, you will find that instead of X i, the book uses the symbol X, this is what is widely used in statistics.

In this course, we have reserved our i subscript for an independent parameter. So, in wherever we find that x is being used in the course, of our course; we will substitute that by X i. So, for the moment, we will just follow what statistics people have done; we go with X and we say that there are different samples that were there X 1, X 2, X 3, X n, these are the readings that this sample had. So, this is what we may call reading in a sample.

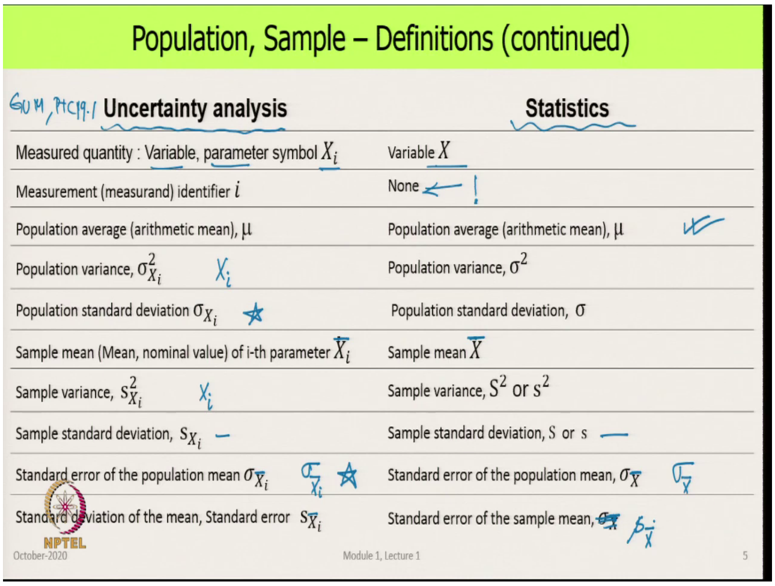
So, in this case this would be number 1, this could be number 2, this could be number 3, this is 4, this is 5, this could be 6, N is 6 in this case. This N number of samples, number of measurements or numbers in the sample, this is called the sample size. Sample variance we will come to the definition little minute; but there are two norms that are being used, some people use capital S square, some use small s square, you will find these two coming consistently in literature.

Variance of the mean is S square divided by N. So, these are some of the important estimators of the sample. We will come back to this in much more detail as the course goes on. But what we have said here is that, we have a population, all the readings are not there; so we cannot work with the population.

We pick a certain set of readings, call that a sample and start doing statistics on that. And our objective is that knowing what is there in this; can I predict what is there in the population that is our ultimate objective.

That means, if I made ten measurements of a particular value from an instrument; using those instrument those measurements, can I predict what would be the outcome if I made a million measurements? That is the question and that is where statistics comes in to help us.

(Refer Slide Time: 12:46)



So, just to clarify the point I made just a minute back; what we will use in this course in uncertainty analysis and most of the uncertainty analysis literature including ISO GUM and as well as PTC 19.1 and what would you see in text books on statistics. So, this is common and it says GUM as well as PTC 19.1; we use the symbol X i for a variable or a parameter, which means an independent measurement.

In statistics this is generally used up and given the symbol X. The measure and identifier which is the subscript is i; that means what is which is the number X 1, X 2, X 3 which is some pressure some temperature, some flow like that, in statistics this is not there. They only deal with one variable; they do not deal with multiple variables; not until you go much later in the course.

Population average, the arithmetic mean we give the symbol mu; here also the symbol is mu. So, this is consistent, population variance here is sigma square. Here we give a clarification that it is sigma square subscript i; that means it is population variance of the parameter X i.

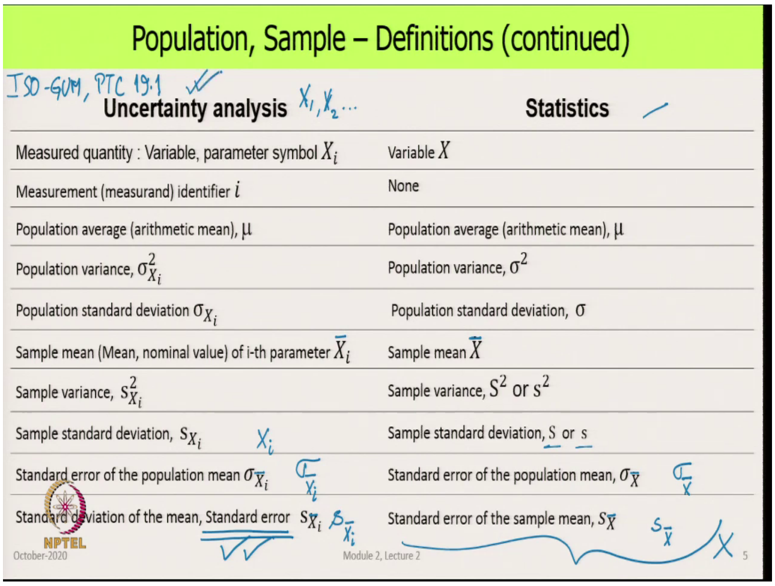
So, that is a subscript coming there. Population standard deviation similarly here it is sigma; here it is sigma with subscript X i which tells you that, this is the population standard deviation for the X i parameter. Sample mean of the ith parameter X i bar, sample mean is X bar. So, we have put the identifier i here; here there is that identifier is not there.

