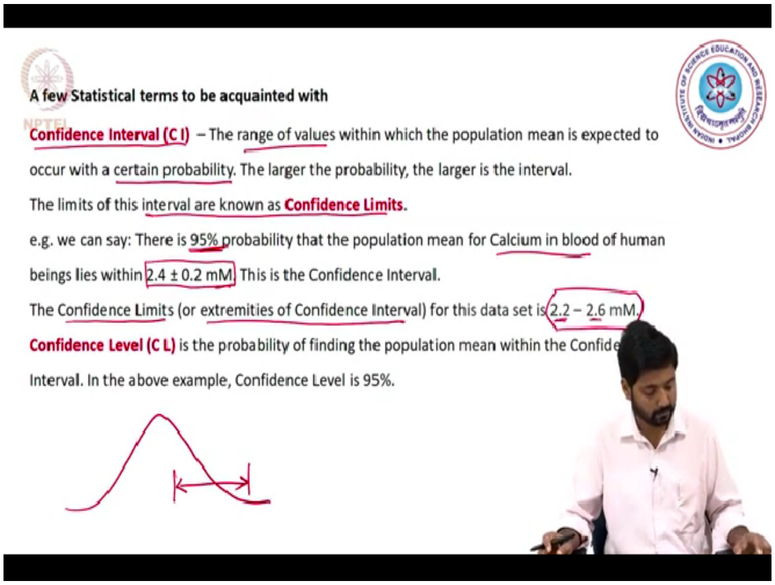
Quantitative Methods in Chemistry

Prof. Dr. Aasheesh Srivastava, Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research-Bhopal

Lecture-14
Introductory Statistics Part 02

(Refer Slide Time: 00:28)



A few Statistical terms to be acquainted with

Confidence Interval (C I) – The range of values within which the population mean is expected to occur with a certain probability. The larger the probability, the larger is the interval.

The limits of this interval are known as **Confidence Limits.**

e.g. we can say: There is 95% probability that the population mean for Calcium in blood of human beings lies within 2.4 ± 0.2 mM. This is the Confidence Interval.

The Confidence Limits (or extremities of Confidence Interval) for this data set is 2.2 – 2.6 mM.

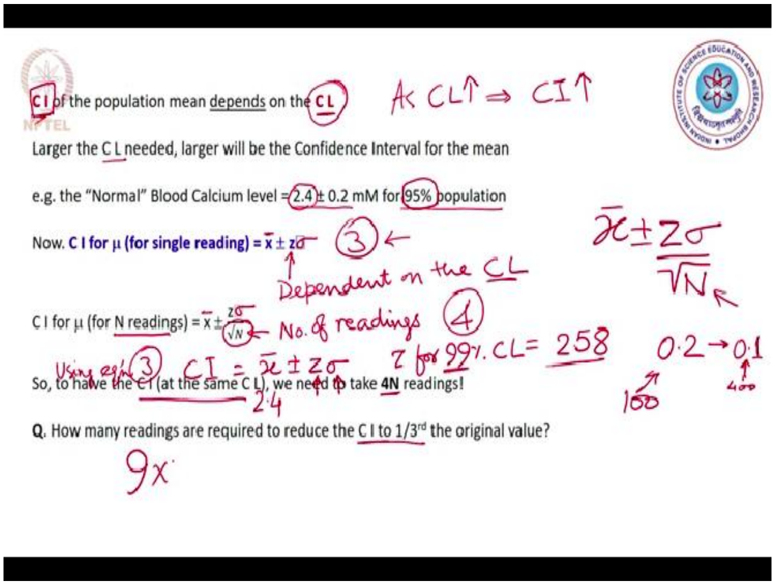
Confidence Level (C L) is the probability of finding the population mean within the Confide Interval. In the above example, Confidence Level is 95%.

Now a few statistical terms that we will be dealing with and which we would want to get acquainted with are the confidence interval. So, the confidence interval is the range of values within which the population mean is expected to occur with a certain probability. So, it is the probability of the population mean occurring within a range of values and these range of values is what we call as the confidence interval.

This will become clear in the example that I will take in a little while. So, please be here with me and the limits of this interval is known as the confidence limits. So, if I draw a Gaussian profile, I can find a confidence interval and the limits of this interval is what is known as the confidence limit to exemplify, let us see, or let us say that there is 95% probability that the population mean for say calcium in our blood lies in the range of 2.4 + - 0.2 millimolar.

