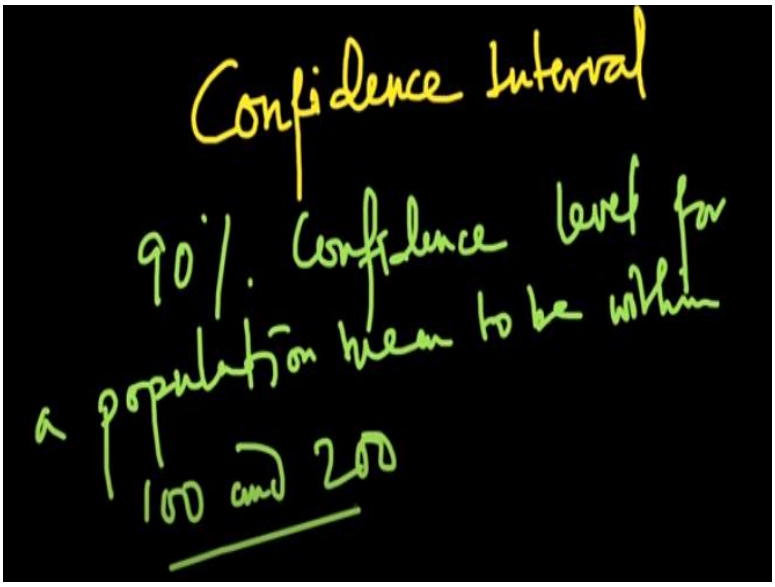
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology-Kharagpur

Lecture - 37
Confidence Interval

Hello and welcome back to the lectures on Applied Econometrics.

(Refer Slide Time: 00:30)



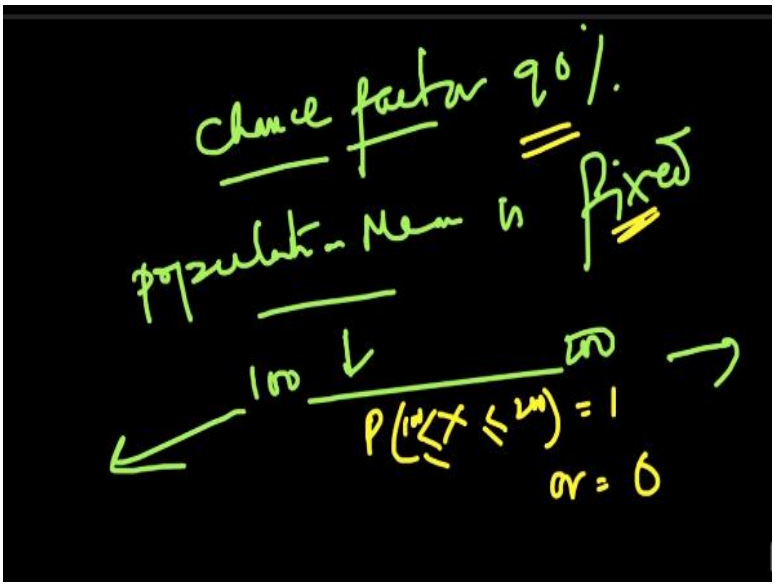
We have been talking about decision making when it comes to hypothesis testing. And we talked about several tools that we use when it comes to decision making or during the hypothesis testing. And those tools or basically the concepts that we use are level of significance, confidence level, P value, power, you know the properties of the distribution and something called confidence interval.

And we will see how we actually derive a confidence interval from you know the other tools like confidence level, okay. So let us begin with that. Now when I say confidence interval, there are something we have to understand and basically understand what it is not, alright. So let me explain that. So when I say confidence level, so basically confidence interval and confidence level are pretty close concepts.

And I will just explain that in a while. Now let us say I say that a 90% confidence level, okay. A 90% confidence level, what it means is that let us say the statement goes like that, a 95% confidence level for a population mean to be within 100 and 200, okay. So that is the confidence level attached to this statement that my population mean is within 100 and 200. And I have a confidence level of 90%, okay.

Now what it does not mean? What it does not mean is that it does not mean 90% chance that the population mean falls between 100 and 200.

