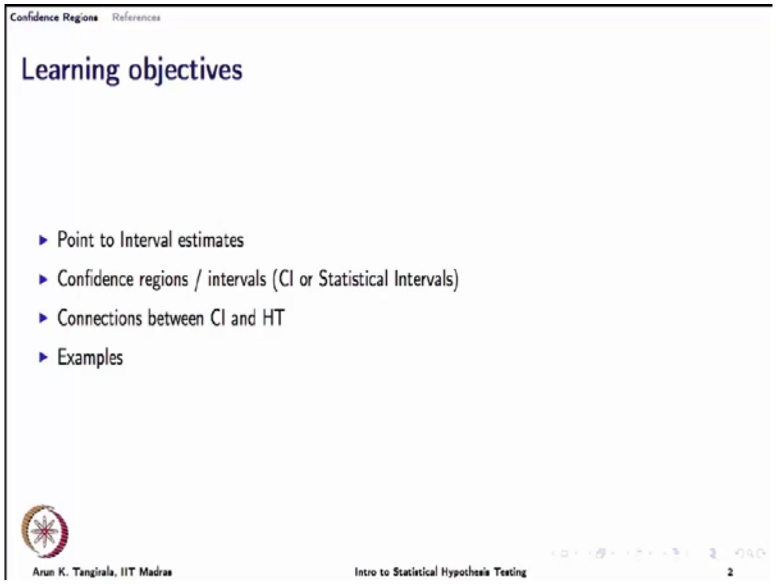
Introduction to Statistical Hypothesis Testing
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture-13
Confidence intervals and Hypothesis Testing

Hello friends, welcome back to the lectures on Introduction to Statistical Hypothesis Testing. In today's lecture, we will learn an existing concept known as the confidence interval or confidence region or another name is statistical interval. And once we will learn this, we will also learn the connections of this confidence interval with hypothesis testing that is we will learn how to conduct hypothesis test using the concept of confidence intervals. Specifically, what we do is when we construct confidence intervals, we are actually constructing what are known as interval estimates, until now, we have been talking of point estimates; that is estimating the truth as a single number.

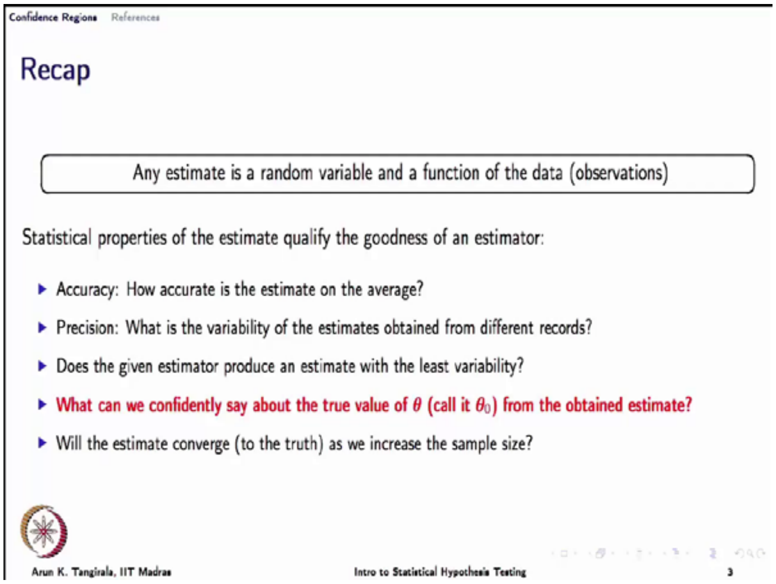
Of course, truth is also a single number and therefore, we decided the estimate should also be a number, but remember the truth is deterministic and estimate is a random variable, because it is derived from data. And therefore, the estimate could be one of the possible values that I have obtained; that means, there are possible values of which one value I have obtained as an estimate. And I still do not the truth even after estimating the parameter could be mean and so on. Therefore, I would be interested in giving or I would be required to give a region in which I believe the truth is and that region is essentially called the statistical interval or the confidence interval, we will learn as we go along.

