Elementary Stereology for Quantitative Microscopy

Prof. Sandeep Sangal
Prof. S. Sankaran
Department of materials Science and Engineering
Department of Metallurgical and materials Engineering
Indian Institute of Technology, Kanpur
Indian Institute of Technology, Madras

Lecture – 06

Probability Distributions

So, continuing from the last lecture we were discussing distributions, now the thing is we need to estimate values from our experimental data. So, for example, if you are making measurements of point fraction, you have to estimate an average point fraction, you want to estimate what kind of a variation in the data is there. And idea of all of these estimates are that you want to predict the parameters of the population. So, if your population is following a normal distribution you want to predict it is mean, you want to predict it is variance standard deviation etcetera.

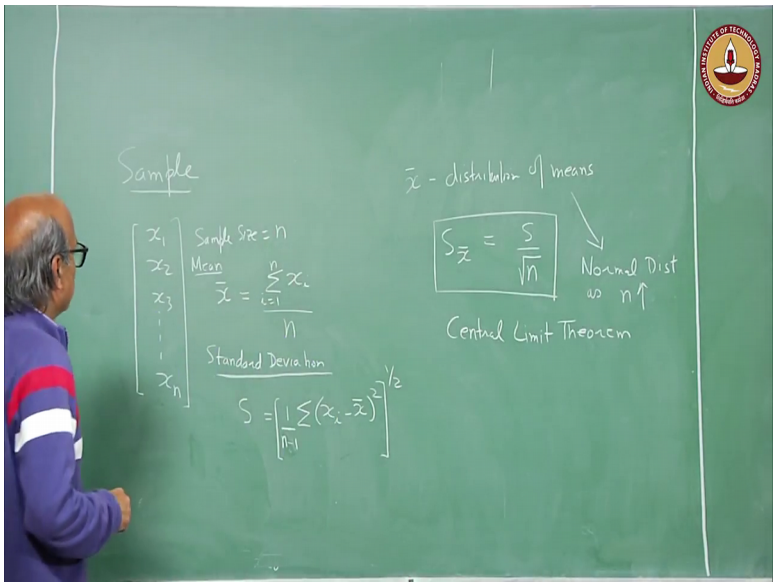
So, how do you do this? Because you are not looking at the entire population you are only looking at a very small part of it, you are just looking at your sample and even in that sample you are looking only at a few areas of your microstructure and from that you want to predict the population. This problem is similar in India to we keep getting elections and then whenever an election comes, there are people who come forward and start predicting things right that, this party will win, that party will win, this guy will become Prime Minister, this guy will lose. So, what do they do, they go to people they take an opinion poll right.

So, opinion poll they do not ask every voting individual in the country, they ask a small sample in different regions of India. On that basis then they analyze the data and start doing a prediction and what are they trying to predict? They are trying to predict how the population will behave when they go for votes, same thing happens in the exit poll how did the population behave, what was their voting pattern? Again they look at a small sample on the basis of this they are trying to predict what was the populations vote? Here also it is exactly the same problem that we are looking at a small part of the material on the basis of the measurements we make in the all part, we are trying to predict it the; its

microstructure the bulk microstructure, what is a distribution? What is it is mean? What is it is variance?

So, the first thing that we do in any prediction is one takes a sample.

(Refer Slide Time: 02:43)



A small sample like in the one of the lectures you were given a nodular cast iron microstructure. And it was from a small area from which we are trying to predict, what is the volume fraction of the graphite nodules? In general terms what you would do is, you would make measurements x 1, x 2, x 3 and let us say go all the way to x n. That is in one region of microstructure, you measured perhaps a point fraction, in another region you measured x 2 point fraction, x 3 point fraction, x n point fraction.

So, this is your sample and we say that your sample size here is n; obviously, intuitively one that larger is our sample size better should be our prediction. Of course, accuracy of prediction also depends on the tools you are using more accurate tools are using you will get better prediction, one of the problems of statistical prediction is sometimes your tools have a bias, what does a bias mean let us say you are measuring distance and suppose your ruler has a bias that your ruler is off by 0.1 millimeters, let us say a 1 millimeter it is off by 1 millimeter; the whole thing is graduated in such a way that it is off by 1 millimeter or so, every let us say 100 millimeters it is off by 1 millimeter. So, there would be a bias depending on whether it is on the downside or the upside, you will make consistently smaller measurements or higher measurements that is called a bias ok.