(Refer Slide Time: 02:31)



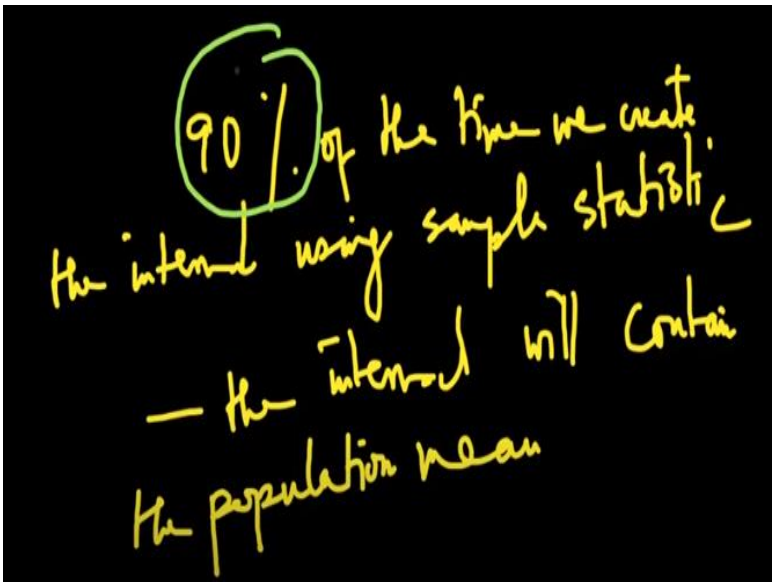
It does not mean that 90% chance, the chance factor, it does not mean there is any chance factor of 90% for the population mean to be within 100 and 200. And the reason is population mean is always fixed, right? Population mean is always fixed, is fixed right, is fixed. So either there are only two possibilities. Because it is fixed there are two possibilities.

Either it can be in the range of 100 to 200. So either it is here or it is not here, or it is somewhere outside. So if it is in here, so I will have a

probability of population mean to be let us say x to be within 100 and 200 is equal to 1 here or is equal to 0, okay. So only two possibilities. It is either 1 or 0. Because there cannot be a chance factor, there cannot be a chance factor attached to something which is fixed, right?

It is already fixed. So that is something we have to remember. So then if it is not involving any chance factor, so what it is really involving?

(Refer Slide Time: 03:56)

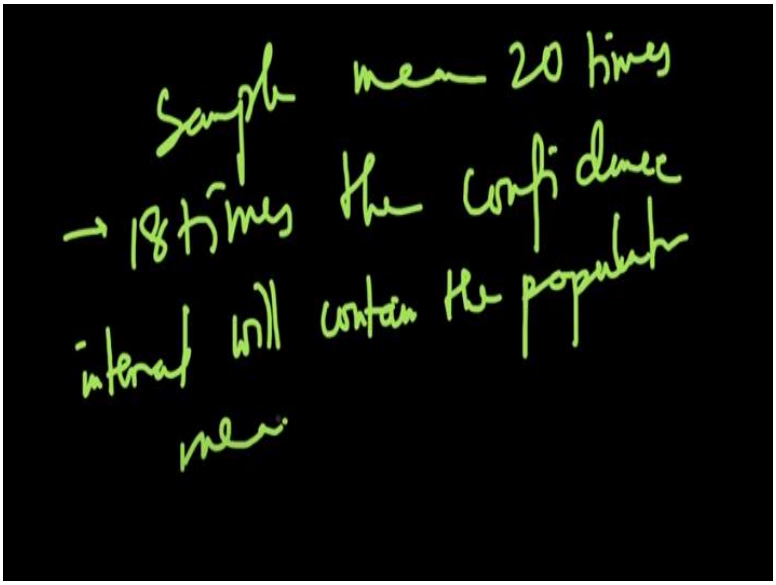


So it is basically how we really understand it is that we say that 90%, so we said 90% right, 90%. So we say that 90% of the time we create the interval using sample statistic. Here it is a mean. It could be a proportion, anything, using sample statistic. The interval will contain the population mean.

The interval will contain the population mean, okay. So basically you have to understand what is varying here is the sample statistic. So every time you get a sample statistic you have certain technique to construct the interval. And when you construct the interval, when I say 95%

confidence level, that means 90% of the time you construct the interval, you will have the population mean within that interval.