So, wherever there is in this case sample variance is S square or s square, there is no subscript here; in uncertainty analysis, we have sample variation variance and we specify as to which parameter or the measure and it pertains to. So, we have s X i square, same thing with sample standard deviation s with a subscript X subscript i. Standard error of the population, it is sigma X i bar; standard error of the population here is sigma X bar.

So, this is sigma X bar, whereas here X bar is not put at all; but here it is sigma X i, whereas here it is sigma X i bar. So, there is a difference between this sigma X i and sigma X i bar. This relates to the mean of the population; this is the mean of the means, the variance of the means.

Standard deviation of the mean s X i bar, here it is sigma X bar sorry this will be s X bar. In uncertainty analysis, we call the sample standard deviation s X i; whereas in the literature in statistics, you will see the symbol S or small s upper case S or small case s.

(Refer Slide Time: 16:38)



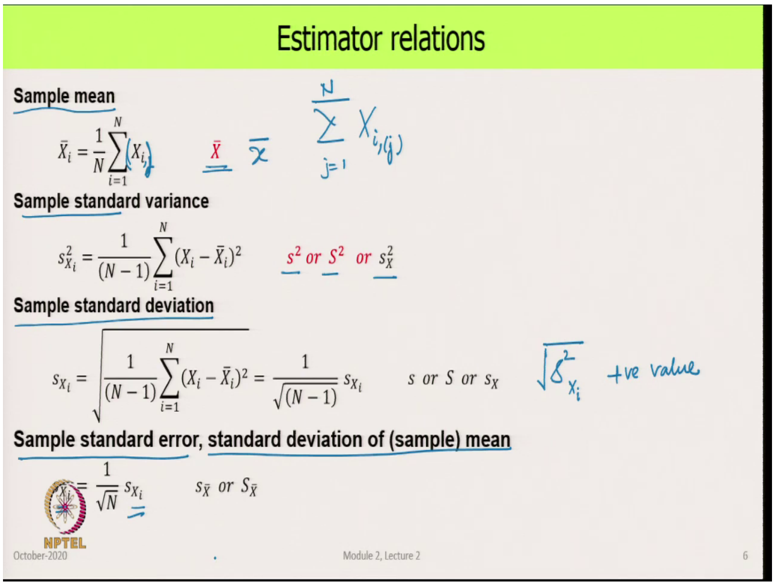
We are clarifying that the sample standard deviation is for the parameter X i. The standard error of the population mean, which is sigma X i bar; sigma is population, X i bar is the mean of the ith variable. So, this is the standard error of the population mean; whereas in literature we will just find, sigma X bar or sigma X bar like that.

And finally, standard deviation of the mean which is called the standard error; this is of considerable importance to us. This has the symbol s X i bar which is s X i bar; whereas in literature, in statistics text books, you will find s small s or capital S just X bar. So, the main thing we are doing in uncertainty analysis compare to what is there in statistics is largely that, we are clarifying that it is not just one default variable as statistics takes it.

But we have saying that we have many parameters in an experiment and each one of them is the variable as far as statistics goes and we differentiate the statistics of each variable with a symbol, so that is what we have done here. Whenever we will largely not follow this part; right now only for some explanation, I am going into some of these parts.

Subsequently and in the notes everywhere we are consistent in using the uncertainty analysis definition, which as I said is consistent with ISO-GUM and PTC 19.1. So, that is what we will do.

(Refer Slide Time: 19:10)



