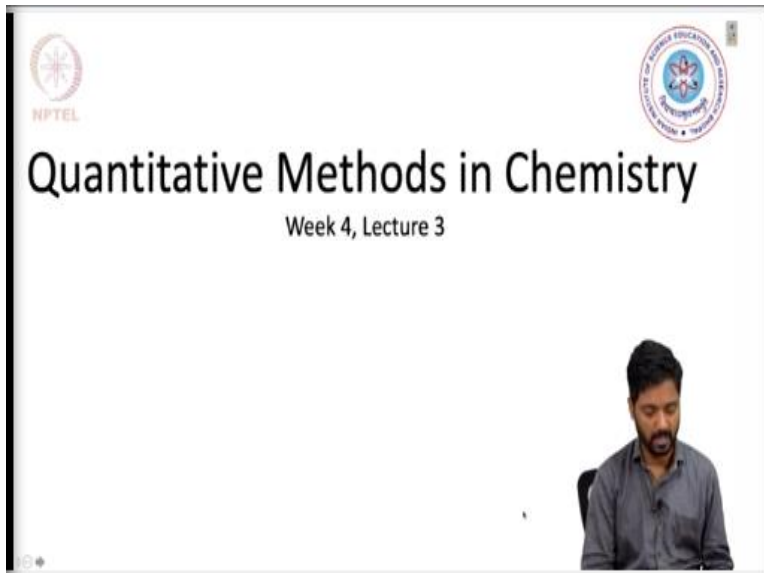


Quantitative Methods in Chemistry
Prof. Dr. Aasheesh Srivastava, Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research-Bhopal

Lecture-17
Pooling of data

(Refer Slide Time: 00:28)



Hello and welcome back to lecture 3 of week 4 of this course quantitative methods in chemistry. This week we have so far learned about how we can apply statistics to test various hypothesis that we need to test. And we get introduced to the concepts of Z test, T test, and also how outliers can be identified for a given dataset. So today we will get introduced to 2 important concepts related to this which is the pooling of the data to get a better estimate of the population. And also how we can apply a paired T test to compare 2 methods on a same sample set, so let us get started.

(Refer Slide Time: 01:20)

The concept of pooling data for enhancing the predictability
NPTEL

For samples collected from the same population can be pooled to improve reliability of predictions and obtain better estimates about the population.

Pooling all the four samples give a better estimate of the population And improve reliability of the standard deviation estimated thus.

So, let me first talk to you about the concept of pooling data to enhance predictability. So, this concept is illustrated by this Venn diagram. So, before that it is again very important to understand that this pooling of the data can be applied only for samples that are collected from the same population. And these samples which are denoted by different colors here for example, S1 denoted by blue and S2 denoted by red and so on and so forth.

They has been collected from this same population which is the blue circle outside and the different sizes of these S1, S2, S3 and S4 denote that they contain different number of datasets or different readings. So how do we pool all of this data to get a better estimate about the population. So this data set can be collected at different time points, it can be collected by different peoples.

So this makes it easier for us to collect a larger amount of data by distributing our resources. Now once we have collected this data, we can utilize pooling of the data to get a better estimate about the population by this pooling protocol. So what is this protocol is what we will learn now and you also need to understand that, by pooling this we improve the reliability of the standard deviation that we estimate through the pooling. So let us quickly go through the protocol of pooling such datasets.

(Refer Slide Time: 03:17)

Sample S_1, S_2, S_3, \dots (t)
 No. of readings N_1, N_2, N_3, \dots

$$S_{\text{pooled}} = \sqrt{\frac{\sum_{i=1}^{N_1} (x_i - \bar{x}_1)^2 + \sum_{j=1}^{N_2} (x_j - \bar{x}_2)^2 + \sum_{k=1}^{N_3} (x_k - \bar{x}_3)^2 + \dots}{N_1 + N_2 + N_3 + \dots - t}}$$

So if you are dealing with samples, which we call as S_1, S_2, S_3 so on and so forth. And we are dealing with say t such samples and in each sample, the number of readings is N_1, N_2, N_3 and so on and so forth, what we can do is, the protocol for pooling this data is that we can estimate. So our S_{pooled} will take into consideration the individual deviances which have been squared. Now these will be divided by the individual readings in each dataset or any each sample.