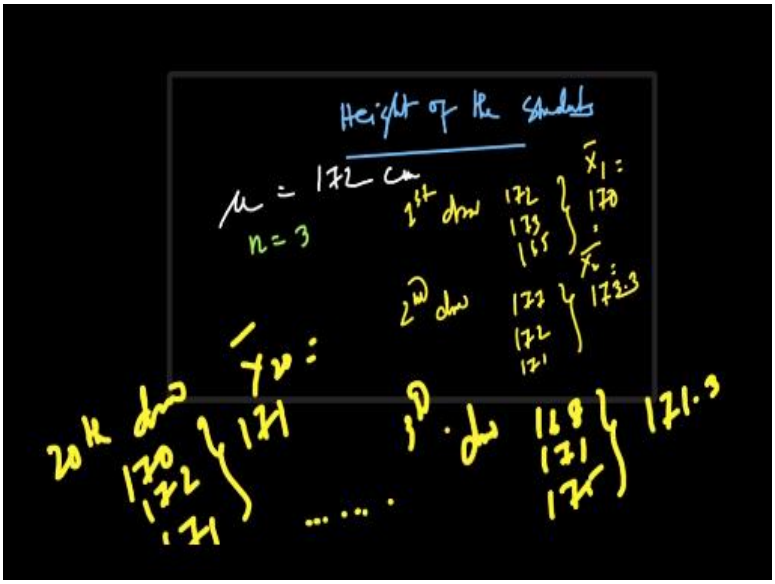
(Refer Slide Time: 05:21)



So let us say I create, I take sample mean 20 times. So what it would mean if my confidence level is 90%? 18 out of this 20 time, 18 times the confidence interval will contain the population mean or population statistic, whatever that is, population mean here, okay. And I will illustrate that with a diagram in a while.

Now, so I hope you understand it. So especially if we say that it is 95% confidence level, so then there will be 19 times out of the 20 times you create that interval that will contain the population parameter, okay. Okay, great. So now let us try to actually illustrate that with an example. And let us say we are, you know sort of getting the height of the students. Let us take an example, let us say height of the students.

(Refer Slide Time: 06:34)



And we want to basically get the confidence interval height of the students. I will basically pick samples and out of using the sample, we will create the confidence interval, okay. And whatever we have just discussed we will just demonstrate that. Let us say I have a population parameter, let us say I know about this population parameter. And I usually denote the population parameter with a mu, let us say it is a mean.

So mu is let us say 172 centimeter. So that is the height of the students, let us say in a school or whatever classroom or you know in a village in a town, whatever that is, let us say in a town, okay? Let us say this is age, maybe 12 to 14 in a given town. I am getting the mean height of the boys or boys and girls. So I get that, let us say that is 172 centimeter.

We usually we do not know the population parameter but let say we know it okay here. Now what I do is I actually, I cannot really go to every household and actually measure everybody's height. So what I do is I basically, you know take three samples, three individuals I take, like n = 3 my sample size, and I take three individuals and I get the mean, okay.

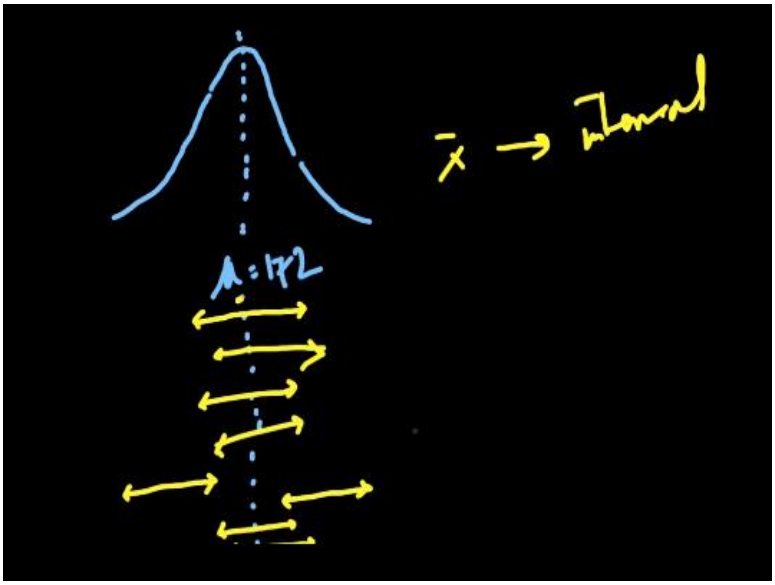
So let us say the first draw, first three candidates I get, I will actually write down. So let us say first draw, I will use red color, first draw. What are the different observations I have? Let us say I have 172, 173 and 165. And if I take average, I think I will get the mean is equal to 170, okay. So that is basically my x bar. So that is the different observations I am getting.