(Refer Slide Time: 01:56)



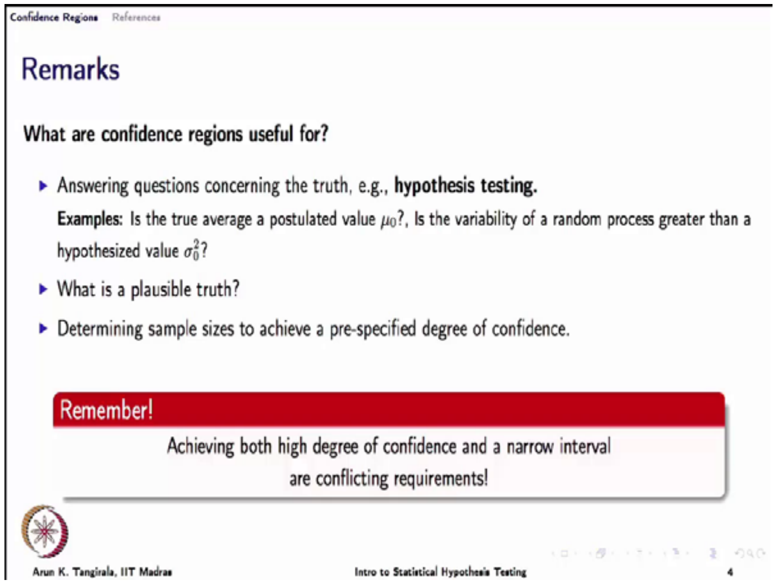
So, what we are going do is move from point estimation to interval estimation. Of course, we are going do in this in an elementary manner; more sophisticated way of constructing interval estimates is through a Bayesian analysis, which we do not discuss in this course. And then from this interval estimates, we construct a confidence region. So, the first step is obtaining an interval for the estimates that means, we are constructing many possible estimates however, from a single data set from that many possible data sets sorry estimates we construct the confidence region. Once we learn how to construct confidence regions, we will learn also the connections between confidence intervals or confidence regions, we will use this terms interchangeably and hypothesis testing which is the focus of this course. We have learned how to test hypothesis using the concept of critical values or p values until now, today we will learn a different way and you will hopefully like this approach. And of course, we will embellish our lecture with some examples.

(Refer Slide Time: 03:04)



To recap with regards to estimation, we must recall that any estimate is a random variable and it is a function of the data that is the observations. And because it is a random variable, there are many statistical properties that are of interest to us, but we are only interested in those statistical properties that are that will tell us something about the truth and we have talked about the accuracy that biased, precision, variability, efficiency and so on. In this lecture, we will be interested in this question that we had raised earlier as well on what can we confidently say about the truth, once we have obtained and estimate. As an example, I want to estimate the mean of some random process, random phenomenon, and I collect data that is I take a random sample compute the sample mean fine, but the sample mean will give me some value; from there, what can I say about the truth that is the goal in the construction of confidence regions. That is to be able to say something about the truth obviously, we cannot say precisely or accurately what the truth is that is not possible. Therefore, we will have to limit ourselves through a region in which we believe the truth resides, again there we may not be able to say with 100 percent confidence that is the limitation of estimating the truth from data.

(Refer Slide Time: 04:45)



Now, once we construct this confidence region, there are many uses through it and one of the prime uses of course, is hypothesis testing. So, once I construct a confidence region, I can test whether a postulated truth is admissible or not, is it possible right. This should not give you the impression the truth is a random variable. All we are going to be able to say is with some confidence or with a lot of confidence, whether the postulated value is a likely truth; truth itself is a deterministic thing, but our ability to answer that question has some uncertainty to it. So, remember that likely is not for the truth likely is for our answer. And also as I said, what is a possible truth; we have raised this question once or twice. Once I failed to reject a null hypothesis then the question comes about as what could be the truth, if I fail to reject the null hypothesis.

For example, in the comparison of yields of the reactors, the example that we discussed, we wanted to compare if yield of reactor A is a same as the yield of reactor B; that means, whether the difference is 0. And if you recall that example through a proper hypothesis test, we were able to say with the data that was collected that know there is not enough evidence to believe that the yields are identical. In fact, the evidence is pointing towards different yields then the question arises as to what is the difference between the yield of reactor A and yield of reactor B. Now, that the truth postulated truth of 0 has been rejected would the difference be 1, difference be 2 or 3. 4, we do not know. Confidence regions would allow us to answer that without going through a repeated hypothesis testing and that is one very nice use of confidence regions.