For example, N_1 for sample 1, N_2 for sample 2, N_3 for sample 3 so on and so forth. But it is important to remember that we need to have the degrees of freedom in the denominator. So since we were dealing with t number of sets, we will have to insert t here and subtract this t number of sets from the total number of readings in all the datasets. So the total number of readings in all the data sets will be given by $N_1 + N_2 + N_3 + N_4$ and so on and t will denote that number of samples that have been pooled. Let us quickly take an example with real numbers to understand this concept.

(Refer Slide Time: 06:00)

Sample I	Sample II	Sample III	(Dev1) ²	(Dev2) ²	(Dev3) ²
5.15	7.18	6.04	0.0009	0.0054	0.0058
5.03	7.17	6.02	0.0081	0.0040	0.0031
5.04	6.97	5.82	0.0064	0.0187	0.0207
5.18		6.06	0.0036		0.0092
5.20		5.88	0.0064		0.0071
\bar{x} 5.12	7.11	5.96	0.0254	0.0281	0.0459

$$S_{\text{pooled}} = \sqrt{\frac{0.0254 + 0.0281 + 0.0459}{5 + 3 + 5 - 3}}$$

So suppose we are dealing with sample 1, sample 2 and sample 3 and they have the readings as 5.15, 5.03, 5.04, 5.18 and 5.20 that means it has 5 readings in the sample 1. Sample 2 has only 3 readings which come out to be 7.18, 7.17, 6.97. Similarly, sample 3 again has 5 readings as 6.04, 6.02, 5.82, 6.06 and 5.88. Now these 3 samples have been obtained from the same population, so we need to pool this data and I will now explain to you how this is to be done.

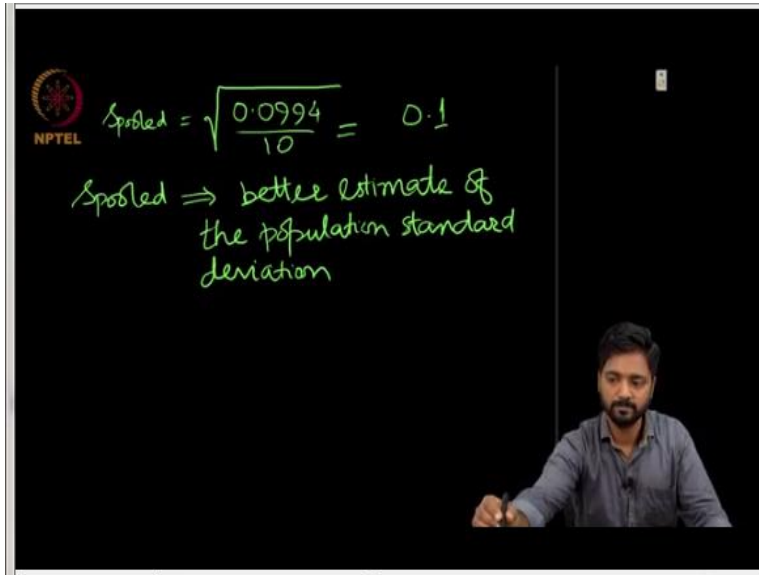
So if I take the average value of the samples, for the sample 1, this comes out to be 5.12, for sample 2 the average value of this sample is 7.11. Similarly for sample 3, this value comes to be 5.96. Now if I calculate the individual deviations, so for example, for sample 1 we talk about deviation 1, then these deviations can be squared for calculating the S pool value and these squared values will come out as.

So essentially these numbers are nothing but 5.12 which is the \bar{x}_1 value, which is subtracted from the individual readings in sample 1 and the result that we obtain has been squared. Similarly we can go about calculating the deviation 2 squared and that will come out to be 0.0054, 0.0040, 0.0187 and we can submit these. So this submission will come as 0.0254 while here for the second sample division squared summation will be 0.0281.