So let us first draw, so x bar 1, and that is equal to 170, okay. Now I have a second draw. And I see that, you know three observations I get 177, let us say 172 and then I have 171. And I think if I take a mean, it is going to give me a value of 173.3, right? So basically 170 and then 7 plus 3, 2 plus 1, 10, and 10 by 3 is 3.3. So that is my x 2 bar. I get this. I will reduce the size of the window.

So third draw, I will go for a third draw. And in this third draw, let us say I have 168. Then I have 171, then I have 175, let us say. And if I take a mean of all this, I will get I guess this is going to be 171.3 I think, 171.3, okay. And this way, let us say I draw it 20 times, okay. And let us say that in the 20th draw my observation is 170, 172 and 171, all right?

And the average is going to be 170 let us say 170, is going to be 171 right, 171. So x, x 20 bar is going to be 171. So I got all the different sample means right, and I will construct the confidence interval using the sample means. And let us do that. Let us do that.

(Refer Slide Time: 10:42)



So let me now actually draw a distribution where I already have, let us say this is the original distribution of all the samples. And let us say this is a true population parameter mu is equal to 172, okay. Now when I create the confidence interval, how I really do that? How I do that is, let us say every time when I draw this different samples, so let us go back to the numbers here.

So the first draw, I got 170, okay. I got in the first draw, I got 170. In the second draw, I got 173.3. In the third draw, I got 171.3. In the fourth draw, I got 171, sorry, it is the 20th draw. But other draws also we got several values. So when I say it is a 90% confidence interval, so what happens here is, 90% confidence level, so what happens is, so in the first draw, I got 170 right?

So I create a confidence interval. So there is a formula, right? There is a formula using x bar, you can actually have an interval, and we are just going to see the formula in a while. So what I create, I create an interval. Let us say this is the interval I got. So essentially, I will find that this mu

too belong in this confidence interval. Second time also you get some x bar, and you create some interval, let us say it is like this.

But you know the interval might, you know like, change its location in the sense that whatever is your x bar, but end of the day, that interval is actually containing the value mu, 172 is within this interval. Similarly third draw, fourth draw, and when it is 90%, so I can say actually, what is happening here is that out of this 20 draws, 18 times I am actually getting the mu within this interval.

So there might be one case where my interval is falling here, there might be another case where my interval is falling here, it does not mean like it has to be in the right or left. It could be in the same side or whatever. But you know what we actually mean is that out of the 20 times 18 times my interval is containing the mu, okay. So that is basically the idea of confidence interval.

And that is how it is related to the concept of confidence level, okay. All right. Okay. So here, we see now, now we will see how we really create this confidence interval. And essentially, you have to remember that we use the mean.

(Refer Slide Time: 13:34)

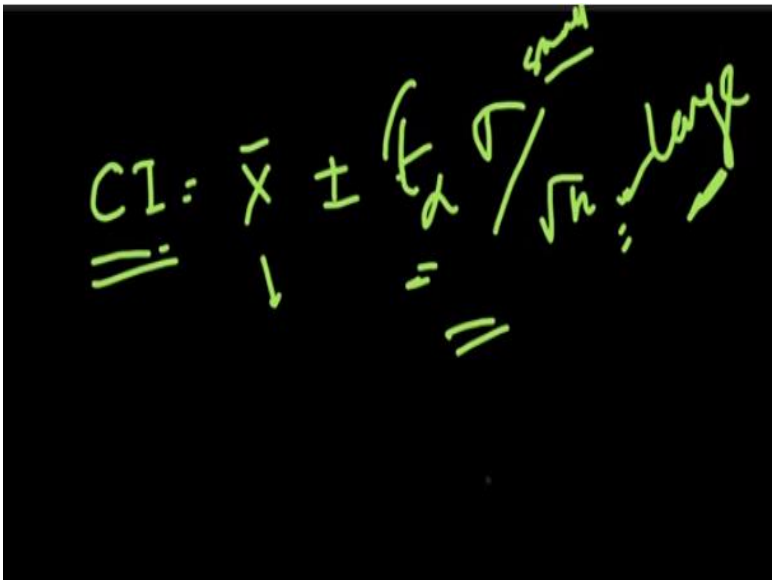
Let me use a different page here, sample mean. What you need is a sample mean, sample SD. We do not know the population standard deviation. So we use the sample SD to basically approximate for the population density, okay. And using this we basically create the, and we of course need to have a confidence level. We need to have the confidence level.