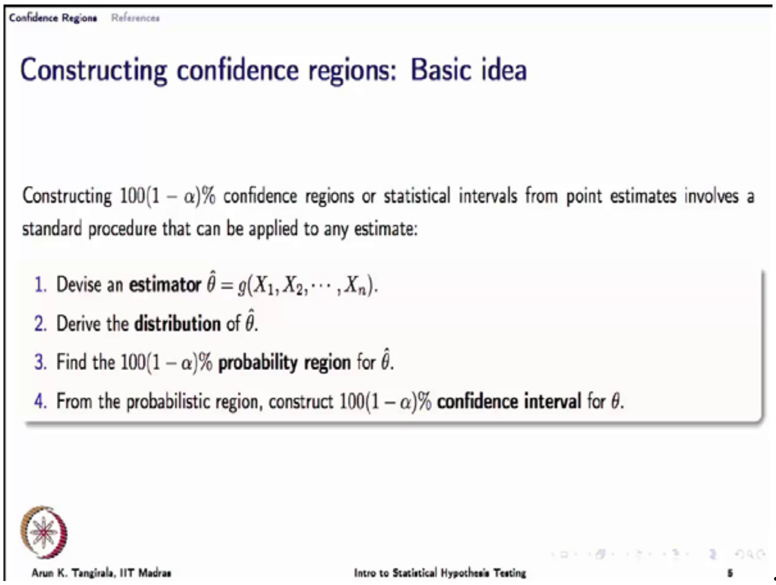
And then, there are other uses pertaining to the experimental design aspects. For example, for a tolerable error that is between the truth and the estimate, average error, I can ask how many minimum samples sorry minimum observations I should collect that is what should be the minimum sample size to achieve a certain degree of error in any estimation exercise. And of course, for certain applications like or certain estimation problems, this is an easy question to answer, because we will have analytical expressions for the width of the confidence interval and the sample size as we will see shortly. And for certain other estimation exercises, one may have to do adopt a trial and error approach.

Before we plunge into the confidence region construction, an important point to remember is that achieving high degree of confidence that is to be able to say with the high degree of confidence, what does it mean I will not be able to anyway say with 100 percent confidence that the truth is in some interval. In fact, if you ask yourself, what is the interval in which the truth is with 100 percent confidence? For example, let us say the room temperature, the temperature of the room in which you are sitting if I ask you what can you say what is a region in which a true room temperature is with 100 percent confidence, it has to be between minus infinity and infinity I mean that is obviously an answer. But that is practically useless confidence region what we want to be able to say is with 100 percent confidence that the room temperature for example, is between 0 and 10 degree Celsius can we do that unfortunately it is not possible.

So, what do we do in practice well we back of and say look I cannot say with 100 percent confidence that the truth is in this finite interval. And I obviously want a finite interval then what do I do at least I should one limit should be finite the other limit may be infinite, but one limit at least should be finite. So, what do I would, I would sacrifice on the degree of confidence and I say let me say something with 95 percent confidence, 99 percent confidence, we will understand shortly what a 95 percent confidence means 99 percent confidence means, but qualitatively we do understand at this moment. So, we want to be able to achieve high degree of confidence; and at the same time, we want the interval the width of the interval to be as small as possible the interval to be as narrow as possible; for example, I would be like to be able to say that the room temperature is between 22 degree Celsius with 99 percent confidence.

Now, unfortunately, these are conflicting requirements that is as I tend to increase the degree of confidence generally for a given data set for a given phenomenon, the width of the intervals starts to increase; obviously, because a 100 percent confidence interval is the widest. So, if you want an narrow interval, you will also have to shrink the degree of confidence. And, a good estimator is that which among all the estimators will give the narrowest interval for a given degree of confidence that is what we mean by an efficient estimate, very good.

(Refer Slide Time: 11:06)