Similarly for the third sample that the deviation 3 squared value can be generated and let me quickly fill in the numbers here. So these are the numbers of the deviations is squared for sample

3 and when we sum all of these, this number comes out to be 0.0459. Now from these values, we calculate the S pooled as square root of $0.0254 + 0.0281 + 0.0459$ this divided by 5 readings for sample 1, 3 readings for sample 2, and again 5 readings for sample 3. However since we are dealing with 3 samples we will have to subtract 3 from here.

(Refer Slide Time: 11:40)



And as you go through this calculations you will obtain that s pooled value. Now comes out to be square root of 0.0994 divided by 10 which is ultimately going to be 0.1. So this S pooled value is better estimate of the population standard deviation. So I hope you would have understood how this pooling of the data is to be done and what is the protocol to be employed to obtain S pooled value or the standard deviation value for this pooled dataset taking into consideration the various readings in individual samples.

Now let us move to the second topic, which is of applying paired t-test to samples which are being analyzed by 2 different methods.

(Refer Slide Time: 13:30)

Paired t-test for comparing two different Methods applied on the same data set


Suppose two different methods are employed on the same sample, then the differences in their results can be compared using the paired t-test

Sample #	Method 1	Method 2
1	500	480
2	292	305
3	343	325
4	445	460
5	480	500
6	395	370

Note:

- Both Method 1 and Method 2 are being applied on the same sample.
- There can be difference between the samples

Q. At 95% confidence, can we say that Method 1 gives different results than Method 2?



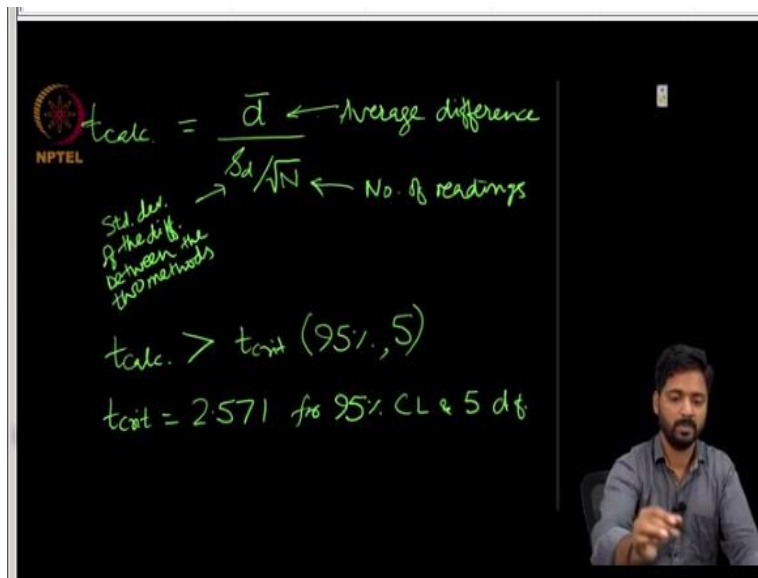
So for that let me go back to the presentation paired t-test for comparing 2 different methods, which are applied on the same dataset. Now again it is important to remember that a paired t-test can be applied only when the same dataset is being compared by 2 different methods. And one example of that is suppose there is a blood test that is done on a patient. So suppose method 1 shown here is the reference method.

And method 2 is a new method that is being developed to estimate the same analyte. So what we do here is, we take samples such as sample 1 or 2 or 6 and we use these samples to estimate the analyte concentration using method 1 or method 2. Now there can be differences between the 2 methods and also there can be differences between the individual samples. So the question that we want to answer here is, whether the 2 methods which are now for as the reference and the new method give as the same or similar results.

Or there are differences between the old method and the new method. So let us take this example which is shown on the slide, where we have these 6 samples and we have applied method 1 and method 2 on them. And generated the analyte concentration whose numbers are given in the table. So what I want you to note here again is that both method 1 and method 2 are being applied on the same sample and there can be differences between the individual samples. And the question that we would want to answer for example is that at 95% confidence can we say that method 1 gives different results than method 2.