Now, this is the confidence interval, and the confidence level here is 95%. Okay, so when we talk in terms of the confidence limits, which is the extremities of the confidence interval, the confidence limit for this data set is shown here, which is 2.2 to 2.6 millimolar. In other words, it is expected that 95% of the human population will have their calcium levels within this particular range of 2.2 to 2.6 millimolar.

(Refer Slide Time: 02:41)



CI of the population mean depends on the CL   As CL↑ ⇒ CI↑

Larger the C L needed, larger will be the Confidence Interval for the mean

e.g. the "Normal" Blood Calcium level = 2.4 ± 0.2 mM for 95% population

Now, **C I for μ (for single reading) = $\bar{x} \pm z\sigma$** ③ ←

Dependent on the CL

$$\frac{\bar{x} \pm z\sigma}{\sqrt{N}}$$

C I for μ (for N readings) = $\bar{x} \pm \frac{z\sigma}{\sqrt{N}}$ ← No. of readings ④

Using eqn ③ $CI = \bar{x} \pm z\sigma$ z for 99% CL = 258   0.2 → 0.1

2.4   100   400

So, to have the CI (at the same C L), we need to take **4N** readings!

Q. How many readings are required to reduce the C I to 1/3rd the original value?

9x

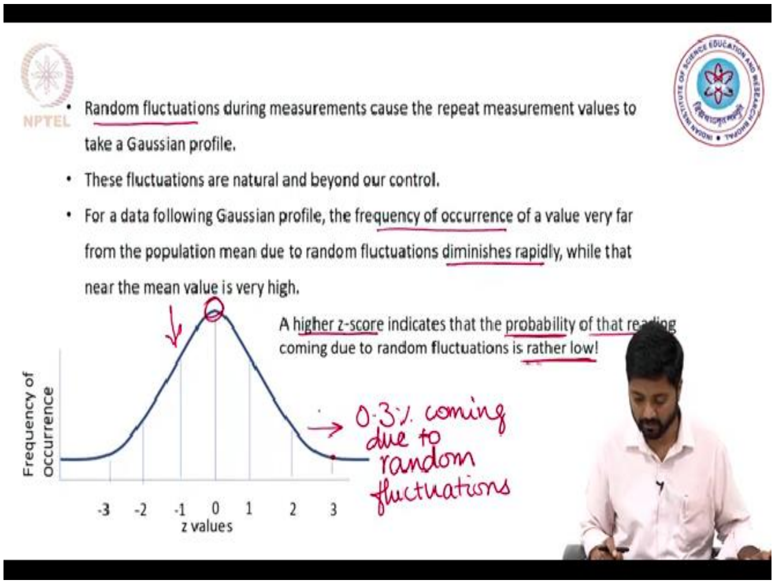
Now, as we can see that the confidence interval depends intricately on the confidence level that we want to report our data with. So, the confidence interval is abbreviated as CI and the confidence level, which is the probability of finding the mean is put in terms of the confidence level CL. So, if we want to for example, if what we said in the last slide was that for 95% population, the normal blood calcium level is 2.4 + - 0.2 millimolar.

Now, for single readings the confidence interval for mu for a single reading is given as the mean value of that reading and + - z and sigma, z is the z score or z statistics which we have introduced already and this will be dependent upon the confidence level of our reporting. Now, if the measurement is done for n multiple readings, then the confidence interval can be reported as x bar which is the mean value of these readings + - z sigma by root n.

Where n denotes the number of readings being undertaken okay. Now, let us consider the previous example, where we said that the 95% population will have the normal blood calcium level of 2.4 + - 0.2millimolar. Now, using these equations, let us call these equations 3 and 4 how do we calculate what would be the confidence interval for 99% of the population. So, for that we will use equation 3 what we observe is that CI = x + - z sigma okay x bar.

And x bar here is 2.4 and z value and sigma value is what we need to plugin here. Now, z value for 99% of the population can be found here.