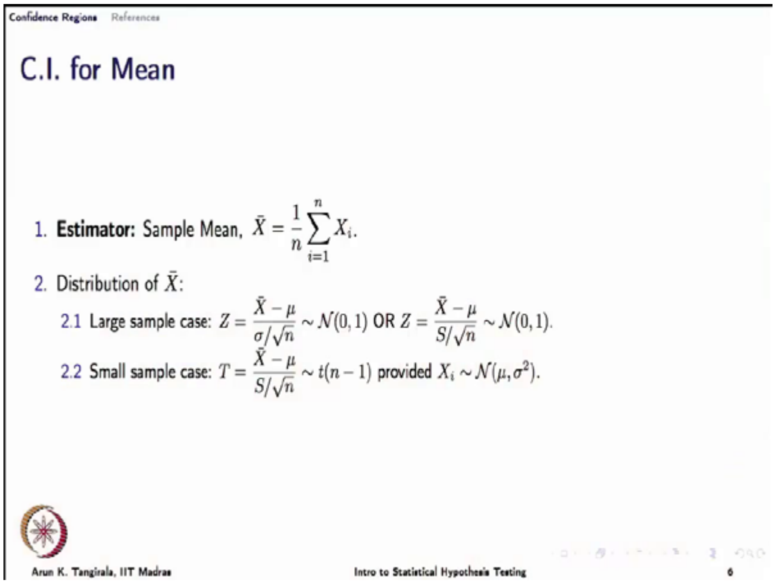


So, now let us quickly understand how to construct a confidence region remember this is a very introductory lecture on confidence regions, but has probably all the elements that you need to construct a confidence region. Any other problem of that involves confidence region construction, we will essentially involve the same steps, but we are not going to discuss too many scenarios we are only going to discuss how to construct confidence regions for the parameters of interest. And then I will just show on a few examples the connections between confidence regions and hypothesis testing. So, let us go through this basic procedure for constructing a confidence region for any situation any estimation problem.

The first thing is to devise an estimator this means that the confidence region obviously, depends on the estimator naturally so, because the width of the confidence region depends on how much error an estimator is making in estimating a parameter and that can change with the estimator. So, first choose your estimator. For example, in the estimation of mean, your estimator could be sample mean, could be sample median, could be sample mode whatever you choose to do so. Once you have chosen an estimator, then the next step is to derive the sampling distribution of theta hat. Again, we have gone through that in the lectures. And then from the sampling distribution of theta hat find out the 100 1 minus alpha percent probability regions for theta. What is this alpha this alpha definitely is related to the type I error? So, there is a connection between confidence region and hypothesis testing. For now, ignore the connection alpha you can think of as a compliment of the degree of confidence that is if I want to construct 95 percent confidence region then alpha would be 0.05. If I want to construct 99 percent confidence interval re region then alpha would 0.01.

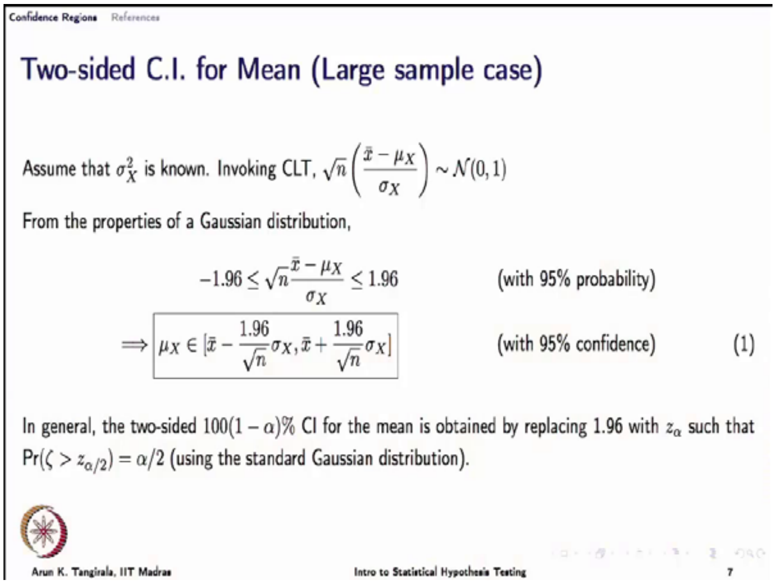
So, first I construct a probability interval for theta hat; theta hat is a random variable. So, it makes sense to use the term probability region. From where that is from this probabilistic region, I construct the 100 times 1 minus alpha percent confidence region for theta. So, notice the statement is very clearly in step three we are constructing a probabilistic region for theta hat; and in step four, we are constructing a confidence region for theta. Always remember the confidence regions are for the truth not for theta hat. Theta hat is a point estimate unless we make a calculation error; we are sure about theta hat for that data for that estimator. What we are doing is we are constructing confidence regions for theta. Why do not we say probabilistic region for theta, because in all our analysis theta is a deterministic quantity; it is not a random variable. Unless you move to base in analysis where theta is also treated the truth is also treated as a random variable; in classical analysis, the parameters of interest are deterministic variables therefore, we do not use the term probabilistic region, because there is no randomness in it better term would be confidence region. It reflects our confidence clearly on being able to isolate the truth in a certain region that is what is meant by confidence region.

(Refer Slide Time: 14:33)



So, let us understand this with an example. Suppose I wish to estimate mean and I choose sample mean as the estimator we know what this sample mean is. And we also the distribution of sample mean under different conditions. Let us take the simplest the large sample size case even samples could follow out of a Gaussian distribution, does not matter large sample case we know that X bar follows a Gaussian distribution provide at the data comes from a random sample by CLT.