So here you might recollect that since it is only about the estimation or prediction of differences between the 2 methods, a 2 tailed test can be applied. And we will also again have a null hypothesis that the 2 methods do not give different results or the results that they generate are statistically similar. And any minor variations between method 1 and method 2 is the result of random fluctuations during the measurement. So now, let us apply the paired t-test to compare these 2 methods, so let me again go back to the board.

(Refer Slide Time: 16:49)



The image shows a blackboard with handwritten notes in green and white. In the top left corner, there is a logo for NPTEL. The main text on the board is as follows:

$$t_{calc.} = \frac{\bar{d}}{s_d / \sqrt{N}}$$

Annotations for the formula:

- \bar{d} ← Average difference
- s_d / \sqrt{N} ← No. of readings
- s_d ← Std. dev. of the diff. between the two methods

Below the formula, the following text is written:

$$t_{calc.} > t_{crit} (95\%, 5)$$
$$t_{crit} = 2.571 \text{ for } 95\% \text{ CL \& } 5 \text{ df}$$

So again here please remember that we will apply the same protocol except that the t calculated value here will be d bar divided by s d by root N. Now d bar here is the average difference between the 2 methods, s d is the standard deviation of the differences between the 2 methods. And of course root N, the N is the number of readings being compared. So let us quickly apply this to estimate or predict whether the method 1 and method 2 in our example are generating different results or not.

So they will be generating different results if our t calculated is greater than that the t critical value at 95% confidence level at which we have to make the prediction and 5 degrees of freedom.

(Refer Slide Time: 18:47)

t Distribution: Critical Values of t

Degrees of Freedom	Two-tailed test		Significance level					
	10%	5%	2%	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	31.598		
3	2.353	3.182	4.541	5.841	10.215	12.924		
4	2.132	2.776	3.747	4.804	7.171	8.610		
5	2.015	2.571	3.365	4.032	5.893	6.859		
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	2.812	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	4.487		
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	3.644	3.965		
18	1.734	2.101	2.552	2.878	3.605	3.922		
19	1.729	2.093	2.539	2.861	3.570	3.883		
20	1.725	2.086	2.528	2.845	3.532	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		

So if I look back at the t table which is given here this number turns out to be 2.571 for 5 degrees of freedom and 95% confidence level or 5% significance level. So going back to the board, our t critical value will be 2.571 for 95% confidence level and 5 degrees of freedom. So let us do the maths.

(Refer Slide Time: 19:38)

Sample	Method I	Method II	Diff.	(Diff.) ²
1. NPTEL	500	480	+20	306.25 ←
2.	292	305	-13	240.25
3.	343	325	+18	240.25
4.	445	460	-15	306.25
5.	480	500	-20	506.25
6.	395	370	+25	506.25
			+15	2105.50

Av. diff = $\frac{15}{6} = 2.5$

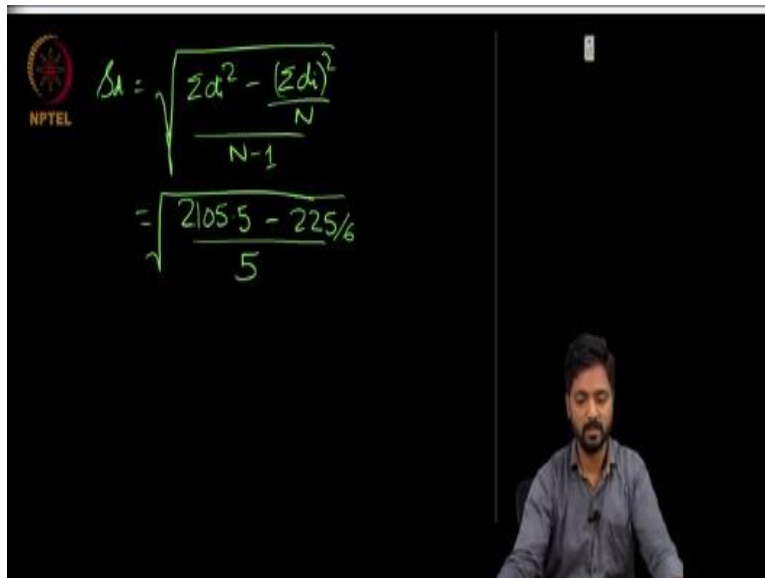
So what we have are sample method 1 and method 2 being applied to the same sample and for sample 1, a method 1 gave a reading of 500, by method 2 gave a reading of 480, sample 2 these numbers for 292 and 305, sample 3 it was 343 and 325, sample 4 it was 445 versus 460. Similarly, sample 5 and sample 6 have these readings as 480 and 500 and 395 and 370. So what

we do is we calculate the difference between the 2 methods and the difference will come out to be + 20 in the first case.