(Refer Slide Time: 05:48)

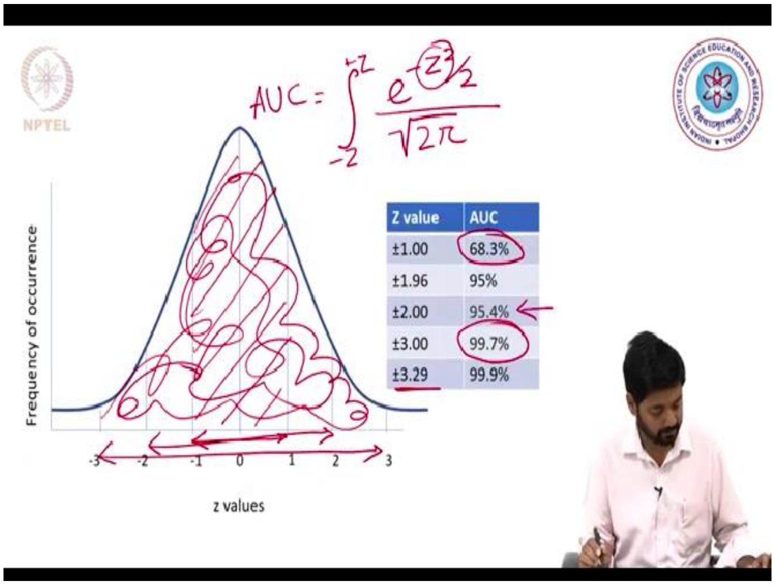


- Random fluctuations during measurements cause the repeat measurement values to take a Gaussian profile.
- These fluctuations are natural and beyond our control.
- For a data following Gaussian profile, the frequency of occurrence of a value very far from the population mean due to random fluctuations diminishes rapidly, while that near the mean value is very high.

A higher z-score indicates that the probability of that reading coming due to random fluctuations is rather low!

0.3% coming due to random fluctuations

(Refer Slide Time: 05:49)

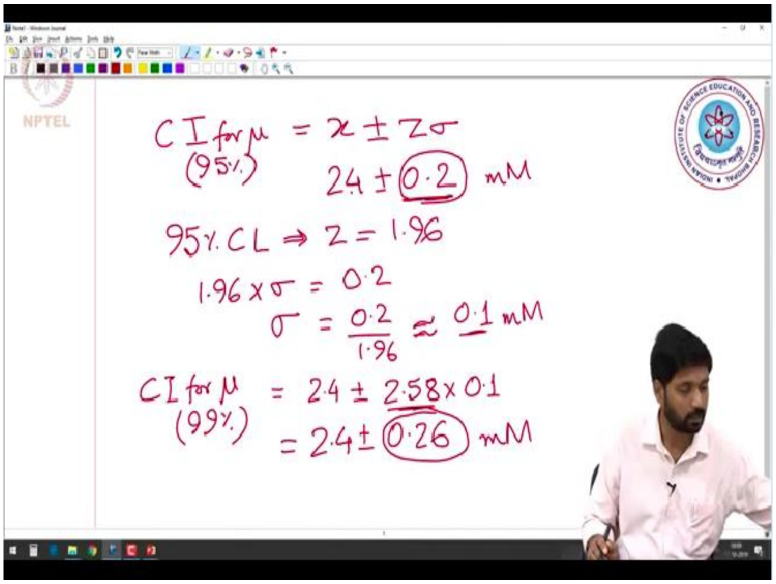


$$AUC = \int_{-z}^{+z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

Z value	AUC
±1.00	68.3%
±1.96	95%
±2.00	95.4%
±3.00	99.7%
±3.29	99.9%

And the z value for 99% population is 2.58, the confidence interval of the population mean depends intricately on the confidence level of our reporting. So, if we want to report with a higher confidence level, we would want to use a higher z value and that would imply that the confidence interval increases. So, as CL increases, this implies the CI also increases. So, which is also written below that for larger confidence level we will require a larger confidence interval for the mean.

(Refer Slide Time: 06:59)



$$\underset{(95\%)}{CI \text{ for } \mu} = x \pm z\sigma$$

$$2.4 \pm \boxed{0.2} \text{ mM}$$

$$95\% \ CL \Rightarrow z = 1.96$$

$$1.96 \times \sigma = 0.2$$

$$\sigma = \frac{0.2}{1.96} \approx \underline{0.1} \text{ mM}$$

$$\underset{(99\%)}{CI \text{ for } \mu} = 2.4 \pm \underline{2.58} \times 0.1$$

$$= 2.4 \pm \boxed{0.26} \text{ mM}$$

So, what we have seen is that the confidence interval for mu = to x = - z sigma, because as we know that the z can take both positive and negative values. So, this for our data set was 2.4 + - 0.2 millimolar right. So, for 95% confidence level, the z score is 1.96. So plugging in this value, what we get is 1.96 into sigma is equal to 0.2 this value here, the right hand side. Now, when we do the maths here what we get is that the sigma value turns out to be 0.2 by 1.96 which is equivalent approximately 2.1 millimolar.