So which will give us the kind of margin of error you want. So let me just explain that. So sample mean is essentially any, it could be any other statistic. So basically the statistic, we need one statistic, okay. Then we need the confidence level. This is, this will be given. And the sample SD will help us to understand the margin of the error okay, margin of error.

So essentially margin of error would be derived by this two, okay. So we need the sample statistic and then we need a margin of error, which will be derived from sample SD and the confidence level, basically from sample SD to population SD. And this will help us to get the margin of error, okay? Population SD and the confidence level together will give us the margin of error. All right.

So the formula here is that confidence interval is equal to whatever statistic we have plus minus the margin of error okay, margin of error. Now usually we calculate the margin of error with the help of basically our confidence level and the standard deviation of the population. And since we do not know the standard deviation of the population, we approximate it from the sample.

(Refer Slide Time: 15:29)



So the formula that we get is, let us say confidence interval for the population parameter, population mean, we will have let us say t alpha, for alpha level of significance, and this is the sigma by root n, okay. So we basically use the sample statistic s and we sort of use that as sigma and a correction factor of root n, okay? Okay, all right.

Now one thing we have to see, oftentime we use a t statistic instead of a z statistic, even though we actually assume the population to be, the distribution to be normal. The reason of, reason behind this is when we deal with the unknown population parameter, we use the t statistic instead

of the z statistic; z statistic is something that we use when we have known population parameter, okay? Okay.

Now what we can infer from this formula is that if I have a high standard deviation, my confidence interval is going to be wide, right? If I have a high standard deviation, my confidence interval is going to be wide. Whereas, if I have a high n sample size, my confidence interval is going to be narrow right? Now we will prefer a confidence interval to be narrow.

Because the confidence interval is narrow means within that I will be more sort of, you know I will be able to say that within that narrow range, my population parameter belongs, I am kind of more certain where my population parameter is. But if it is very wide, then it could be anywhere, right? It could be anywhere. So that is something, it is not very helpful. But if I have a narrow confidence interval, then it is actually helpful.

And that is the reason why we need, I mean we usually prefer a small sigma, we usually prefer a small sigma and a large a, right, the sample size is large is really good and of low standard deviation is usually good, right? Okay. So how we construct the sample the confidence interval? So by now we understand that we have to choose a sample statistic beta mean or beta proportion.

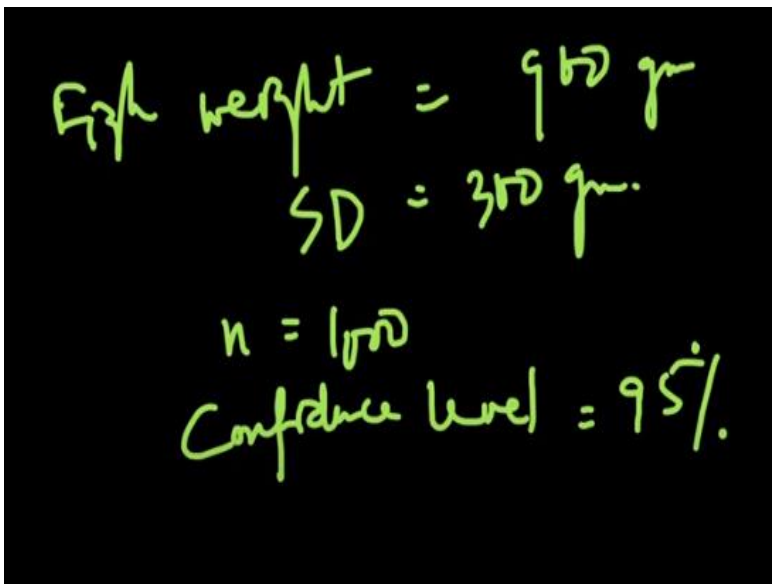
And then we have to choose the confidence level. So depending on the confidence level, I get the t statistic, and I know how the confidence level and how to see the confidence basically t statistic corresponding to a confidence level and we are going to see that in a while when you do an

example. And then we basically get the margin of error. And finally, we actually obtained the confidence interval, okay.

