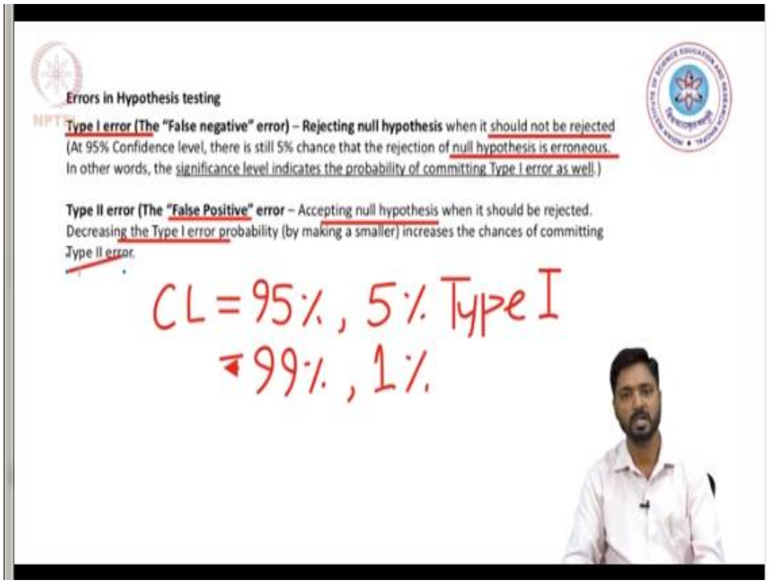
Quantitative Methods in Chemistry

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

Lecture-16
Hypothesis testing and Finding Outliers Part 02

(Refer Slide Time: 00:28)



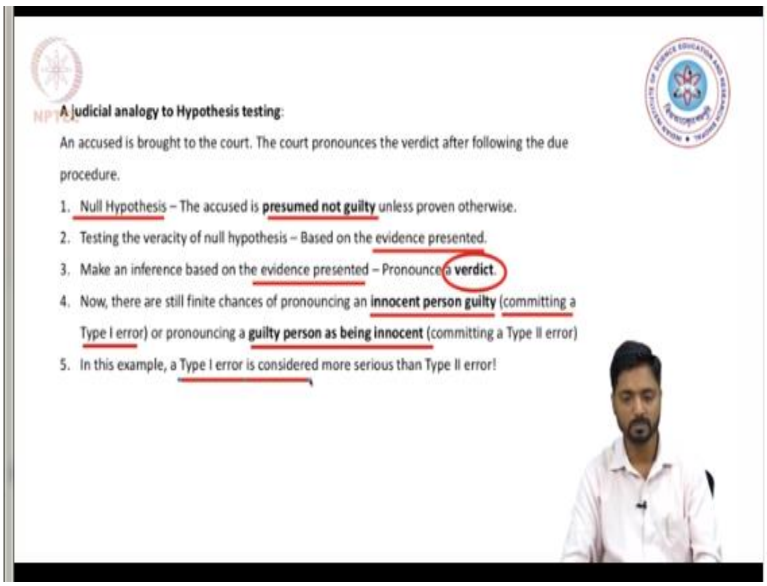
Now let me talk to you about the errors that can creep in when we do our hypothesis testing by applying the T or Z tests. There are 2 types of errors that we can encounter while making our inference, the first one is known as the type 1 error and it is also known as the false negative error. And what it means is that, we are rejecting the null hypothesis when it should not be rejected.

For example at 95% confidence level, there is still 5% chance that we will reject the null hypothesis in an erroneous fashion or we will make an error if we reject the null hypothesis. So the significance level indicates the probability of committing the type 1 error. Now, what can we do to reduce the type 1 error, we can reduce the confidence level and reduce the significance level to say 1%. In that case, we will be still having 1% probability of making a type 1 error.

So just to rewrite it confidence level is 95%, we still have 5% chance of making a type 1 error. However, if we increase the confidence level to 99% we reduce this probability to 1% of making type 1 error. Now type 2 error on the other hand is quite opposite, it is also known as the false positive error and it is accepting the null hypothesis when it should be rejected, so what we can do is.

We can sometimes decrease the type 1 error probability and by increasing the confidence level, however when we do that, we increase our chances of committing type 2 error. So there needs to be a sort of a compromise made between the 2 errors that we can commit and that will depend upon the context at hand.