So, we are assuming in all of this that there is no bias; what we are saying is that your measurements will be subjected to statistical errors or random errors which some of them will be plus and some of them will be minus. So, I have now a sample of size n so, my what is my mean value? Well my mean value is I just add up all of the values have obtained all the n values have obtained and I divide by n; this gives me the estimate of the mean of the what I am assuming here is, this gives me the mean value for this sample. But then I am making another statement here that this is I am trying to say is an estimate of the mean of my population if I had measured, all possible areas I could measure ok.

So, this becomes mean and the second most important parameter that one measures is the spread of your data through standard deviation. I am going to use the letter S for this to distinguish it from sigma, sigma will keep it as the population standard deviation and S is my sample standard deviation and this should be equal to what? How do I get from this data? Now again you have to do sum of the squares of the deviation about the estimated mean x bar. And you will typically see sometimes from 2 kinds of formula written for this, you would see very often a forward you will see in the denominator n written there the number of data points and then the whole thing is taken to square root, if you do not take it to the square root then this is an estimate of the variance S square.

Now, here for large N this is ok, but when n becomes very small then this is an incorrect expression because then this does not give you what is called as an unbiased estimate of the standard deviation, to get an unbiased estimate of the standard deviation you must divide by n minus 1 and not n strictly speaking it should be divided by n minus 1.

There are one can view this that why we are dividing this by n minus 1, yes n minus 1 represents the number of degrees of freedom in your data. In this expression you have already estimated x bar so, that the number of degrees of freedom has gone down by 1 you had n degrees of freedom it becomes n minus 1 degrees of freedom for the simple reason that independently you can choose n minus 1 values, but the nth value then will be fixed because you have already have x bar.

So, only n minus 1 independent values can be chosen so, this becomes degrees of freedom, another way of looking at it is just imagine that your sample size was 1 ok. If a sample size was 1 what would be the standard deviation, what would be denominator is

clearly 0 if a sample size is 1 x and what would be the numerator only 1 sample. So, x bar and x would be same so, it will be 0 upon 0, which means it is a indeterminate quantity for a sample size of one you cannot determine a standard deviation from one sample data point you cannot have a spread. So, this formula is expressing that as well that you cannot calculate a standard deviation out of one data point you have to have more than one data points.

And so, we calculate the standard deviation this way and then we are also making a statement here that this standard deviation represents the standard deviation of the population so, far from a sample you can get these 2 parameters. Now, the next question that arises is that what kind of a interval I can place on my mean value because you should realize that suppose I do one set of experiment I get a sample of size n so, an I get up some value of x bar and standard deviation.

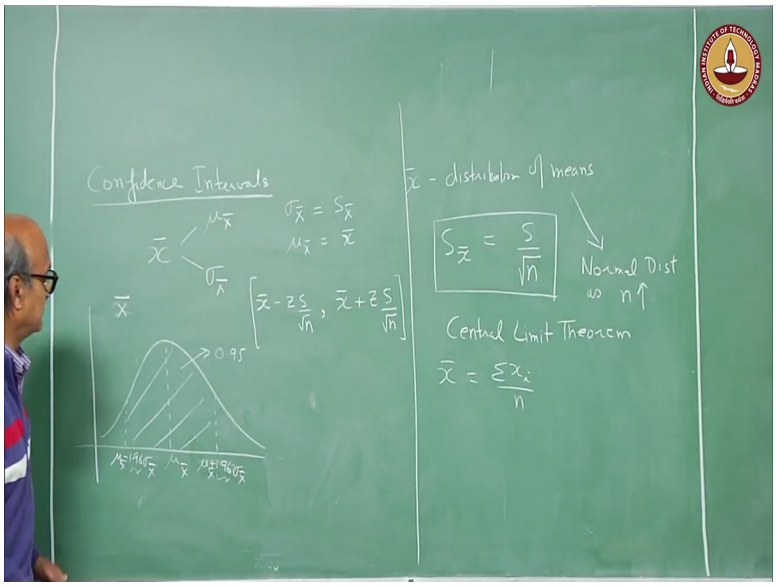
Now suppose I repeat this experiment, I do another I take another sample out I take more areas I measure more point fractions I get second sample, will I get the same mean value? No I will get a somewhat different mean value and imagine that I can take a third sample I will get yet another mean value a fourth sample yet another mean value and in the process what I will get is, I will get a distributions of means ok. I will get a distribution of means this itself will follow a distribution of means and it can be shown I will not be showing in this particular course because of time constraints. It can be shown that the mean of the distribution of means would also be predicting the mean of that original population ok.