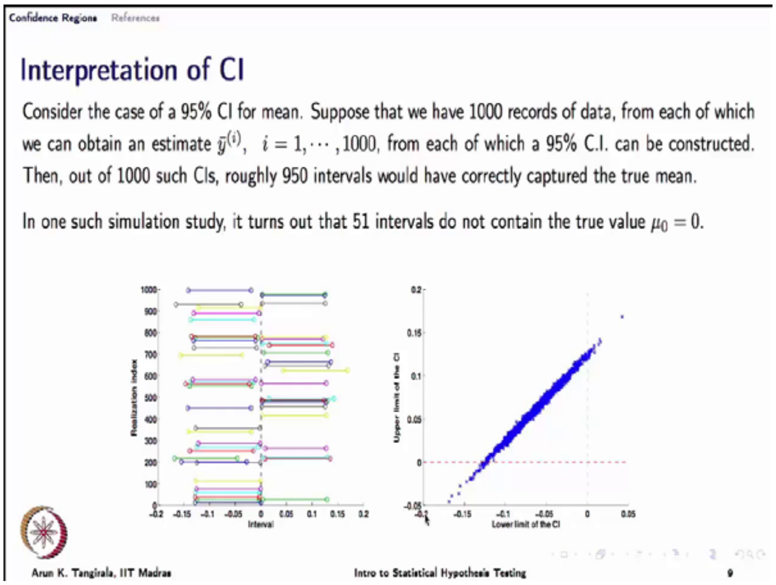
(Refer Slide Time: 15:03)



Suppose that is the case then we know that the standardized statistic follows a standard Gaussian distribution which means now if we look at the equation here is a from the properties of a Gaussian distribution, we know that the standardized Gaussian variable takes on values between plus or minus 1.96 with 95 percent probabilities. So, now we are in step three; step one was choosing the estimator which is sample mean; step two was writing the sampling distribution which we wrote already saying that the standardized sample mean follows a standard Gaussian distribution. Step three is to write the probabilistic interval, and we are only one step away from constructing the confidence region which is step four and that is your equation one here, which shows the confidence region for the truth which is mu X. How did I arrive at this interval? Very simple algebra, take each of the bonds here the upper bond and the lower bond on the probability region for z and then rewrite it in terms of mu.

For example, you would write route n times x bar minus mu by sigma is less than equal to 1.96, from which you would be able to say that mu is greater than x bar minus 1.96 by route n times sigma. So, you obtain the lower bond by taking the upper bound for the z; lower bound for mu is obtained by taking the upper bound for z. And likewise, the upper bound for mu is obtained by looking at the lower bound for z that is obvious because here mu as a negative sign on it in the z. So, what this interval is the 95 percent confidence interval for mean means there is 95 percent chance now again that the truth is in this interval that the true value is a deterministic quantity that means, there is a five percent chance that this interval may not contain the truth. What is this 95 percent chance and five percent chance business by the way this is called the two-sided confidence interval; very soon, we will talk about one-sided confidence interval.

(Refer Slide Time: 17:27)



Let us understand this two-sided confidence interval carefully. It is very easy to misinterpret confidence intervals. In order to understand this 95 percent confidence interval that the truth is outside this interval has a 5 percent chance. Let us set up a simple simulation experiment you can also do this on your laptops, I am not going to show you this in R, but maybe you can take it as an exercise, write a small script that will do what I am showing you on the slide. What I have done here is I have taken thousand data points that is from a sampled thousand observations from a Gaussian population of zero mean let us say I know the truth using R nom you can say in R. I have randomly sampled thousand observations and computed the sample mean and I have repeated this thousand times.

Now every time I randomly draw a sample thousand observations, I get an x bar; and from that x bar, I can construct this 95 percent confidence interval. What is it mean, for every data record, I have 95 percent confidence interval. I can also have an 99 percent, but we will only stick to 95 percent for now that means, when I have thousand records of data, I can construct thousand confidence regions one per data record. When is 95 percent confidence interval means that out of this thousand confidence intervals that I have constructed from thousand data records roughly about exactly speaking if I had repeated this infinite times then may be a million times then it would be almost exact, but thousand let us say roughly about 50 times 50 of the 1000 confidence intervals will not contain the true value. What is a true value? It is 0; and what I am showing you here in that figure are those 51, because I am only repeated 1000 times you are not going to get exactly 50 intervals that will miss the truth, but 51 intervals it turns out will miss a truth.