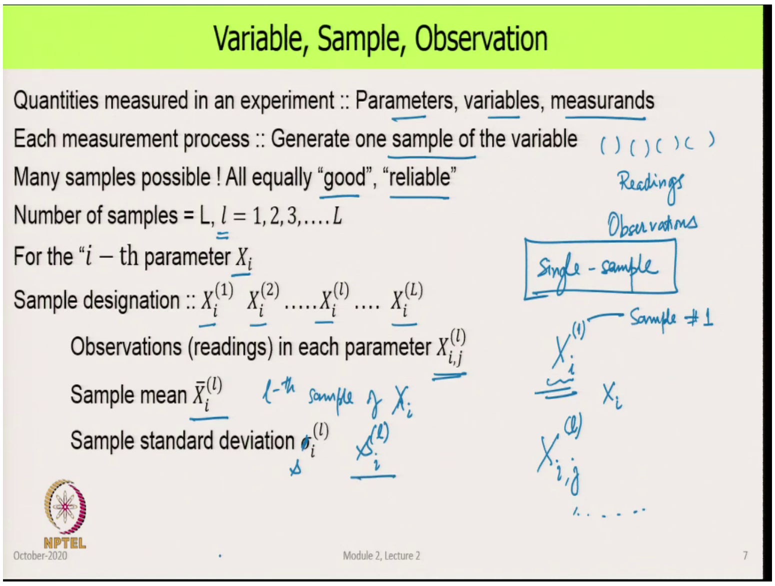
Now, here are some of the formula which are there in any course in statistics, beginning statistics may be even in school you have seen it. So, to recap that, I am just putting it up here again; the first definition is the sample mean X i bar, in statistics we use the symbol X bar or may be some people use the symbol X bar.

And this is 1 upon N summation of all the readings X i; that this is a little confusing here, because X i we have already used for a particular parameter, we could use X i with one more descriptor, we can call it say j, j going from 1 to N. So, this is X i, j; sample standard deviation s square of X i; then 1 upon N minus 1 X i minus X i bar square for all the values there.

In text books, this is s square or capital S square or in some cases s X square. The sample standard deviation X i bar is the positive square root of the sample standard variance, which is nothing but the square root of this, which is square root of S square X i the positive value.

And the sample standard error or the standard deviation of the mean s X i bar is 1 upon N s X i. So, this is the standard deviation of a sample which consists of all the sample means and this is the sample mean itself, the mean of the values of the sample. So, these are basic definitions that we have learnt elsewhere; we will use it in all the calculations that come, we will not spend any more time on this one.

(Refer Slide Time: 21:47)



So, now what is the basis of why we make an interval estimate? So, in an experiment, we have measured something a quantity, which we call a parameter or a variable or a measurand. Each time we went and make a measurement of that particular parameter, we got some readings or observations as we call them.

And they form a set; this set is called a sample. So, one such set of readings is one sample or as you will find in literature, even in uncertainty analysis, this is a single sample. And as we saw earlier, many samples are possible; we are not biased, we treat all of them as being good and reliable and there is no obvious reason to say that we should discard any one of those samples.

So, we are making that assumption. So, we said that the number of samples is L and we denote the variable as small l, which is going from 1, 2, 3 all the way to l capital L. So, for the ith parameter which is X i, we get different samples and we are using now this notation to denote that it is coming from a particular sample.

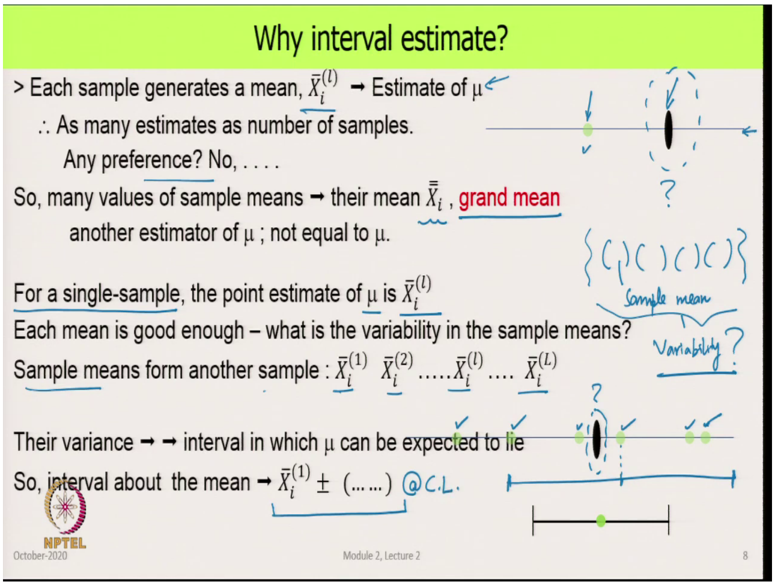
So, we have X subscript i in bracket super script 1, which tells you that this is the first sample or you can call it sample number 1 for the parameter X i. So, X i was some quantity over measuring say pressure and you took ten measurements of the pressure; those ten readings are the readings of the sample.

So, like this we have many samples coming in, L number of samples; observations in each parameter is X i, j for L. So, what it means is that, X i is the parameter we are measuring j equal to 1, 2, 3, 4, 5, 6 whatever is the every individual reading X i, j which we saw looked at in the earlier slide. And when all of these is referred to a particular sample, we put the subscript l, so that is what this is coming in.

And the mean of this particular sample is the arithmetic mean is X 1 bar with the superscript l, which tells you that this is the lth sample of the variable X i.

And using the same formula that we saw in the previous slide, we can calculate the sample standard deviation which is the sigma subscript i l which is lth sample of X i or rather this will be s i to the power in bracket l. So, sigma as we said is for a population; so we should be using small s over here. So, this is s i l, which is the sample standard deviation.