But what about it is variance or standard deviation, the standard deviation of the distribution of means would be different and this can be shown to be again in this, this is beyond the scope of this course. The standard deviation of the means can be shown to be the standard deviation obtained here divided by square root of the sample size. If this is a very key relationship that one gets and the second more important thing that comes out is that as n becomes larger my probability distribution of the means tends towards a normal distribution.

This comes from what is known as the central limit theorem in statistics ok, then there is a theorem called the central limit theorem which says that s n increases the distribution of the means tends towards us towards a normal distribution. So, this goes towards the

normal distribution as n increases and s n tends to infinity this becomes the normal distribution, regardless of how the individual x distribution is regardless of how that distribution is the distribution of the mean move towards the normal distribution so, for large n there is no problem in considering the distribution of x bars as a normal distribution. Now just with this thing in mind now we can develop confidence intervals so, there is a that this is how we will do that.

(Refer Slide Time: 14:22)



So, what we are trying to do here is, we are trying to produce an interval which has a lower bound and an upper bound and we say that we expect that our mean should lie in this interval with a certain probability ok. So, if I say that I want to associate it with the probability of 0.95 that is 95 percent of the time my samples will have a mean in that interval. So, if I drew 100 samples 95 of them I expect them to lie in that interval how do we calculate that is actually quite straightforward with the background that we have from the previous lecture as well that let us assume that x bar follows a normal distribution.

So, x bar follows a normal distribution it has it has a mean of mu it has a standard deviation of sigma then we have already seen that if this is my distribution, this is my mean and if I go let us say 1.96 sigma to this side that is I go mu plus 1.96 sigma and mu minus 1.96 sigma, then this area is 0.95 or 95 percent of the total area.

Now, here and this is a distribution of means. So, this represents the distribution of the random variable which is X bar. So, if I know an estimate of mu and if I know the

estimate of sigma then I should be able to and this is the let me let me actually call this as X bar so, that ok. So, if I then I will be able to this becomes my confidence interval that is what this means and how do I get this in this sigma X bar is nothing, but S x bar I have estimated from my sample data and mu X bar is nothing, but X bar that I have calculated that is this is what I have calculated. So, both of them I know then I get my confidence interval.

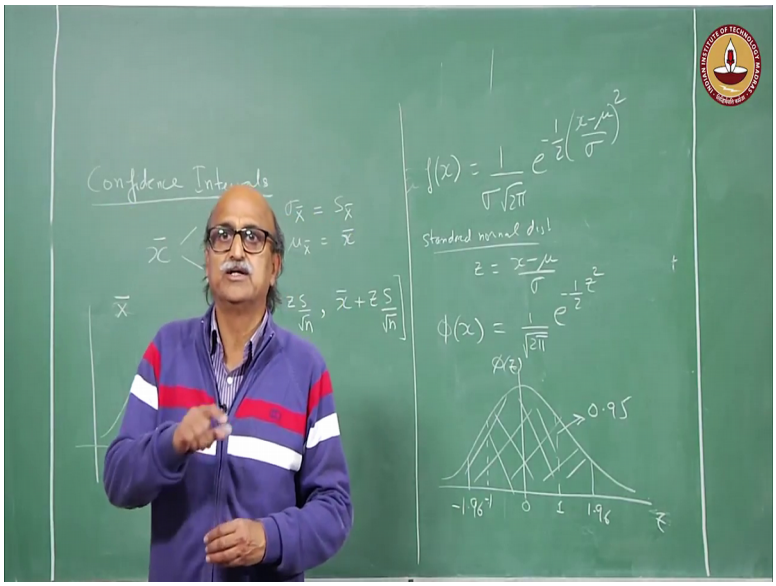
Now, this is 95 percent confidence interval, but then I can work with other intervals as well I say I do not I want higher certainty of my interval. So, then I said you know I want to be sure 99 percent of the time. So, I would like to work with 99 percent confidence interval then this value will change it will not be 1.96 it will be some other value. So, in general then if I were to write a confidence interval then this would be the lower bound would be x bar minus some value z times S divided by square root of n ok.