(Refer Slide Time: 03:56)

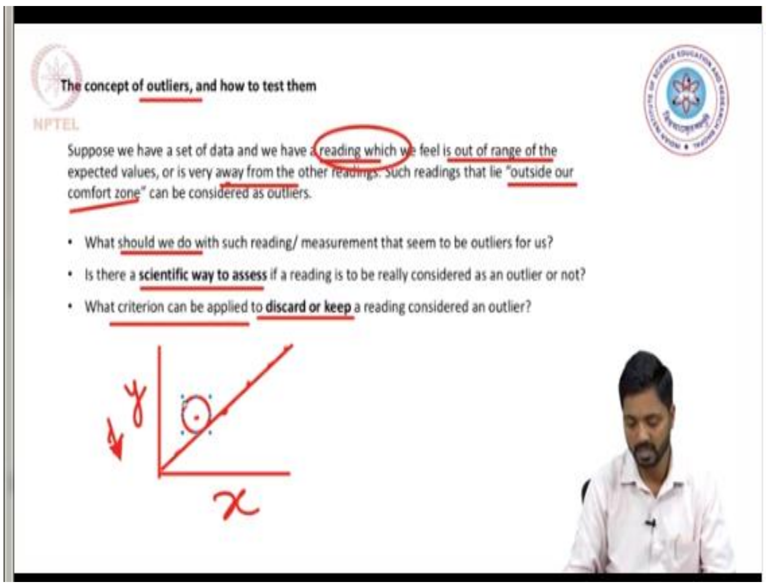


An example for that is the judicial analogy of hypothesis testing, so when an accused is brought to the court. The court pronounces the verdict after following a due procedure which is very similar to the hypothesis testing that we discussed just now. So for example, it will start with a null hypothesis, it will presume that the person accused is actually not guilty, unless proven otherwise.

Then based on the evidence presented, the veracity of the null hypothesis will be tested and an inference made based on the evidence provided, and that inference is what we called as the verdict of the court. Now, when the court pronounces it is verdict based on the available evidence, there are finite chances of either pronouncing an innocent person guilty that is committing a type 1 error or pronouncing a guilty person as innocent, which is considered as the type 2 error.

So in the first case, we are essentially rejecting the null hypothesis of the person being not guilty by committing a type 1 error. While in the second case, we are accepting the null hypothesis when it should not be accepted or when the person is actually guilty but is being pronounced innocent. Now we all understand that in the jurisprudence, the type 1 error is considered to be more serious than type 2 error.

(Refer Slide Time: 05:58)



Now, towards the last part of this class, let us talk about the concept of outliers and how do we test them using statistics, so how do we define outliers. Outliers are the values or readings which we just feel based on our experience or our presumptions, we feel that they are out of range of expected values. Now this may or may not be true or the value at hand is quite away from the other readings.

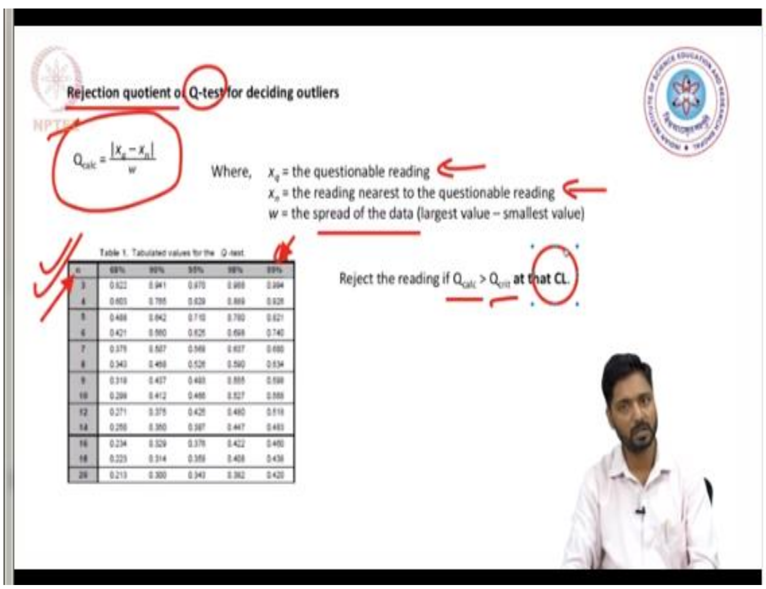
Let us take 1 simple example, where we are simply plotting the absorbance value. Now towards the last part of this class, let us look at the concept of outliers and how do we test them. So suppose we have a set of data and we have a reading, which we feel is out of range of the expected values or it is very far away from other readings. Such readings which lie outside our comfort zone are considered as outliers.