Now let us do one example to basically see how we really construct that, okay. So let us say the problem we have in hand is. Let us say we have a sample, let us say we are we have like, you might be knowing the hilsa fish in, you know Padma river or, you know the in Ganga river also, I think mostly in Padma river. Let us say it is Padma river. Padma is famous for hilsa fish.

And we are interested to know, we are interested to know the mean weight of the hilsa in this year, okay. During monsoon season, people just go for you know sort of fishing and hilsa is really a high value commercial, high commercial value and you know a lot of people actually end up fishing like small hilsa fish and so forth. So we are interested to know the mean weight of hilsa fish in Padma river this year, okay.

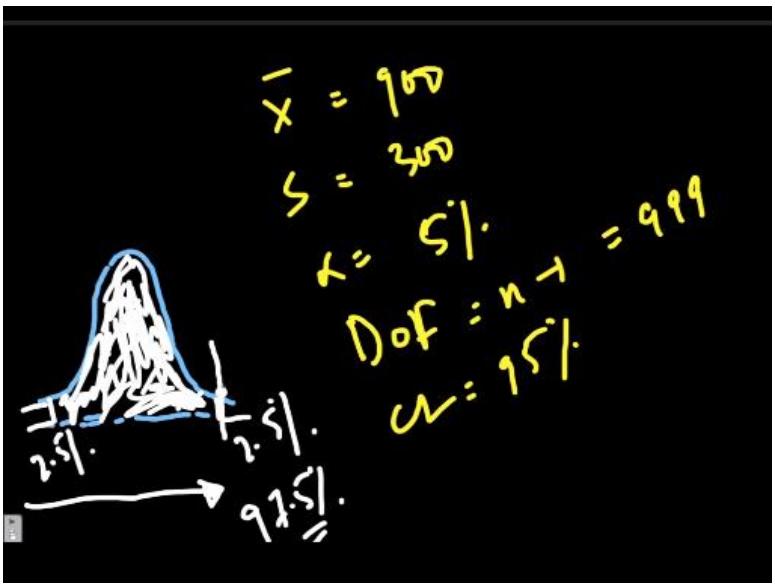
(Refer Slide Time: 19:12)



Let us say we have fish weight, let us say we have collected some samples, and we get a fish, mean fish weight is 900 gram, okay. And the standard deviation for the fish, you know all the different figures we got is 300 gram all right? Now and let us say we have actually say you know we got around 1000, fish, 1000 hilsa. And we got this, we got this you know this different values from mean and standard deviation.

And let us say we have a confidence level given to us is, let us say 95%, okay. Now we have to create a confidence interval for the, for basically obtaining the weight of the basically hilsa population in Padma river, okay. So if I write down in all this different information in terms of notation.

(Refer Slide Time: 20:14)



So I write x bar is equal to 900, s is 300, alpha is equal to, this is the level of significance, which is one minus confidence level, is 5%. And for because we know that we are going to use a t statistic, we really do not know the population parameter. So I have to also have the degrees of freedom. And if you remember t statistic, the degrees of freedom is going to be n - 1, which is 999, right? Okay, great.

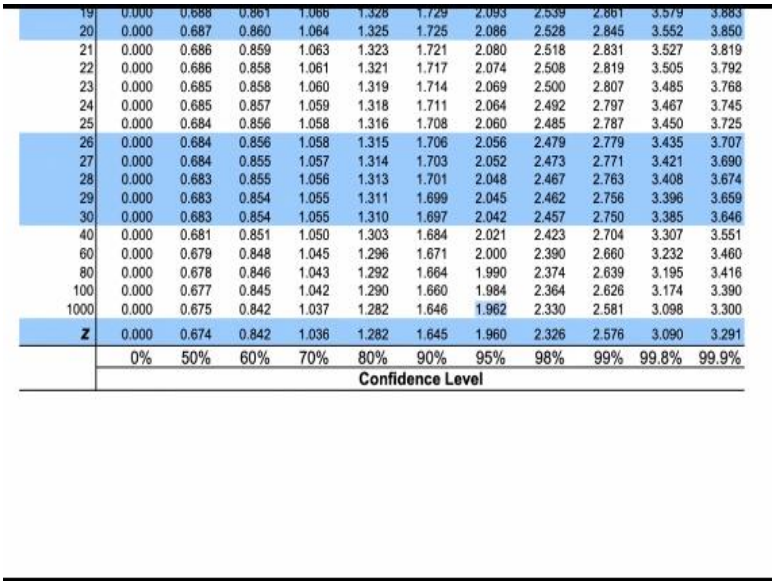
So and of course, we are given the confidence level CL is 95%. Now let me actually try to get, try to see the let us, I have too many windows open here. But let us see. Let us see. We want to get the t table. Let us get a t table, okay; t table let us say t table PDF. We get a PDF here; t table, I got a t table here. So I, we know how to see a t table. We have to see the degrees of freedom.