So method 1 reports more than method 2, in the second case method 2 is reporting more, so this difference becomes negative 13. Similarly, for sample 3, this is positive 18, sample 4 it is again negative 15, it is also negative 20. Because method 2 is reporting more, and finally for the sixth sample the value is + 25. So, if we sum up all these differences, then the sum of total sum of differences is +15, taking into consideration all the signs that are present.

We also do the difference squared and that number comes out as, so how does this number of 306.25 comes. So what we have done is we have calculated the average difference, which is nothing but 15 by 6, which is 2.5. And for the first reading, we have subtracted 2.5 from 20 and square that number. So 17.5 square will be 306.25 when we do this exercise down the table, the numbers turn out to be 306.25, 506.25 and again 506.25 and you can sum it all of this 2105.5. So, from these numbers, we will be able to calculate the standard deviation of the differences.

(Refer Slide Time: 23:17)


$$s_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{N}}{N-1}}$$
$$= \sqrt{\frac{2105.5 - \frac{225}{6}}{5}}$$

St value will be square root of and when we plug in these values this would be square root of 2105.5 - 15 square that is 225 divided by 6 and all divided by 5.

(Refer Slide Time: 24:08)

$$S_d = \sqrt{\frac{\text{Dev}^2}{N-1}} = \sqrt{\frac{2105.5}{5}}$$

$$= 20.52$$

$$t_{\text{calc}} = \frac{\bar{d}}{S_d} \times \sqrt{N}$$

$$= \frac{2.5}{20.52} \times \sqrt{6}$$

$$= 0.298$$

And square root of deviations squared by $N - 1$ which is nothing but 2105.5 divided by 5 square root and this number will ultimately come out to be 20.52. So our standard deviation of the differences between the 2 methods is 20.52 and the average difference was 2.5. With these let us calculate the t calculated value which is \bar{d} by s_d the into root N . Now \bar{d} is 2.5 divided by 20.52 into square root 6, this upon solving will come out to be 0.298.

(Refer Slide Time: 25:25)

$$t_{\text{calc}} = 0.298 ; t_{\text{crit}} = 2.571$$

$$H_0: \text{Method 1} \sim \text{Method 2}$$

$$t_{\text{calc}} < t_{\text{crit}}$$

$$H_0 \text{ hold true at } 95\% \text{ CL.}$$

Now, going to the next page, we have the t calculated value as 0.298 while our t critical value was 2.571. And remember that our null hypothesis here was that method 1 gives similar results as method 2. And when we test the null hypothesis here, we observed that that the t calculated value is actually less than the t critical value. That means, the null hypothesis holds true at 95%

confidence level or method 1 and method 2 giving us statistically similar outputs or similar results.

So, you can see that you can apply disparity test on methods which are being applied on the same samples. Now, I have also collected a couple of other questions for us to discuss, which will sort of summarize our understanding of this course for this week. So, let us quickly go to these questions.