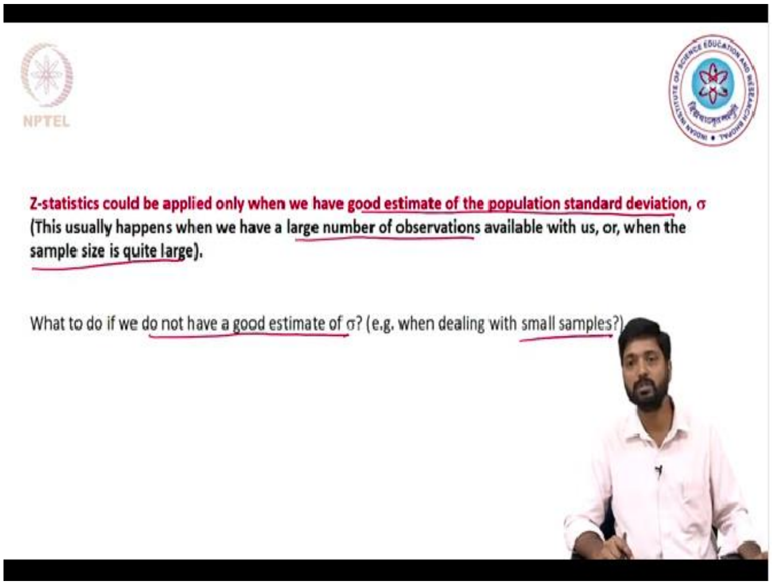
So, 0.1 millimolar is the standard deviation value for this data set. Now, when we plugin this value, so, this all result was at 95% confidence level, if we want to increase the confidence interval for mu to 99% we will have this value as 2.4 + - 2.58 which is the z value for 99% confidence level into 0.1 which is the standard deviation for this data which we just calculated here. Now when we do that, we get the confidence interval for mu at 99% confidence level as being 2.4 + - 0.26 by rounding it off.

So, what we already see is that the at 95% confidence level the spread was 0.2 millimolar at 99% confidence level, this confidence interval need to be increased or expanded 0.26 millimolar values. Now, let us go back to our presentation. So, what we observe is that if we are doing our measurements n times or if we are repeating our measurement n times the confidence interval can be written as x bar + - z sigma by root n.

So, this equation is very similar to equation number 3 which is shown here, except that it incorporates the root n term which is coming from the n repeat measurements that we have performed. So, what this equation also tells us is that the confidence interval at the same confidence level can be half if we take 4 times the readings. So, for example, if our readings work or if our confidence interval was 0.2 in the previous case based on say 100 readings.

If we take 400 readings, this can be reduced to 0.1 millimolar by taking 400 ratings okay. So a question can be that if you want to reduce the confidence interval to one third of the original value, how many readings do we require and yes the answer is that we will require 9 times the original readings to get reduced confidence interval which has been reduced to one third original value.

(Refer Slide Time: 11:50)



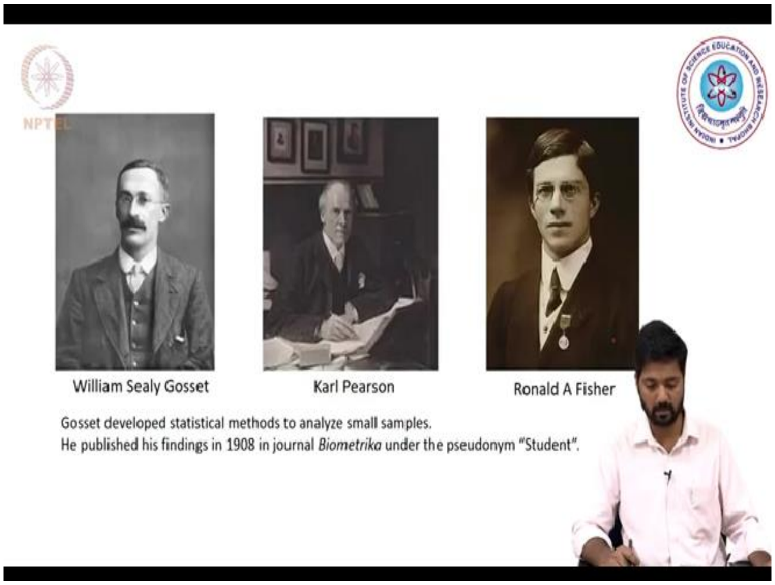
Z-statistics could be applied only when we have good estimate of the population standard deviation, σ
(This usually happens when we have a large number of observations available with us, or, when the sample size is quite large).