In fact, you can repeat this; you can just go to R and use replicate collect the sample means, and from the sample means, you can construct the confidence regions. In your case, it may be 49, 51, may be 50, 52 at times possible, but not more than that. In my case, I have 51 confidence intervals missing out the truth. How do I know that? Because, the truth is in the middle line here 0 and the lines that I am showing you here with the circles at the markers are the confidence regions. These lines are either completely to the left of 0 or to the right of 0. What are those lines? They are the intervals; those 51 different intervals that I have constructed obtained out of the thousand data records have missed the truth; 0 is not contained in that interval. The remaining 949 intervals would contain 0.

Now, the question is not whether 0 is to the extreme point towards a right of the confidence region, towards the left of the confidence region, there is absolutely no comment on that. There is nothing to interpret there; you cannot say that the truth is most likely going to be the left extreme or the right extreme nothing like that. We know all we can say is the truth is contained in that interval that is all, where it is in that interval it is not possible to say; that means, any value in that interval is a possible truth, you see we have answered one of the questions. We do not know what the truth is, so any value in that interval is a truth, so which means if 0 is in that interval 0 is also a truth. And, if 0 is what you have postulated as the truth and it falls within the confidence region then your null hypothesis is not rejected that is the basis of using confidence regions for hypothesis testing.

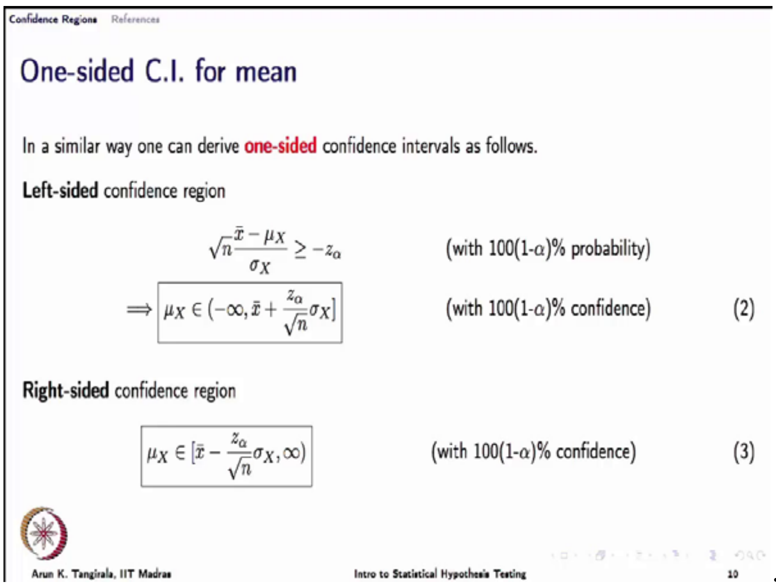
There is another way of viewing these intervals, where you plot the lower and upper limit on the x and y-axis respectively. And you see here you cannot count, it is very difficult here, but roughly about 51 points are outside this quadrant, this second quadrant you may say. The 949 points that are within this second quadrant will capture 0, the truth in the confidence interval. The remaining 51 are either in the first quadrant or in the third quadrant, alright. Notice that here on the x-axis is 0 is not for the x value 0 is not at the left corner, but the 0 is where the green dash line is for the x-axis. And for the y-axis, a 0 is not at the corner again it is at the red dash line. So, read this plot carefully. These are two different ways of showing the confidence region.

So, hopefully now you have understood what that 95 percent of confidence region means, what that 5 percent chance of missing out the truth is. That means, if for our starts are bad, if the data is that realization for corresponding to which the confidence interval is not contained the truth then it is bad luck, yes, we have missed out the truth. But, the chance of obtaining such a realization is 5 percent that is what this 95 percent confidence region means. Obviously, if you want to minimize the chances of missing out the truth, what do you have to do, you want to increase 99 percent confidence, right; you want to be able to say more confidently that means, you want to minimize the chances of missing out the truth in the interval, but the sacrifice that you make is on the width of the interval.

Why, go back to the derivation here instead of writing a 95 percent probability region write a 99 percent probability region then this would become around 2.58, the left and right side would be minus 2.58, 2.58; that means, the interval is now roughly 6 over root n times sigma or you can say 2 times 2.58. So, the width of the interval has increased from 4 sigma by root n to 6 sigma by root n, but you have now fewer chances of missing out the truth, but the interval is wider. Now, you can think if I want to construct 100 percent confidence region, what would be the confidence interval minus infinity to plus infinity. As you increase the confidence level, you want to increase the confidence level you have to sacrifice on the width of the interval.