Now what should we do with such readings, is the question that we would want to deal with and if there is a scientific way to assess that the reading being considered as an outlier is indeed an outlier or not. In other words, what is the criteria that we can apply to discard or keep reading, which we consider as an outlier. Now take an example, where we are suppose to get a straight line between y and x.

So we see that about 5 of the 6 points that I have drawn, actually fall very well along a certain line, but 1 point which is encircled now, seems to be lying outside this expected value of linearity. Now should we be keeping this data or should we be reporting this data or not, is the question that we are dealing with. It has to be remembered that it is considered scientifically incorrect.

And in fact in certain cases unethical to delete or to not report outliers unless we test them properly and are sure that they are indeed outliers. Now what is the criteria or how do we assess that reading is an outlier.

(Refer Slide Time: 09:23)

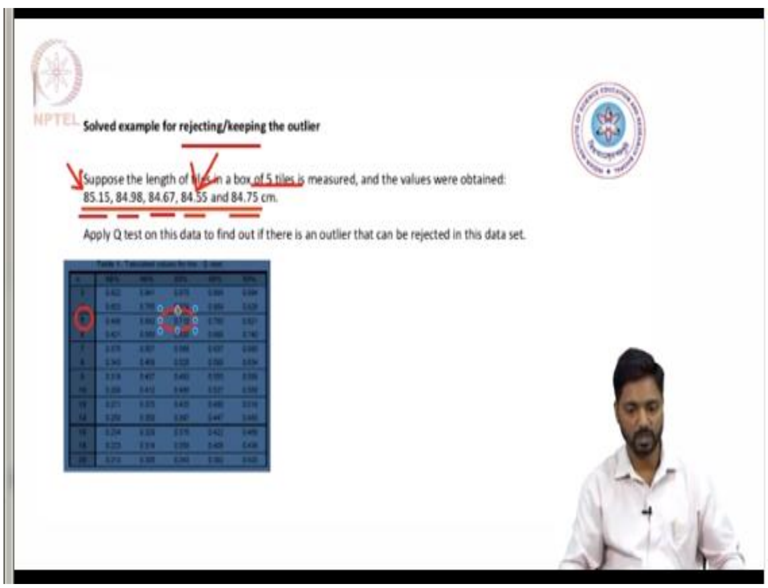


We utilize what is known as the rejection quotient or a Q-test to test, whether a reading is to be considered as an outlier. The formula is given here, where we calculate the Q-value of a data set, so x Q is the question enable reading x n is the reading which is very near to the questionable reading. And w is the spread of the data which is with us, so spread is nothing but the difference between the largest value of a dataset and the smallest value.

And we use Q table similar to the T table to assess or to make a conclusion, whether the reading at hand is indeed to be considered as an outlier or not ok. So the Q table again lists the confidence level with which we can report the results and it also talks about the number of readings in a dataset to choose a critical value of Q. So to rephrase a critical value of Q will depend both on the confidence level with which we want to state our result or conclusion and the number of readings present in a dataset.

And as previously, we compare the Q calculated with the Q critical at a certain confidence level and either we reject the reading as an outlier **r**, otherwise we do not.