What to do if we do not have a good estimate of σ? (e.g. when dealing with small samples?)

Now, it is very important to keep in mind that the z statistics can be applied only when we have very good estimate of the pollution standard deviation sigma. And when do we have that we have that when a really large number of observations are available to us. So, we have a very good estimate of how a population is behaving or in other words, when the sample size is quite large, usually that is not the case.

So, this question baffled initial statisticians in terms of what do we do if we do not have a good estimate of sigma or to rephrase it, what do we do when we are dealing with small sample sizes.

(Refer Slide Time: 12:50)



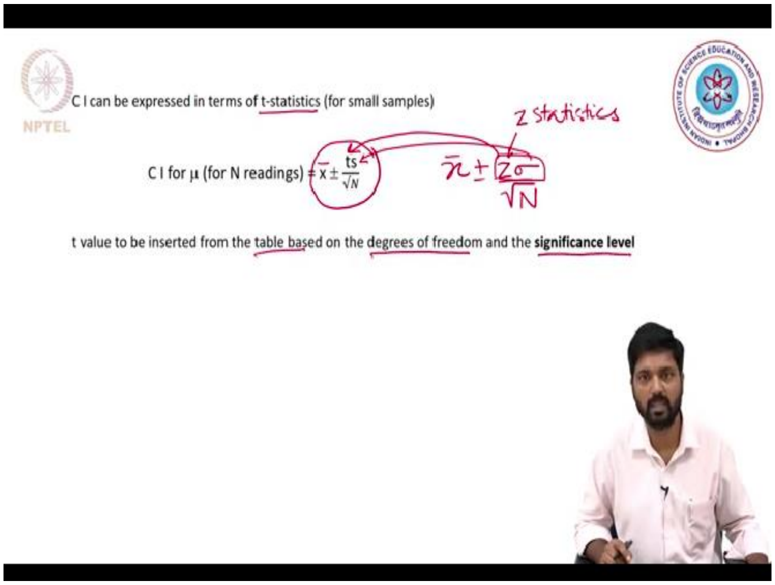
Gosset developed statistical methods to analyze small samples.
He published his findings in 1908 in journal *Biometrika* under the pseudonym "Student".

Now, this question was very effectively answered by William Gosset, who is shown here from our Wikipedia image and he along with 2 others solvers of statistics Karl Pearson, he is the same person in whose name Pearson coefficient is given. And similarly Ronald A Fisher in whose name other statistical tables are available. So these three people laid the foundations of early statistics which saw very rapid inclusion, both in academics as well as an industry.

So, what William Gosset did was he developed statistical methods to analyze small samples. And there is an interesting anecdote to be told here. William Gosset when he was working on his analysis of small samples, he was working for a beer company which is known as the Guinness Yes, it is the same one, which gives the Guinness Book of World Records to us. So Guinness had put an embargo that the scientists working for it will not be able to publish their findings, if they have word beer in it, if they have word Guinness in it, or if they have their own surname.

So what William Gosse did was to sort of overcome this embargo he published his results with a student name student in a journal, which is known as Biometrika. So he analyzed the beer samples and then he came up with a statistical tool that allowed him to analyze small sample sets and make predictions on a larger population. So what William Gosse did was to come up with a different statistics which is known as the t-statistics.