Now obviously, here we assume sigma to be known and this interval is only applicable when you are using sample mean as an estimator. What does this mean if I use sample median I have to go through this exercise all over again, and the interval can be quite different that means where things changes at step one itself. The moment you change the estimator, step two is effected the sampling distribution is effected, step three is therefore influenced by step two and step four is influenced by step three. So, all are interconnected. What this means is the confidence region realize heavily on the estimator. And this procedures that we have followed for deriving the confidence interval for mean is exactly the same for all other parameters variance, be difference in means, be it proportions, be it ratio of variability whatever may be the case, whatever may be the parameter that you are estimating the procedure is same. Therefore, I am not going to repeat in interest of time and also it will be quite boring to do so. To for different situation that is for variance and proportion, ratio of variances and so on, you simply have your sampling distribution lecture next to you, write the probability interval and derive the confidence region. Go through that exercise for a few parameters, in fact, I have given the results here the results are also available in textbooks by Montgomery and (Refer Time: 26:03) or any other standard textbooks, but I have collected those results for you. And you can cross check if you have obtained the confidence regions correctly. Our interest would be more on relating confidence interval to hypothesis testing.

(Refer Slide Time: 26:22)

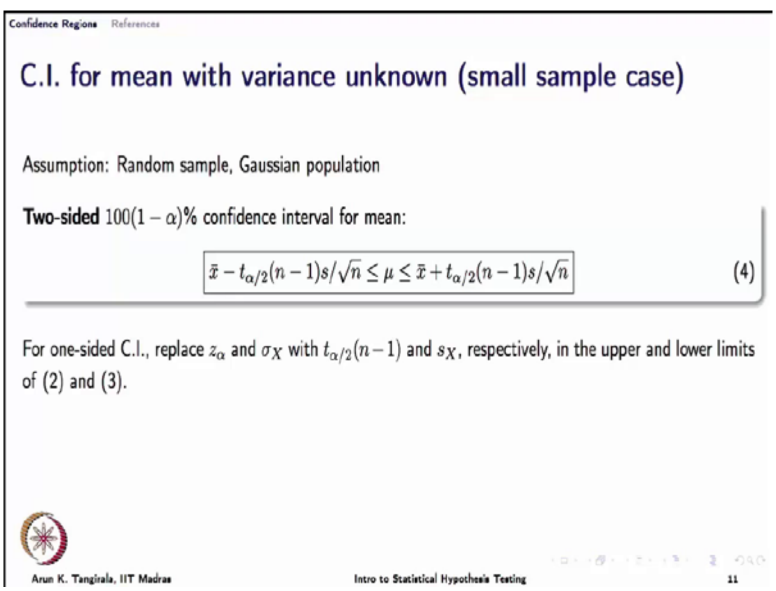


Before we do that lets quickly complete this understanding of confidence region where we now discuss the one-sided case; earlier we derived the two-sided confidence region. We will now derive the one-sided confidence interval for mean. Again the procedure is the same; assume sample mean is your estimator we know the sampling distribution. Now, let us say, I want to construct what is known as a left sided confidence region that is when I am constructing the low tail test I would like to actually construct the left-sided confidence region. So, we know in a lower tail test, the probability that z is greater than or equal to minus z alpha is 100 times 1 minus alpha. Remember alpha is the probability of the observed statistics falling within the rejection region, and the rejection region is z less than or equal to minus z alpha, so the probability that z is greater than or equal to minus z alpha is 100 times 1 minus alpha percent; from which, you derive the left-sided confidence region for the mean. It is fairly obvious that the left extreme is minus infinity and the upper bond is x bar plus your z alpha earlier we had z alpha by 2, now it is z alpha remember by root n times sigma.

For example, if alpha is 0.5 z alpha would be 1.645. This now can be used to conduct hypothesis test of the lower-tailed type. If the postulated value falls within this confidence region, then we fail to reject the null hypothesis. If the postulated value is greater than the upper bond then we reject the null hypothesis, very straight forward.

And likewise, you can derive the right-sided confidence region, where you say that the probability that z is less than equal to z alpha in a right-sided test the critical region is z is greater than or equal to z alpha. Therefore, the probability 100 minus 1 alpha percent probability region is z is less than or equal to z alpha; and from which, you derive a lower bond for mean. And once again these are used in testing hypothesis of the you know right-tailed or right-sided or upper-tailed test, where you say if the postulated value is less than the lower bond that I have derived, then I reject the null hypothesis. If it is greater than the lower bond, then I fail to reject the null hypothesis; that is all to it.