This represents a lower bound which is coming from here all I have done is substituted mu X bar and sigma x bar and S x bar x S bar is this and the upper bound is x bar plus z times S divided by square root n. So, this is my confidence interval ok, where the value of z will be chosen in such a manner that we get the desired level of confidence I may work with only 90 percent confidence interval.

So, how do I get this value of said now so, these values of z can be obtained from statistical tables which I will just show, but before I show you that table I want to bring in those tables are created by what is called as the standard normal distribution. Have you heard of the standard normal distribution, it is actually nothing it is simply a normal distribution with mean of 0 and standard deviation of 1 and from that data for that standard you for that particular distribution can be used to calculate the value of z and essentially if you look at the if you look at the formula for standard deviation.

Now, sorry if you look at the formula for the normal distribution it is.

(Refer Slide Time: 20:50)

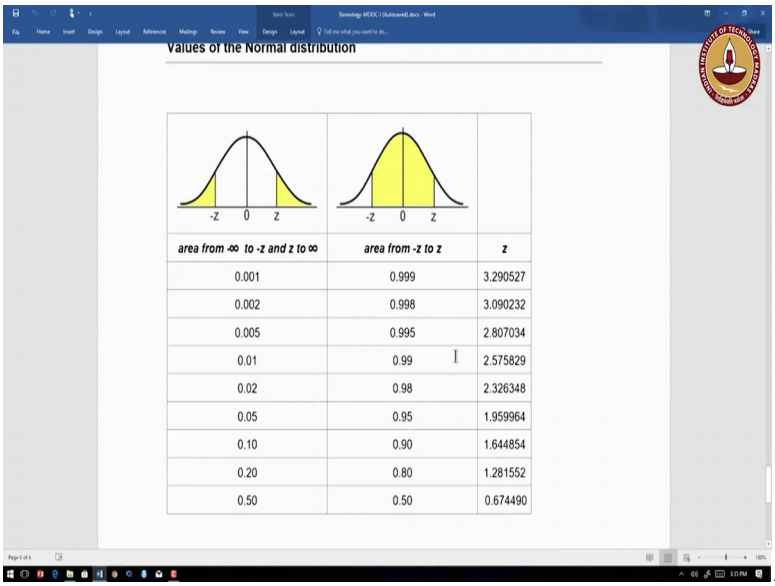


This is the probability density function for this if I want to convert this to a standard normal distribution do the following substitution z is equal to x minus mu upon sigma ok. If I do this distribution the density function phi x, I am representing for the standard normal distribution is simply 1 upon square root of 2 pi e to power minus half z square, which essentially means that you have 0 mean if I put 0 here and if I put standard deviation 1 then you will be left with this.

So, if I look at that plot and this is how the standard normal distribution will show up this is 0 0 mean so, this is my probability density. Now if I look for 95 percent of the area on either side of the mean total area if this area has to be 0.95 then what should be this value, this value should be mean plus 1.96 of the standard deviation, but standard deviation is 1. So, this is actually z I should write here. So, this is 1.96 and this is minus 1.96, if I choose the only 1 z is 1 plus 1 minus 1 then this area is 0.68 mean minus 1 time standard deviation to mean, plus 1 times standard division mean being 0 so, it is simply this ok.

So, now I would be able to get z values from here and which I can plug into this relationship to get my confidence interval for different levels of confidence. So, I need the p p t.

(Refer Slide Time: 23:44)



So let us look at this table of standard normal distribution and it is essentially giving me values of z for different areas under the curve. So, you look at this right curve this gives me area from minus z 2 plus z, the left one you can ignore this is giving me basically total of the areas in the left tail and the right tail. So, you just have to basically look at the look at this particular column giving me the area from minus z to plus z, these on the left side is only complimentary area it will be 1 minus that.

So, I want to look at 0.95 what is the values z for 0.95 area value is 1.96 right, if we take it to second decimal place. If I choose 99 percent 0.99 then the value of z becomes 2.57 or 2.58 if I choose 99.9 percent the value becomes 3.29 depending on what kind of a error you want to put. If you want to call it as an error or you want to put to your data you can you how you want to put a very high probability and then you will go for higher area under the curve, but if you look here what is going to happen to your interval that also you should realize that as your z value is increasing your interval will become bigger.