(Refer Slide Time: 15:06)



C I can be expressed in terms of t-statistics (for small samples)

C I for μ (for N readings) = $\bar{x} \pm \frac{ts}{\sqrt{N}}$

z statistics

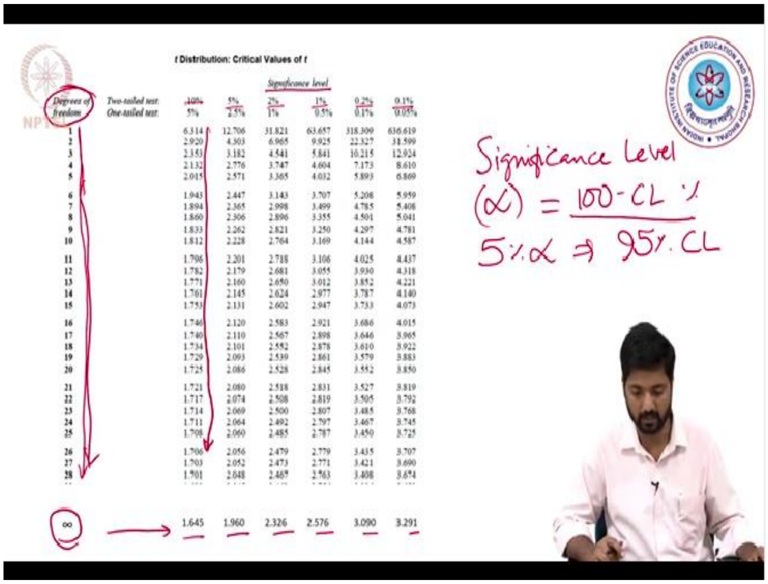
$\bar{x} \pm \frac{z\sigma}{\sqrt{N}}$

t value to be inserted from the table based on the degrees of freedom and the **significance level**

So, t-statistics is quite similar to the z statistics. So, if I look at the confidence interval, the same can be expressed very similarly, just to recollect it was z + x bar + - z sigma by root n and when we were dealing with the z statistics and when we are dealing with the t-statistics, this equation only changes that the top part to the z is changed to t and sigma z is move to t and the sigma is converted to s and the equation essentially remains same.

Now, the t value can be inserted from a table which is widely available in a variety of textbooks and it utilizes 2 terms one which is known as the degrees of freedom, which again has been already introduced and the other one is the significance level, we will understand this in a little while, but before that let us look at how a t table looks like and how we can insert values from the t table.

(Refer Slide Time: 16:37)



t Distribution: Critical Values of *t*

				Significance level			
Degrees of freedom	Two-tailed test:	~~10%~~	5%	2%	1%	0.2%	0.1%
	One-tailed test:	5%	2.5%	1%	0.5%	0.1%	0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
∞	→	1.645	1.960	2.326	2.576	3.090	3.291

Significance Level

$(\alpha) = \underline{100 - CL\ \%}$

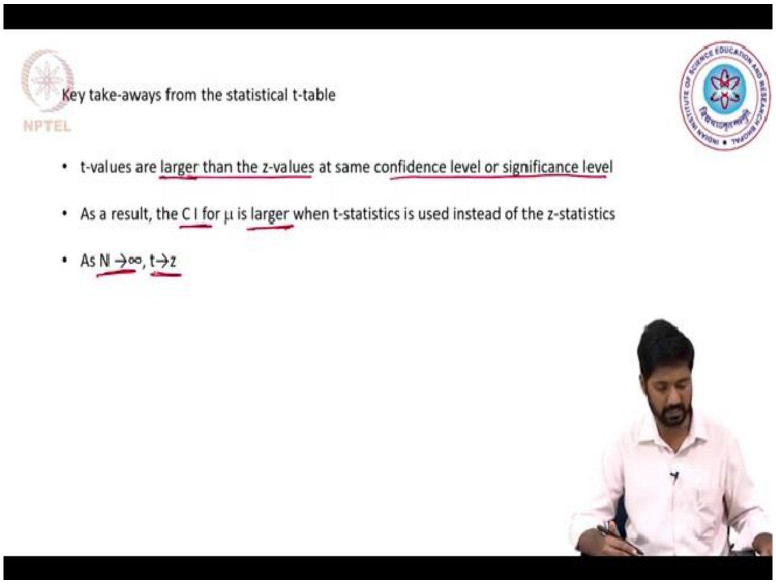
$5\%\,\alpha \Rightarrow 95\%\ CL$

So, a t table would look something like this, it has degrees of freedom and it has significance level different significance levels are given here. On top based on the fact that we are dealing with a two tail test, we will discuss that also in the next class. So be here with me for a little while, but what it gives us are based on the degrees of freedom for example, let us say here we have degrees of freedom 1 to 28.