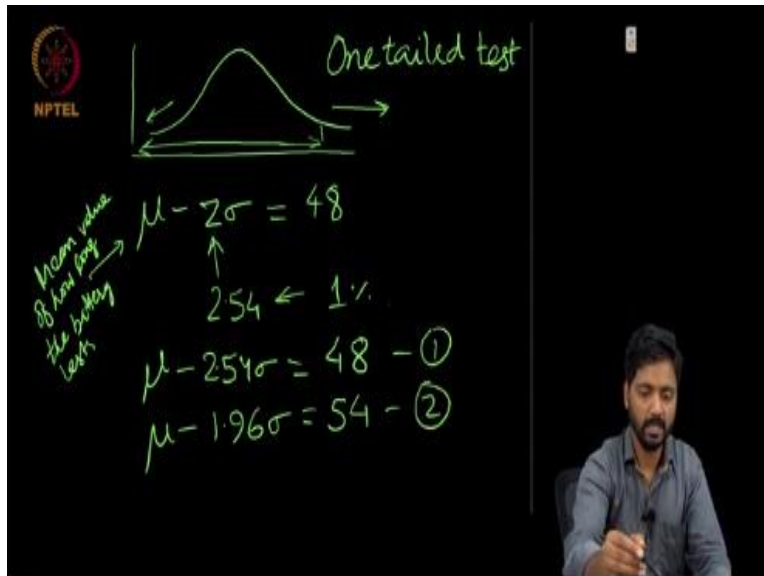
(Refer Slide Time: 26:56)

Q1. A company warranties its batteries for 48 months so that only 0.5% of its batteries fail during the warranty period. However, in 54 months, 2.5% of the batteries fail. What is the average life of the batteries manufactured by this company?

Q2. A medical test data is presumed to be in normal range if it lies in 95% Confidence Interval in a two tailed test. If a sample registers a low "out of range" value, how likely is the sample to indicate a disease that inflicts 0.1% population?

So, let us take this example of a battery company that warranties its batteries for 48 months. So that only point 5% of it batteries fail within this warranty time period. However, immediately after the warranty time period gets over that is in another 6 months there is of a significant reduction in the number of working batteries and 2.5% of the batteries fail to perform. So, with this data, the question that we are dealing with is what is the average life of the battery is being manufactured by this company. So, let us go back to the board and understand how this problem is to be approached.

(Refer Slide Time: 27:51)

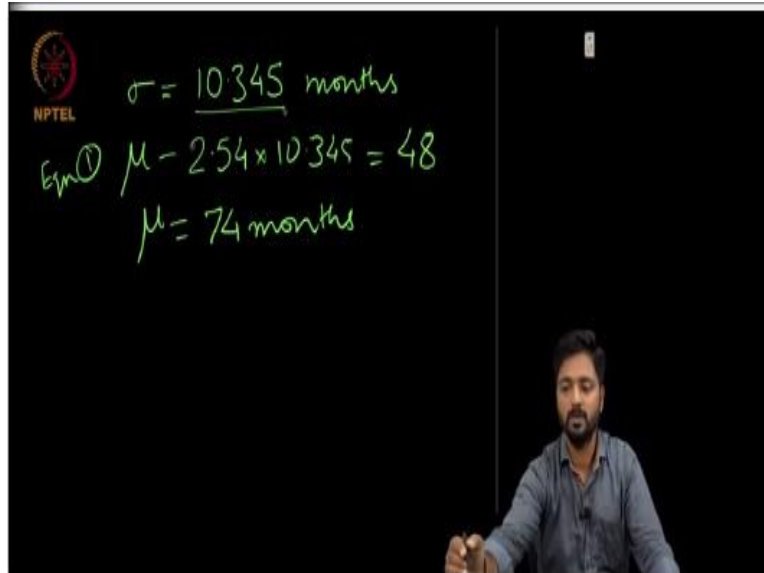


Now here since we are only concerned with batteries that fail during the 48 month or 54 month period, we are concerned only with the 1 tail of the Gaussian profile. So, we are only concerned with this portion of the Gaussian profile, any battery that keeps performing beyond 48 or 54 months is considered to be good. So, in other words when we say that 0.5% of the batteries are failing, we are essentially implying a significance level of 1%.

So, just to repeat this point, we are dealing with a one tailed test here. So, what we get from this data is if μ is the mean value of how long the battery lasts. Then $\mu - z\sigma$ in this case is 48 hours 48 months, when the z value in this case is 2.54 corresponding to a significance level of 1% in the 2 tail test. So, if we plug in these values, we get $\mu - 2.54\sigma = 48$ similarly, since 2.5% of the batteries are failing.