(Refer Slide Time: 25:23)



Now, why do we need an interval estimate? What we have is that, each sample is generating a mean, which we denote by this symbol. And this mean is an estimator of the population mean which is given here as mu. So, if you want to pictorially represent this, here is the line which denotes basically values; this thing here is where the true value lies which we never know let us assume it is there somewhere.

And this symbol is the number we got from the sample mean, which is X i bar l. We do not know this part here. And our objective is that we know this and we say well, what can you say about this; this is what we are going to look at in the next series of arguments.

There are as many estimates of the population mean mu as there are the number of samples, where each sample is generating 1. We have no specific preference; so there are many values of sample means and we can do two things with it, treat it as yet another parameter.

So, all these now are individually a sample mean. This forms a set; the average of these values is X i with two bars on top, this is called the grand mean. Now, if you have lot of samples, we can calculate the grand mean; but we have working in the case where is almost all uncertainty analysis, we have only one sample.

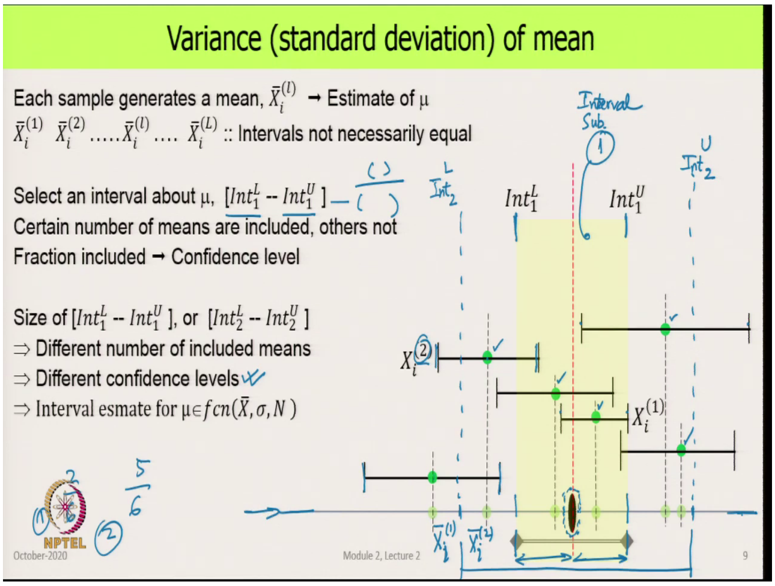
So, this could be another estimator of mu; but not equal to mu and we will not be using this, because we do not have too many samples to work with. So, we go back and say look, I am going to be working with the single sample; the point estimate of the population mean is this, each mean is good enough.

What is the variability in these this set? This is the question. That means, in a as we seen in this line here at the bottom; what we have got is that, each sample gave us a mean which is this and we have listed all of them over here. And we say that, on this number line now; this is what we were trying to predict, which we do not know and we got all these values.

Each one came from a sample and so we now know all of these values; but of course we do not know where this is and that is the objective of using multiple samples, we get this data and now we want to measure this. So, in the end what we end up doing; we say that look, about any one of these means, sample means, I will specify a range and I will tell you that at a certain confidence level, mu lies in this range.

This is what we have actually want to do; because from there we will see in uncertainty analysis, we connect that with the standard error. So, our final result will be of this form that the population mean mu lies in a certain interval and we will see this at a particular confidence level.

(Refer Slide Time: 29:28)



So, with that let us see what is happens next. We make a plot and say that this is a number line, right at the bottom here; this is what we do not know. We have just put it here; you could have put it anywhere else. And what we showed in the previous picture is that, we had these values; each one of these this represents say X 1 bar of the first X i bar of the first sample and so on.

And this represents X i bar of the second sample. And just to show clearly what the interval are, we have taken up each one of them and shown them separately. So, we take the first one and say look the value is this much; this dot shows the value and we specify a certain interval about that which is equal on both sides.

How this interval is specified, we will come to that little later; let say that we have specified an interval. We do the same thing for every sample, for the second sample which is right here, we do the same thing; this is its mean value and this is the range of the interval estimate.

Like that you do for the third, fourth, fifth and all so on; some intervals are small as you can see in this one from here to here, some intervals could be big, like you are seeing in this particular case. And so, we end up creating; if you had lot of samples, we end up creating a lot of such points.

So, then we say how can I predict where this thing lies and for that what we say is that, we will make a band which has got two limits; a lower limit and an upper limit, so this is an interval 1 interval 1 upper limit.

So, I am saying I make an interval with two limits and this is what I show that I have these two limits, at the top this is interval number 1 subscript 1, this is the lower limit of that interval, this is the upper limit of that interval and so I have got this band. And I look at this band and say, how many of the sample means lie in this range.