What would be the t values at different significance level. Now significance level is denoted as alpha and is nothing but 100 minus the confidence level in percent. So, when we say 5% significance level, we are implying 95% confidence level for two tail test. So, what we see here is that our as our degrees of freedom increase the t value at a particular confidence a significance level also decreased considerably.

Okay and another thing to keep in mind is that at infinite degrees of freedom, our t value will merge into the z value. So the numbers written here in the last line are all the z values at a particular significance level. And the significance level I have already explained is usually 100 minus the confidence level.

(Refer Slide Time: 18:44)



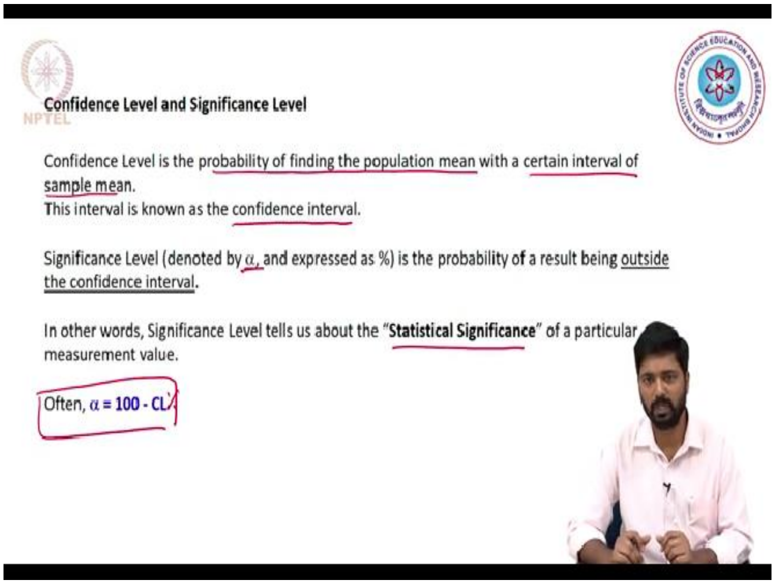
Key take-aways from the statistical t-table

- t-values are larger than the z-values at same confidence level or significance level
- As a result, the C I for μ is larger when t-statistics is used instead of the z-statistics
- As N →∞, t→z

So, the key takeaways from the statistical t table are that the t values are larger than the z values at the same confidence level or significance level and as a result, when we calculate the confidence interval using the t-statistics, that confidence interval would be larger or more spread out compared to the same confidence interval calculated using the z statistics. And another important thing that we saw just a little while ago is that as the number of measurements increase to infinity, the t value merges with the z value or t-statistics merges with the z statistics.

So, in other terms, what this table tells us that our predictability about the population increases as the number of readings are increased and a limiting case of t-statistics at infinite reading is the z statistics.

(Refer Slide Time: 20:03)



Confidence Level and Significance Level

Confidence Level is the probability of finding the population mean with a certain interval of sample mean.
This interval is known as the confidence interval.

Significance Level (denoted by α, and expressed as %) is the probability of a result being outside the confidence interval.

In other words, Significance Level tells us about the "**Statistical Significance**" of a particular measurement value.

Often, α = **100 - CL**

So coming back to the confidence level and significance level, it is again important to emphasize that confidence level is the probability of finding the population mean within a certain interval of the sample mean. So, we can take a small sample from the world population, the sample mean will be different from the population mean, but if we analyze it properly, we will be able to make predictions about the population mean based on what the sample mean is and at what confidence level do we want to make that prediction.

Now, the interval within which the population mean is supposed to lie is known as the confidence interval. The significance level which is denoted alpha is expressed as percentage and it is the probability of the result being outside the confidence level. Now, in other words the significance level tells us about how significant or statistically significant that measurement value is.

And as has already been told that the significance level alpha for a two tail test is 100 minus the confidence level in terms of percentage. So, in the next class, what we will do is, we will utilize some of these concepts to start calculating the confidence interval of the mean. And slowly we will get introduced to what is known by the one tailed and the two tail tests. And we will use these tests to test the hypothesis that we are projecting. So all that will be covered in the next class and thank you for being with us. Good day.