Along with that we have to see the value of the t statistic. And here since I have, this one you have to remember. So this is basically if we consider a two-tailed test, the 95% what is given here. So essentially, if I draw a distribution, it will look like this. Let me try it again. Okay. Something near normal let us say.

So if I draw it, I will have so 95% is this region right, this region under this curve. And since it is symmetrical one, the 5% will be spread across both these sides and this is going to be, let me actually make it more prominent. So I fill this and this part is going to be 2.5%. This part here is going to be 2.5%. Now when I see the t statistic, I will see the t table. So I will see that they have given a cumulative probability.

So cumulative probability for 95% confidence interval, I will basically consider only up to this and that will be from here, right? So that is going to be, that means cumulative probability is going to be for 97.5%, right? Because I am talking about the cumulative probability now. So it is only up to here, the probability value will be taken up to here right, which is 97.5%.

(Refer Slide Time: 23:27)



19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

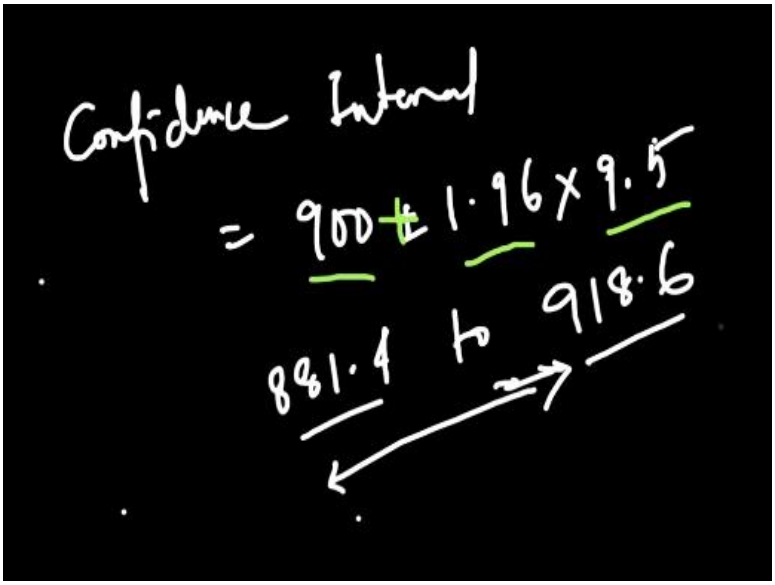
Now let us look at the t table and we need to look at the t table for cumulative probability now for t statistic is equal to t .975. And we know how to see a t table. So let me just actually go to that. So in case of t table, we need to basically look into the cumulative probability and the degrees of freedom, okay. And the degrees of freedom in this case is going to be as we have seen is a 999.

Now we can look at the df, which is on this row and all this df values we see it is from 1, 2 and then it has gone up to 100 and then suddenly it has jumped to 1000. Now since our df is 999, it is close to 1000. So we can basically approximately take the value corresponding to the df of 1000. And the value is going to be, now the value is going to be, so this one. Let me just again, so t .975.

See if I just go down here 4000 it is going to be 1.962. You can approximately say less than 1.96, okay, and let us just go back to our calculation table. Let us go back to our calculation table. And here, and we see that for t stat t is of 0.196. We have we let me get those you know

sigma by root n value which is 300 by 100, square root of 100 which is 30 by root 10. And which is essentially around 9.49 or 9.5, okay.

(Refer Slide Time: 25:16)



Now you take this and then you do the rest of the calculation. So you basically use a formula. You have your x bar, which is 900 gram for the average weight and then you substitute the value and then you have the t statistic is 1.96 and then sigma by root n is 9.5. Now if you just do a plus and minus here, so you basically get the interval and that is going to be 881.4 to 918.6.

So that is how you basically get the confidence interval. So with this we will end this lecture here and in the next lecture, we are going to see couple of more examples of confidence interval.