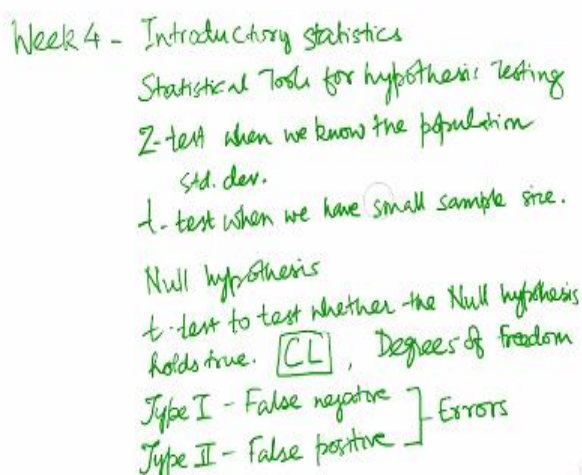


Quantitative Methods in Chemistry
Dr. Aasheesh Srivastava
Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research – Bhopal

Lecture – 49
Course Revision (Week 04 and 05)

So thank you Bharathwaj. I think it has been very interesting and exciting 11 weeks of discussions on various aspects of quantitative methods in chemistry and I am also very pleased by the extent and the diversity of the coverage of the topics that we could achieve in the last 11 weeks. So we have been summarizing this week what all we learned in this course. (Refer Slide Time: 00:57)



Week 4 - Introductory statistics
Statistical Tools for hypothesis testing
Z-test when we know the population
std. dev.
t-test when we have small sample size.
Null hypothesis
t-test to test whether the Null hypothesis
holds true. CL , Degrees of freedom
Type I - False negative
Type II - False positive } Errors



And for week 4, our focus was mainly on introductory statistics and how we can utilize a variety of statistical tools for hypothesis testing. So this is very important because we realized in this week that the random fluctuations in our measurements can often result in the measured values being very far away from the mean value of the dataset and we need to apply statistical tool such as the Z-test.

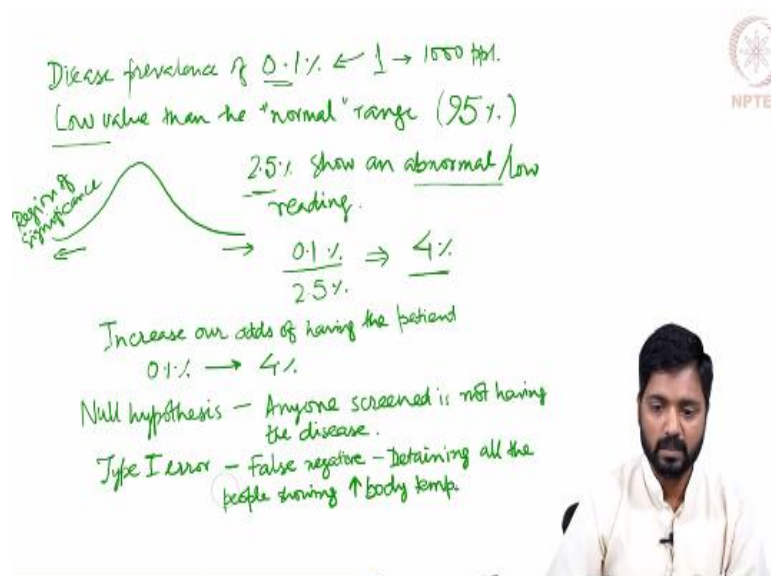
When we know the population standard deviation and the t-test which is also known as students t-test and we discussed that it was developed by William Grosse who gave it for dealing with small samples. So the t-test is applied when we have and usually in our day-to-day lives we will be dealing with the t-tests just because typically the number of samples that we deal are finite.

And in those cases t-tests turns out to be a more appropriate test to be undertaken for testing our hypothesis. We also got introduced to an important concept of null hypothesis and how we can apply the t-test to test whether the null hypothesis holds true and we talked in terms of the confidence level this is a very important concept that we also got introduced to. So the confidence level will in turn decide the confidence interval.