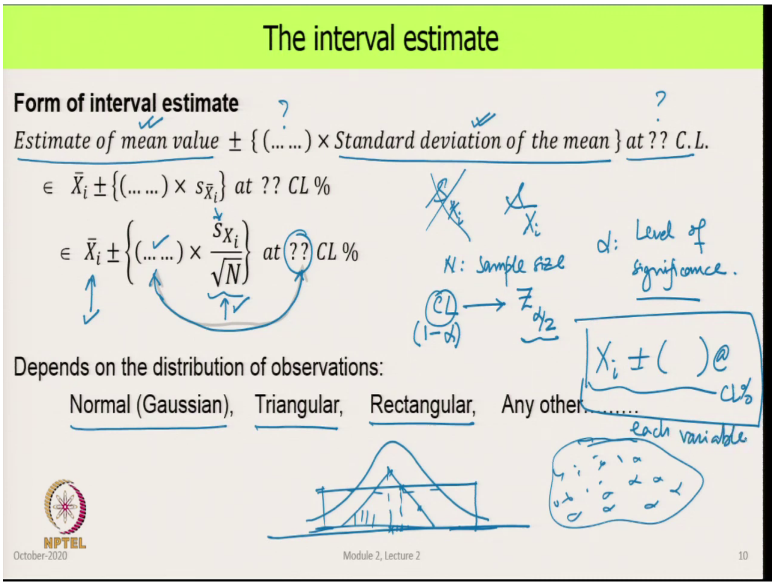
And so, in this case we see that there is one here and there is one here; these two lie in this band. And so, I will say that in this first interval, I have a certain fraction of the means out of the total number of sample means that lie in this; this tells us what is the confidence level.

So, in this case it was 2 out of 6. So, one third of the means lie within this, this is what is being told to us. Now, fixing this value equal on both sides, this is our provocative; somebody can choose a smaller value, somebody can choose a bigger value and so we could have a different interval.

So, somebody else says that, I am going to pick up an interval like this and then we draw these lines all the way up like before. So, this is our second interval Int number 2, this is Int number 2; this is the lower limit of this interval, this is the upper limit of this interval. And now what we see, in this interval, we have more of the means that lie in this; in this case it is these first two are there, then three, four and five, this one is still not there.

So, we say that in the second interval, I have 5 out of 6; in the first interval, I had 2 out of 6. This was interval 1, this is interval 2. So, depending on how wider interval we use, we will get greater and greater confidence that, this the population mean would lie in this with; because we did not know this, we have used this and said that where does this thing lie.

(Refer Slide Time: 34:37)



So, that gives us the idea of the interval estimate and the form of the interval estimate is estimate of the mean value plus minus something multiplied by the standard deviation of the mean at a particular confidence level. This is again unknown. So, from every sample or even if you have just one sample; we got one thing, which is the mean, which is the mean from the sample and we got the standard deviation that is this one.

And we say that, the population mean belongs to this plus minus this s X s X i bar; not S X i, S X i is the sample mean, this is not the case, we are looking at S X i bar which is the mean of the standard deviation of the means. And this is nothing but connected in statistics in this particular way.

And what it does is that, this multiplication factor; this is S X i upon square root N, N is the number of samples this is the sample size. And this multiplication factor which is the number, what number it takes; this depends on what confidence level we take. So, although this equation shows that there are two unknowns, there is actually only one; that is if we specify the confidence level, we automatically get this multiplication factor, which we will call say Z alpha, confidence level is 1 minus alpha, this is Z alpha or Z alpha by 2.

We will see that in little while, where alpha is level of significance. So, what it tells us in the practical case, I measured something for X i parameter; I took a some number of readings, we got the mean, so this is taken care of. We calculated the standard deviation from this, this is taken care of. Now, I will select what level of significance or what confidence level I want. And once I select that, I will get this value, which I will put over here, multiply this whole thing and I will get a number.

So, my final answer then is X i plus minus something which is at a certain confidence level. This is the interval with in which the population mean can be expected to lie; that is our ultimate objective. Now, the question is; is this limited to this readings coming from a particular distribution?

Which is that? In this population, we had lot of numbers; did these numbers have a particular distribution? This is what we are saying. In reality it could take many distributions, it could be Gaussian; that means we take the normal distribution and generate millions of points, that is a population size from which we can then randomly take a few points.