Then the z value in the second case will come down to 1.96 which implies that a total of 5% of the significance level for the 2 tail test. So, where $\mu - 2.54\sigma$ is 48 months and another where $\mu - 1.96\sigma$ is 54 months.

(Refer Slide Time: 30:44)



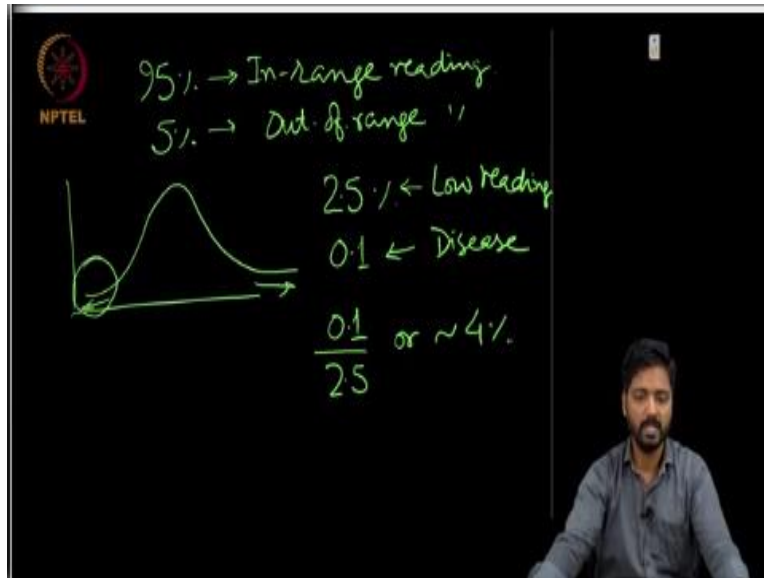
So, when we solve these 2 equations, what we get is that $\sigma = 10.345$ months and when we plug this into $\mu - 2.54 \times 10.345 = 48$ which was our equation 1. Then we obtain the μ value as 74 months approximately. So, 74 months is the approximate life of the battery is being manufactured by this company and with a standard deviation value of 10.345 months. And if you have these numbers, the company can warrant its batteries for 48 months with only 0.5% battery is being failing this time.

Now, let me take another question here, which corresponds to how medical tests are performed and how inferences are made through them. So, typically when we undergo a medical test a blood test or a urine test, there is a normal range which is specified. And any numbers greater than or lower than these normal range are considered to be suspicious. But being suspicious does not always mean that there is a problem with the sample or the patient from whom this sample is taken.

So, in this question there is a medical test presumed to be in normal range. If it lies in the 95% confidence interval, that means 95% of the population will have its numbers within this range. And if a sample registered low out of range value, how likely is this sample to indicate a disease that inflicts 0.1% population. So, suppose there is a disease which we know through previous studies that it inflicts only 0.1% population.

And it gives also in the patients who have these diseases, there is a particular analyte whose value comes out to be low. So we conduct a blood test on a patient and we indeed find that the blood sample gives out a low out of range value, do we say that this sample is coming from the patient who has a disease or not. Or how likely is the patient to have the disease just because this number turned out to be low. So let us do our calculations quickly, just go back to the board.

(Refer Slide Time: 33:56)



And here what we know it is that 95% population would give the In-range reading and 5% of the population will give out of range dating. Now this out of range can be either on the left side or on the right side. So, the numbers can be significantly low or the numbers can be significantly higher than the prescribed range of values that they should have. So, in this case the question says that the patient had a low value.

That means we are talking only about the left tail of this Gaussian profile. So in other words, there is only 2.5% population that will give out a lower reading than the normal values. Now this 2.5% population is that which will give out a low reading and only 0.1% population will actually have the disease. So the probability of a person giving a low rating and also having the disease would be 0.1 by 2.5 or around 4%.

So, only in 4% of the patients who give low reading of this analyte can there be actual disease. So I hope you understood how we apply statistics to understand how blood tests are performed

and how predictions are made through these tests. So this brings us to the end of our fourth week of classes for this course. And in the next week, we will be discussing about how analysis of variants is performed on a large dataset, thank you.