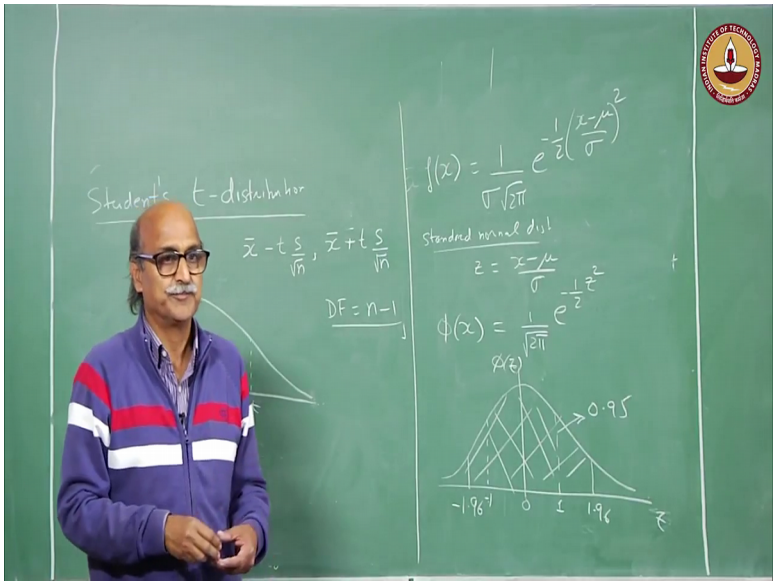
So, if you want to be more and more certain about what you are reporting you should choose a high z value representing higher confidence 99 percent confidence for example,, but at the same time your error bar will also increase. So, you have to judicially choose you do not want to you can not make an error bar very small by choosing a small confidence interval well you know only 10 I am going to report only

for 10 percent surety right so that, that would not work. So, with this is straightforward right is this clear. So, you will be given a demonstration of how to with the data that was collected how to select it.

Now, assuming a normal distribution how well it that is, is if you have large value of n I will come to it what does this large value of n mean, but if you have very small value of n then this assumption of normality or a normally distributed mean values may be incorrect, then one can take recourse to another distribution which is known as the students t - distribution.

Now students t distribution is also a bell shaped curve, but it is somewhat it is somewhat wider than the normal distribution and so, essentially then my confidence interval will become x bar minus t times S upon square root of n to x bar plus t times S upon square root of n instead of z I am going to choose t.

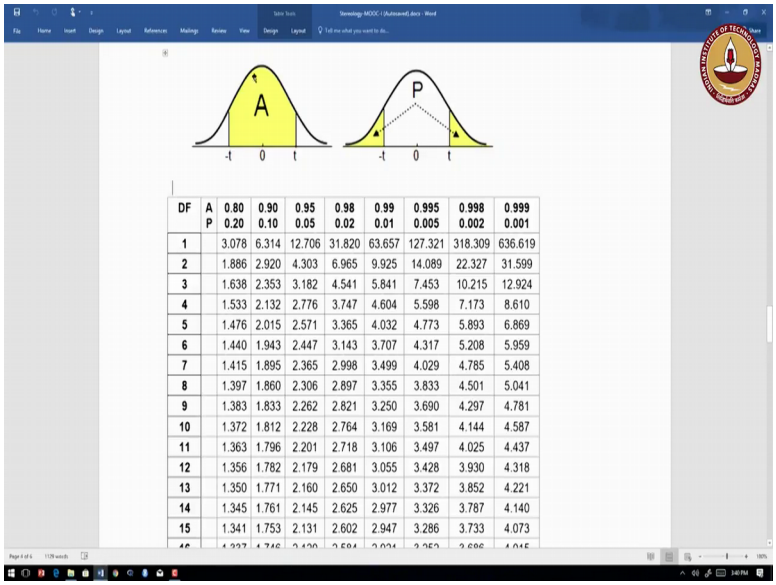
(Refer Slide Time: 27:30)



So, again there would be this is 0 this would be plus t and this would be minus t and you will select a value of t depending on the desired confidence level. In order to get the t values instead of z values if your n is small it is advisable to use the students t distribution then one can go to the t distribution table. Now, this is the students t distribution and let us try to understand this what this students t distribution mean here.