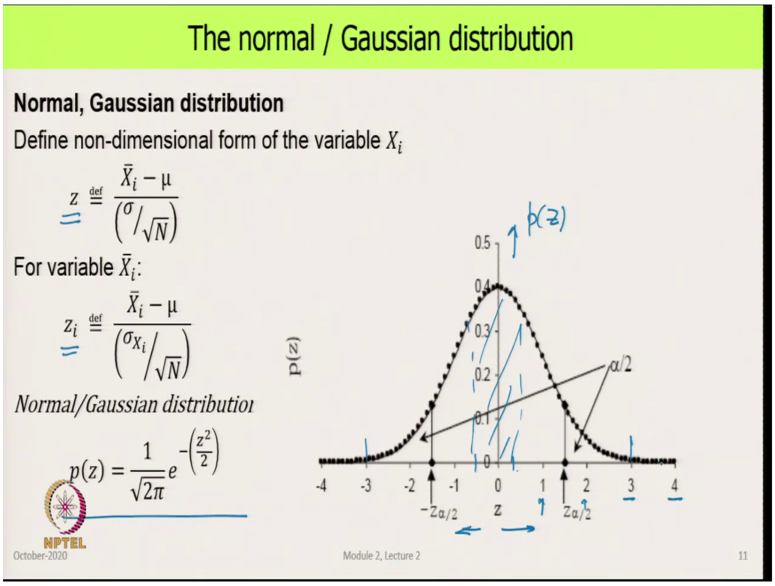
So, we will be getting our sample from a normal distribution. Or this could be a rectangular distribution, which is very easy to imagine; in that if you go to any spread sheet or program and say I want a random number, it generates a random number with equal probability between say 0 and 1.

So, there is no particular preference to any one number being more frequent in that population; that is a rectangular distribution and from that, you pick up some numbers and that is your sample from a rectangular distribution, there could be other distribution like a triangular distribution.

So, that means that, these are the numbers and the probability at that this number will come is given by this value. The Gaussian distribution would have had say value like this and a rectangular distribution would have been something like that.

These are topics for level 2 course, we will not going to that details of this at this point. But what we have got is that, here is a method by which I can get the interval estimate that I am looking for and this I do for each variable in my experiment.

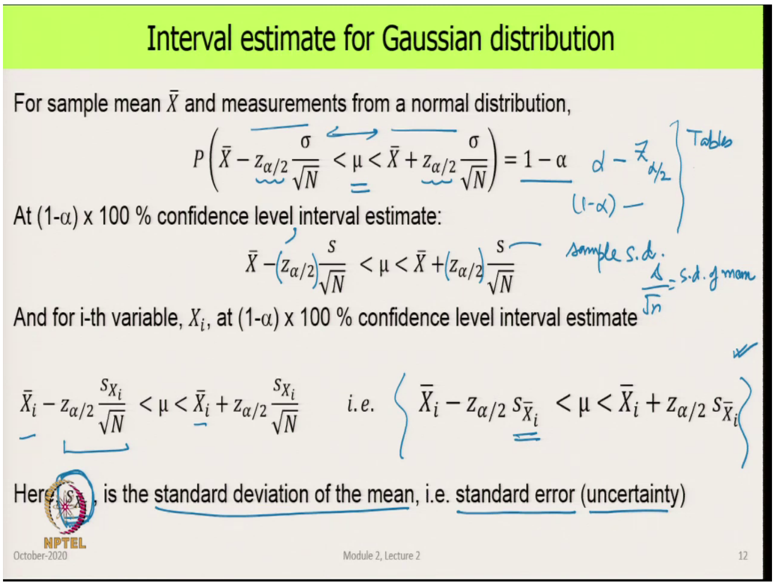
(Refer Slide Time: 39:30)



And just to show quick revision of what is normal distribution or a Gaussian distribution; we can look up any book on statistics, we define a non-dimensional form of the variable in this form. And for the variable X i bar, we have Z i is equal to this and the normal distribution is given by this relation. And if you plot this, this side is Z on the x axis; this side is the probability of that Z and we get a symmetric distribution which is like this with Z values being 1, 2, 3, 4 like that.

So, if we pick up an interval there, the area that is covered under this; this tells you what fraction of the readings lie in this band. If you take another thing say from 3 minus 3 to 3 and ask how many readings live in this; then you will get another number like that, bigger the interval, more the readings get into this. So, that is the Gaussian distribution.

(Refer Slide Time: 40:44)



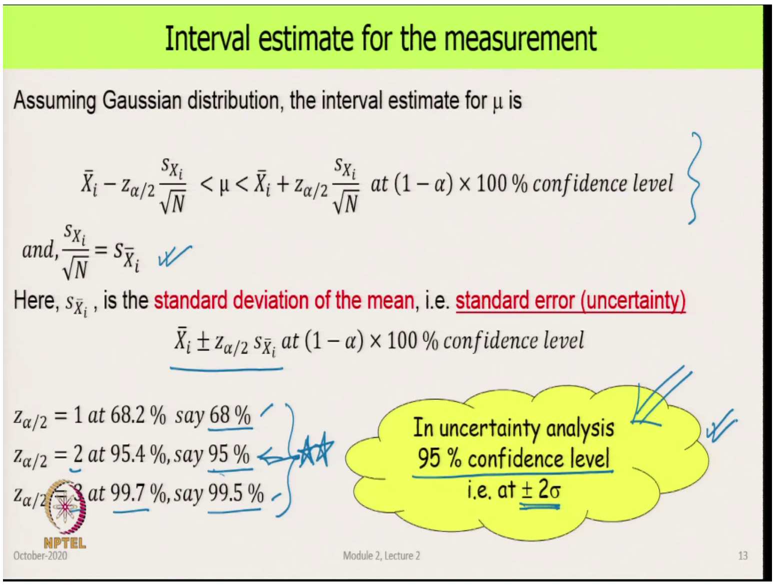
Now, we say that, let assume for the time being that we have a Gaussian distribution. Can be get a more specific formula? In that earlier formula, can I know what that multiplication factor was? And that is what we have done here; we say that the probability that mu will lie in this interval from here to here is 1 minus alpha.