And our inference or the results will change based on what confidence level we are talking about. So our null hypothesis can be true at say for example 95% confidence level, but it may not hold true at a higher or a larger value of the confidence level and vice-versa. So it is very important to keep in mind the concept of confidence level and also the degrees of freedom when we are testing the various hypothesis.

And it is here that we also understood that while testing the hypothesis using these statistical tools we can commit variety of errors basically 2 types actually. So one is the Type-I error which we call as the false negative error where we reject the null hypothesis even though it is supposed to be true. Similarly, there can be Type-II error which is called as the false positive where we accept the null hypothesis when it is actually false. So these 2 errors are to be kept in mind when we are undertaking the hypothesis testing.

(Refer Slide Time: 05:38)



So I discussed in the Week 4 lecture 2 a problem where we said that disease has a prevalence of 0.1% and in this disease a certain analyte would come out of range or have a low value then the normal reading or normal range and we realize that this normal range is defined to

often cover 95% of the population. So when we say that reading is having a low value what we are implying is the left tail of the Gaussian error profile is our region of significance.

Now I have been thinking about this. So essentially what we imply here is that about 2.5% of the population will show an abnormal low reading. So what we figured out in that problem was that if we see that out of the people who show this abnormally low reading which is 2.5% of the population only 0.1% are actually having the disease. So this implies that only about 4% of the people who have the reading of the analyte low will actually be having the disease.

So I have been thinking about this in the context of the recent times where we know that a viral infection is spreading rapidly and consider that actually this disease has a prevalence of 0.1% in the human population and we are screening people for having this disease or not. So one option for us is to screen all the people of a city or a country and that would be really herculean task.

So instead what has been done is to use the thermal screening where the body temperature of the people coming to say airport or railway stations is being measured through thermal detectors and those who are showing abnormally high body temperatures are being considered to be suspect cases. So when we say again here abnormally high we are implying now the right side of the tail of the Gaussian curve.

And presume again that about 2.5% of the people will actually have a higher temperature reading. So the question arises what is the utility of undertaking this thermal screening. So if you see that the prevalence of the disease is 0.01% so for finding one actual patient we will have to screen 1000 people on an average, but if we can prescreen the people based on their temperature rise we will have a 4% chance of finding the people with the disease.

So we essentially increase our odds of having the patient or finding the patient significantly from 0.1% to 4% just by using the increased body temperature as a prescreening parameter. Now we can also discuss what are the errors that we can commit in undertaking this exercise. So we can also discuss the null hypothesis here in this context will be that anyone who is screened is not having the disease.

Now Type-I error in this context would be what we call as the false negative that is when we reject null hypothesis when it is true and that would imply detaining all the people showing increased body temperature. So obviously as we can think that just because one has a high body temperature does not mean that the person is actually diseased, but we are increasing our odds of finding a person who has the disease by undertaking the thermal screening.

(Refer Slide Time: 11:55)

Type II error - letting go a diseased person because they still don't have high temp. onset.

Sources of error?
↓
Thermal screening.

Z-test, t-test, F-test (precisions)
Q-test (Finding the outliers)

NPTEL

Similarly, the Type-II error in this case would be letting go a diseased person because say they are not showing the symptoms or they still do not have high temperature onset. So this again is a very real possibility to happen so if I draw a Venn diagram of the situation this is the whole population that we have and suppose we have the diseased people or the people who have the diseased being represented by this red region.

What we are essentially doing through thermal screening is shown in the blue profile where we are essentially increasing our odds of finding the people with this disease. We will have a greater probability of finding the people who have the disease, but we may still commit a Type-II error in cases where the thermal screening will be unclear or unaffectionate. Now we can also think of what all are the various sources of error that may arise during the thermal screening.

So this part I leave to you to think and figure out what could be the sources of errors that may come when we are screening patients for their high body temperature through thermal screens. So essentially what we covered in week 4 was the utility of the Z-test, the t-test and



we also got introduced to the F-test for comparing the precisions and also the Q test for finding the outliers in a dataset.