(Refer Slide Time: 29:07)

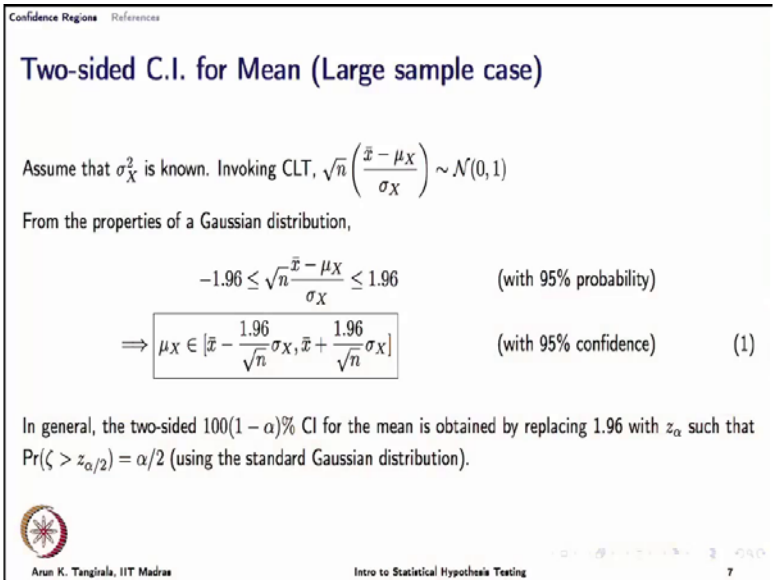


And of course, then you have different situations. Now, I am not going to go through this procedure. The next case is mean with variance unknown we have until now derived the confidence regions two-sided and one-sided for the case of known variance. Now as usual as we have discussed in the sampling distributions and hypothesis testing, we are going through those corresponding scenarios. As I said earlier, we will not go through the derivation, it is a straight forward from the sampling distribution you write the probabilistic region; from the probabilistic region, you write the confidence region. The story is the same. So, from here on wards, it is more of reading out the confidence region even I may not do that I may just point to the equation.

So, here equation 4 gives you the two-sided 100 1 minus alpha percent confidence interval for the mean when the variance is unknown and the population has a Gaussian distribution and the sample size is small. If the sample size is large, you replace the t with z there. So, clearly the choice of alpha plays also a roll in the confidence region that is got to do with your degree with confidence. There alpha had got to do with the type one error; here alpha has got to do with the degree of confidence, but the interpretation of alpha is the same. Now we have in fresh interpretation what we thought is type I error as alpha, now is 100 1 minus alpha percent is a degree of confidence, it is a nice thing to learn, alright.

And again from equation 4, you can obtain the one-sided confidence intervals by replacing z alpha and sigma x either from equation 4 or from this equations 2 and 3, where you replace sigma x by s x the estimated standard deviation and z alpha by t alpha with the degrees of freedom n minus 1 that is all. Otherwise, the story is the same. Now, very often when confidence intervals are discussed even if you look up the book by Montgomery, immediately one tries to derive the relationship between the width of the interval and sample size, remember we ask we said one of the uses of confidence intervals is in experimental design. And, a decision in experimental design is how many observations to collect what should be the sample size.

(Refer Slide Time: 31:36)



So, for example, if I go back to the two-sided confidence interval here, the width let us say for discussion purpose is the width of this 95 percent confidence interval for mean using sample mean as an estimator is 4 sigma by root n that is the error you can say that is the band. Now, if I say that I specify that this the band that I am willing to tolerate at the maximum then what should be the sample size, I say 4 sigma by root n is the b the width of the band, and b is known sigma is known anyway I can calculate n. So, it is a very easy way of determining the sample size for a pre-specified error band or you can say confidence band whichever way you want to look at.

Likewise, in every confidence interval analysis, I can derive a relation between the sample size and the pre-specified that is a user specified error band. Now, it was easy for the sample mean case where the variance is known and sample size is large. But unfortunately, when I am looking at this situation, where the variance is unknown and small sample size, then it becomes difficult because the width is not only dependent on n, but also on t alpha by 2 with n minus 1 degrees of freedom and the standard deviation. They are implicit functions of n; they are not independent of n; earlier z alpha by 2 was independent by n, there alpha was 0.05. So, z alpha by 2 was 1.96. And sigma was known. So, the only unknown was n which appeared explicitly it was easy to derive. However here, it is quite difficult to arrive at an expression for the sample size in terms of the pre-specified band. We will not pursue that, but just to let you know in such situations, you have to adopt the trial and error approach.