So, alpha is the level of significance and we got something coming here, which is z alpha by 2, which is nothing but the value from the normal distribution for that particular value of alpha. We do not need to calculate it; there are tables, there are charts, there are programs, there are automatic things in all the spread sheets. So, we do not need to worry about it.

So, we can get this correlation between alpha and z by 2 or 1 minus alpha and z alpha by 2 from all these things. So, you specify one, the other is automatically known. So, that is a formula that we have from statistics and we say that at 1 minus alpha into 100 percent confidence level, the interval estimate is X bar minus this times s by N into X bar minus plus this into s by N; s is a standard deviation of the sample. S by square root N this is standard deviation of the means.

So, for the ith variable, this is a formula we end up getting, which is X i bar minus this s X i upon square root N; X i bar plus z alpha by 2 s X i bar upon N. Or if you express it in terms of the standard deviation of the mean, the relation becomes this and in uncertainty analysis, this is what we will be using most of the time. Here we give it a new name s X i bar is the standard deviation of the mean, this is what statistics teaches us; this is known as the standard error, which we will in the later on also call it as the standard uncertainty and take it forward in our analysis. So, that is an important thing that we have got and this is of great value to us.

(Refer Slide Time: 43:41)



So, this is the complete form of the answer that we were just looking at for the interval estimate, for the X i the ith variable that is the relation. And so, we have in just one line we will start expressing our result in this form X i bar, which the mean value plus minus z alpha by 2 s X i bar and this is at this much confidence level.

So, here are some representative numbers if z alpha by 2 is equal to 1; we say that 68 percent of the readings lie in that interval. If it is 2, it is 95.4 percent and we take that as 95 percent; if it is 3, it is 99.7 percent, we take it as say 99.5 percent.

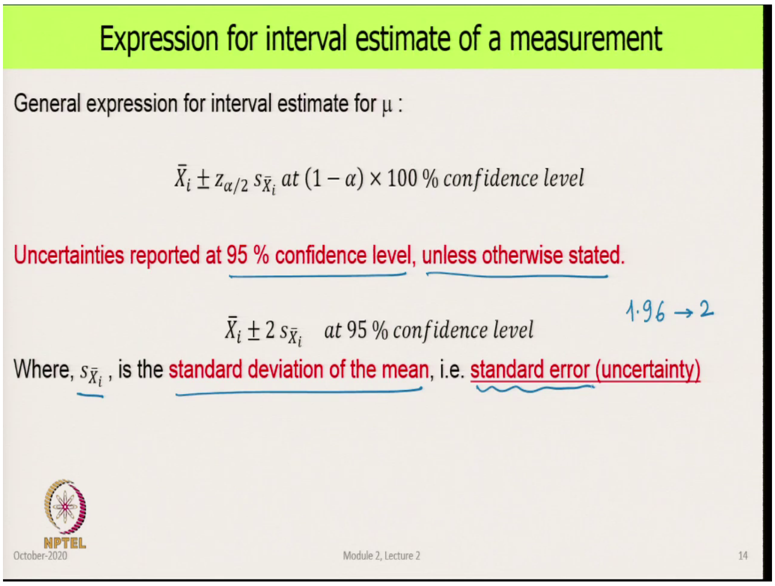
So, in uncertainty analysis, we will work only with this; we would not be bother about z alpha by 2 being 1.5 or 2.3 or something like that, we are going to stick with one of these. And in all uncertainty analysis that we will be looking at, we will report all uncertainty is at 95 percent confidence level.

So, it is this thing that we will be looking at; that means we will be reporting all our interval estimates at plus minus 2 sigma level. This is now an important thing that we have. This will stay with us throughout this course and we will come back to it again and again. So, at this point even if you look up say the specification of an instrument and there will be manufacturers who will quote that, the accuracy is this much in bracket at 2 sigma.

Or they will say that linearity is this much and they may not tell what it is; but if you see which standard it is it conforms to it will be ISO GUM or PTC 19.1, in which case it is all being referred to at 95 percent confidence level.