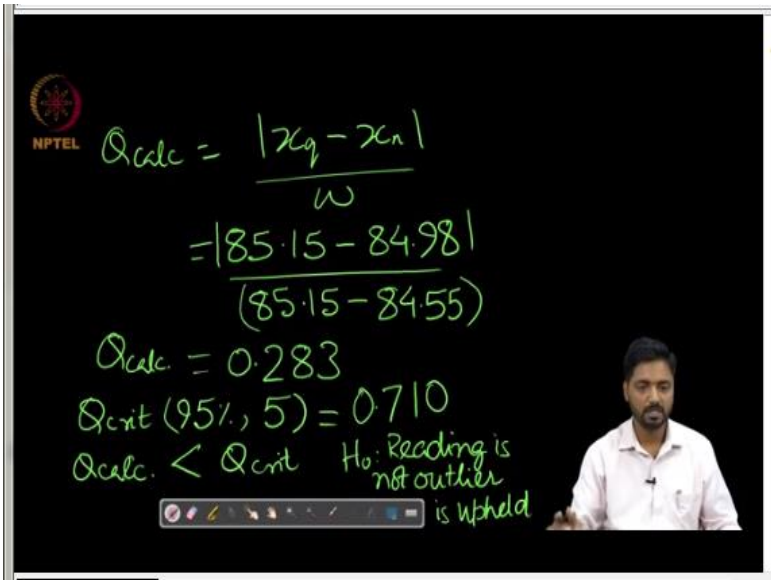
(Refer Slide Time: 11:31)



Let us take a quick example, where we are suppose to reject or keep an outlier. So let us take a **a** solved example to understand how Q table can be applied for rejecting or keeping an outlier. Suppose we are dealing with a box of tiles and there are 5 tiles there and the measured values of the length of the tiles are given as 85.15, 84.98, 84.67, 84.55 and 84.75. Now we need to apply the Q-test on this data to figure out if we are dealing with any outlier or not.

Now there can be 2 readings that can be considered as outliers, the largest and the smallest. In this case, 85.15 is the largest reading and 84.55 is the reading which is smaller. So either of these readings could be an outlier and let us quickly perform the Q-test to test which of the readings if any is to be considered as an outlier, what we said that.

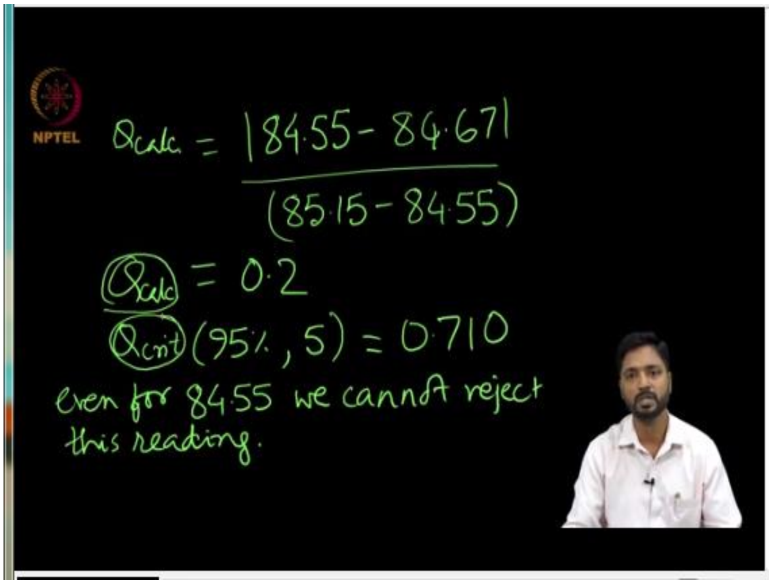
(Refer Slide Time: 13:01)



The Q calculated value is that questionable reading minus it is nearest neighbor divided by the spread of the data, for the data set at hand for the reading which is 85.15. This is how we will do the calculations and this finally solves to a Q calculated off 0.283. Now the Q critical as discussed will depend upon the confidence level, let us say we are talking about 95% confidence level and the number of readings here the number of readings were 5.

Now let us go back to the Q table and figure out what the Q critical value will be, so n as 5 is shown here. And at 95% confidence level, the Q value for n = 5 is 0.71, so we put 0.710 here and what we observe is, that the Q calculated is less than the critical value. In other words, the null hypothesis which is that the reading is not an outlier is upheld. So we cannot reject the tile that has 85.15 centimeters as it is length.

(Refer Slide Time: 15:47)



Similarly when we do the exercise for the smallest value of length which is 84.55, the nearest reading to it is 84.67 and the width of the data or the spread of the data is this. Now when we solve this up, what it comes out is, that the Q calculated value in this case is only 0.2, while the Q critical at 95% confidence level and 5 readings is still 0.710. In other words even for 84.55, we cannot reject this reading ok.

So what you see is that, when we use the Q calculated and the Q critical, we can make a clear judgment whether reading is to be considered as an outlier or not. So with this, we will stop this lecture 2 of week 4 and I thank you for your attention.