So, one you will notice that there is a there is something called as a degrees of freedom. So, this column represents degree of freedom and what is the degree of freedom in a particular in our particular case, where I have collected one set of sample x 1 to x n the degree of freedom is D F is equal to n minus 1. So, you can read off degrees of freedom from the on the left column.

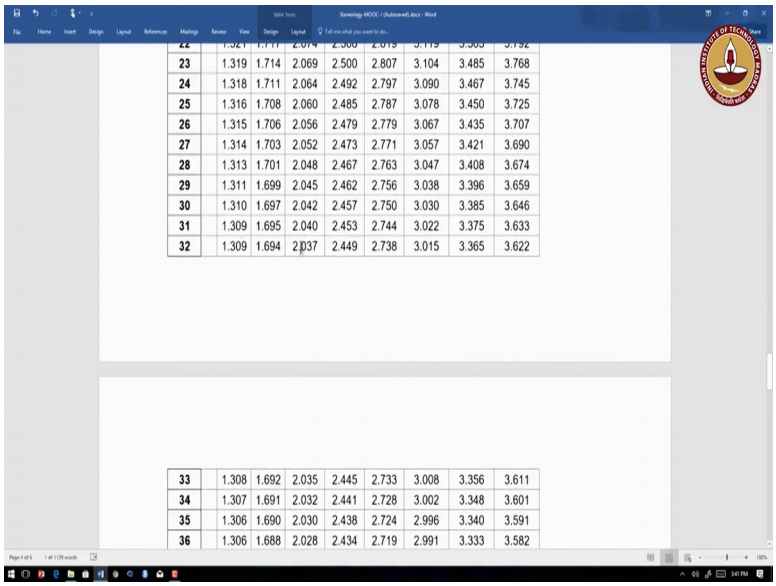
(Refer Slide Time: 29:08)



DF	AP	0.80 0.20	0.90 0.10	0.95 0.05	0.98 0.02	0.99 0.01	0.995 0.005	0.998 0.002	0.999 0.001
1		3.078	6.314	12.706	31.820	63.657	127.321	318.309	636.619
2		1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3		1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4		1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5		1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6		1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7		1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8		1.397	1.860	2.306	2.897	3.355	3.833	4.501	5.041
9		1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10		1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11		1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12		1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13		1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14		1.345	1.761	2.145	2.625	2.977	3.326	3.787	4.140
15		1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073

Now, I want to get this area a let us say I want to get it to 95 percent. So, I have to find the t value area A this table represent the top row and the in the top row the topmost number. So, this represents 0.95 represents 95 percent, in this column down these are all t values for different degrees of freedom. So, if you had only one degree of freedom; that means, you are only 2 sample points with one simple point you cannot do anything if you had only 2 sample points you can see that t value is 12.7 instead of 1.96 your error bar is going to be very large for 2 data points and as the degree of freedom is increasing that mean your sample size is increasing your t value you will find is reducing and if you go down this column of 0.95 I will keep the cursor on that column.

(Refer Slide Time: 30:11)



23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.090	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	1.309	1.695	2.040	2.453	2.744	3.022	3.375	3.633
32	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	1.306	1.688	2.028	2.434	2.719	2.991	3.333	3.582

So, that we know this is the column we are referring to you see the t value is reducing it has become 32 degrees of freedom that with 33 data points I have a t value of 2.037 it is coming closer to 1.96 and let us go down further at 80 you get 1.99 at 500 you get 1.96.

So, it is moving towards the normal distribution and at infinity it is exactly 1.96 for 95 percent. So, going back so, actually if you think back even to is pretty close to 1.96 you are not going to find too much of a difference going by this. If you have data of the order of 40 sample points or 40 degrees of freedom I have a t value of 2.02 and you may find that there is not much of a difference whether I use 1.96 or 2.02. So, maybe we can say up to 40 the sample points you use a you use a normal t distribution beyond that you are to use a value of 1.96 assuming normality.

If there are any questions you would like to ask at this point because this is basically in a nutshell what was confidence interval and how to calculate the confidence interval. So, then in the subsequent lecture now we will see with the data that we collected in a previous lecture how to get you know now those error bar in your actual data points so, with this I will close this lecture.

Thank you.