So, that is how important this is 99.5, 99.7 percent confidence level will come very rarely; in some arguments we will come across this, almost always we will stick with this. And all those statistics goes into all other values of this; for this course that is not a concern for us, we got what we want, which is this thing.

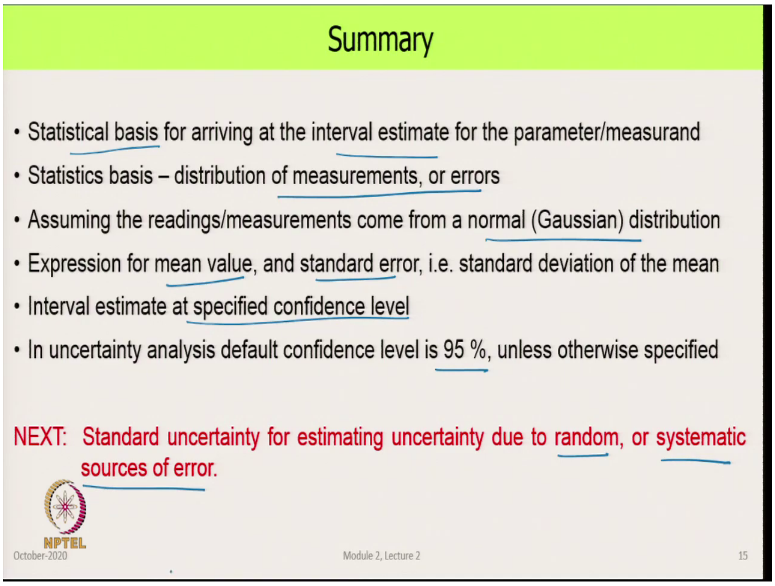
(Refer Slide Time: 46:16)



So, here is what we have the general expression for the interval estimate for mu is X i bar into z alpha by 2 s at confidence level at 95 percent confidence level which we assume is always the case, unless otherwise stated X i bar is 2 s X i bar at 95 percent confidence level.

So, 2 is an approximation, at 95 percent if you go very strictly that z alpha by 2 is 1.96; we are going to use 2, it increases the range, it is being more conservative in reporting the uncertainty that is what. So, we are with it; we are not under reporting something. So, s X i bar is a standard deviation of the mean and something that we will use it in the rest of this course, this is standard error or the standard uncertainty.

(Refer Slide Time: 47:10)



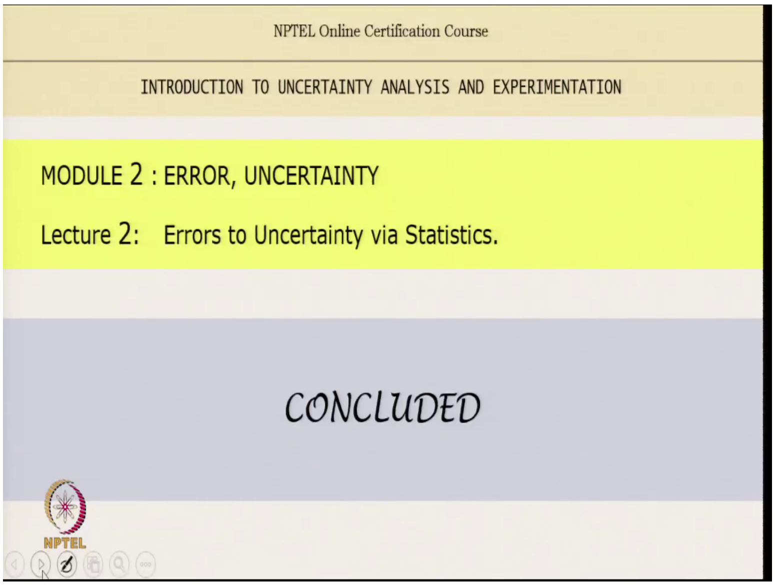
So, with that we come to the end of this lecture. And what we have done is, we have had a very quick look, sort of a revision of basic statistics as to what is the sample, how do we get data from a sample, how do we do calculations on it, and from there we estimate, given interval estimate for the population mean.

So, we have a statistical basis for arriving at the interval estimate; we saw the statistical basis, what is the distribution of measurements or errors. Assuming that means come from a normal or Gaussian distribution; we got an expression for the mean value or the standard error or the standard deviation of the mean and the interval estimate is specified at a particular confidence level, which is 95 percent unless otherwise specified.

So, what we have learnt here is basically is that, what we will do in the rest of this course when we report an uncertainty? We know that we are on sound statistical footing and we are also aware of what approximations we have done, which from a practical stand point are ok and so we do what we do.

So, in the next continuation of this, we will then see that having defined this standard uncertainty; how can we then define various other issues that come in the definitions of sources of error and start defining sub classification of these uncertainties.

(Refer Slide Time: 48:43)



So, with that we conclude the second lecture of module 2.

Thank you.