(Refer Slide Time: 15:00)

Week 5 ANALYSIS OF VARIANCE (ANOVA)
 Single factor ANOVA (One variable)
 Factor (Temp)
 Levels ($T_1, T_2, T_3, T_4, \dots$)
 Response (EE multiple times)

ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F _{calc.}
Between the groups (Factor effect)	SSF	I-1	MSF = $\frac{SSF}{I-1}$	$F = \frac{MSF}{MSE}$
Within the group (Error Effect)	SSE	N-I	MSE = $\frac{SSE}{N-I}$	
Total	SST	N-1		

So this essentially set us well for the next week discussion which was on what is known as the analysis of variance or in short ANOVA. Now the analysis of variance that we discussed in this course was known as the single factor ANOVA that means we have been dealing with only one variable in this. Now as discussed in week 5 this variable can be temperature, the variable can be the person who is conducting the measurement or any other variable that we can think of for example (()) (16:13) can be another variable.

And we had essentially 3 things here. One of course is the factor which is being varied and then we had levels that means how the factor was varied for example if the factor is temperature then the levels will be the various temperatures at which a certain measurement was repeated. So the factor if it is temperature then the levels will be T1, T2, T3, T4 and so on and then we have the response.

For example, this response could be the yield of the reaction or the enantiomeric excess for example and what we will have to do is to perform the reaction at a particular temperature T1 for multiple times. So the change or the differences in the enantiomeric excess within a temperature value will be compared with the changes in the enantiomeric excess at different temperatures T1, T2, T3, T4 and so on.

So based on all these we will be creating what is known as the analysis of variance table and here essentially we saw that we analysed the source of variation then we talked about the sum of squares, the degrees of freedom and finally the mean square values and the F calculated value and in the source of variation we could have 2 sources of variation what we call the between the groups which is known as the factor effect.

And also within the group which is known as the error effect. Consequently, we have the sum of squares due to factors which is enlisted here and the sum of squares due to errors the degrees of freedom in these 2 cases are different if we are dealing with I number of sets or I groups then the degrees of freedom for the factor effect will be $I - 1$ and for the sum of squares due to error the degrees of freedom is $N - I$.

Now from this we can calculate the mean square factor which is nothing, but sum of squares factor divided by the degrees of freedom $I - 1$ and the mean square error which is the sum of squares due to error divided by $N - I$ degrees of freedom. From here we calculate the F calculated value which is nothing but MSF divided by MSE and we can also do a total value here.

So this will become SST which is nothing, but the sum of SSE and SSF. The degrees of freedom total would be $N - 1$ if we are dealing with a total of N readings. So once we create this ANOVA table we get a F calculated value which is then compared with a F critical value.

(Refer Slide Time: 21:07)

*F_{calc.} is compared with F_{crit}
H₀ (Null hypothesis) holds true or not at a particular Confidence Level.
F_{crit} (CL, I-1, N-I) < F_{calc.}
Reject Null hypothesis
Least Significant Difference

$$LSD = t_{(CL, N-I)} \sqrt{\frac{2 \cdot MSE}{N_B}}$$
*N_B ← No. of replicates**



So F calculated is compared with F critical value to estimate or to infer whether the H_0 or the null hypothesis holds true or not at a particular confidence level. Now F critical we estimate by taking into consideration the confidence level value and the $I - 1$ degrees of freedom for the numerator and $N - I$ degrees of freedom for the denominator. Now once we have a F critical value which is less than the F calculated value.

Then we reject the null hypothesis and at this stage we can only say that at least 2 or more groups of the given dataset are having results which are significantly or means which are significantly different from each other to understand or to quantify which of the groups are actually having a difference we use what is known as the least significant difference and we calculate this least significant difference value as $t \sqrt{2}$ into mean square error by N_g where N_g is the number of replicates in each group.

The t value here is having a particular confidence level value and the degrees of freedom are the same as that for the error which is $N - I$. So from here we calculate a value of the difference between the means which is considered to be least significant at a certain confidence level and from here we can estimate which of the groups are actually having different mean values between each other.

Now once we have done all this exercise week 6 and week 7 were essentially devoted to the concept of data fitting and using different models to fit the data and estimate the parameters that require to be estimated and Dr. Bharathwaj Sathyamoorthy covered this for both linear data fitting as well as the non-linear data fitting.