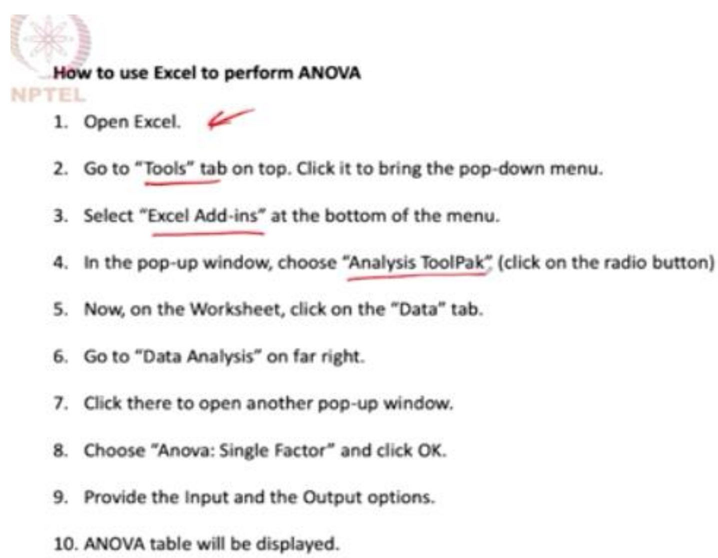
Quantitative Methods in Chemistry
Dr. Aasheesh Srivastava
Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research - Bhopal

Lecture - 21
ANOVA and Least Significant Difference (LSD)

Hello and welcome back to lecture 3 of week 5 of this course Quantitative Methods in Chemistry. So far in this course in this week, we have learned how to apply F test to understand or to compare the precision between two samples, which are being compared and then we have gone ahead and learnt how analysis of variance is to be applied on samples, which are more in numbers than 2.

So samples, which are 4, 5 or 10 or more needs to be analyzed through analysis of variance rather than simple T test, because T tests will accumulate errors when we apply it on multiple samples. So with this understanding, let us start today’s lecture, which will be about how we can use Excel, Microsoft Excel to perform analysis of variance through a computer and then we will move ahead and we will also understand how least significance difference is to be applied on samples, which show that there are significant different in the means of samples being compared.

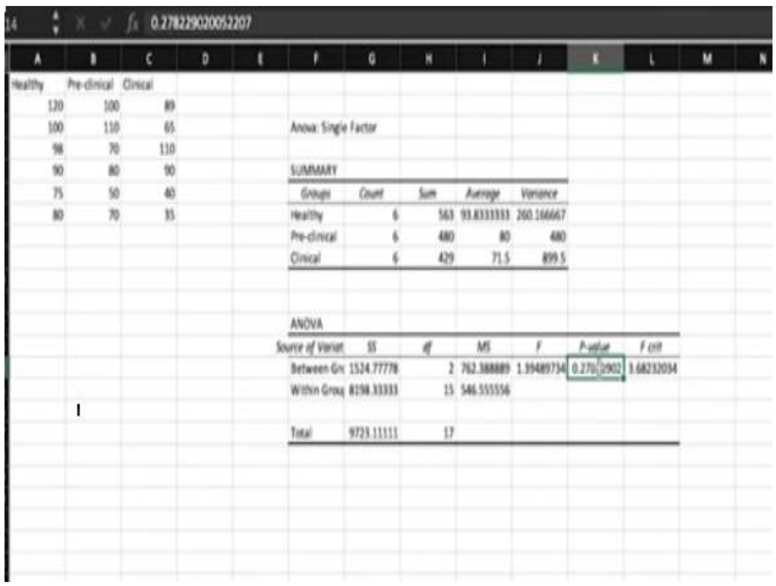
(Refer Slide Time: 01:53)



So let us go ahead and understand how we can use Excel to perform analysis of variance. So the steps are enlisted here. We need to start with opening the Excel and then in the tools tab, we go on top and click on the tools tab and this will bring down a pop down menu, from which we need to choose Excel add-ins and in the add-ins, there would be analysis tool pack, which we need to choose and through a click of radio button.

Now once we have this analysis tool pack, we can perform the analysis of variance. So let us actually go to Excel and perform this task.

(Refer Slide Time: 02:47)



So this is our Excel worksheet and what I mentioned was, we need to go to the tools pack here and come to Excel add-ins and in the Excel add-ins when we click this new window opens where analysis tool pack needs to be checked and ok symbol clicked. Now once we do that, as you might have noticed that the data analysis icon has got activated now and so let us perform analysis of variance by inserting some numbers here.

So what we did in the last lecture, we compared mean potassium levels of healthy, preclinical, and clinical samples and we wanted to address a question whether there was a significant difference in the mean potassium levels of these 3 groups. Let us quickly put in these numbers and perform the analysis of variance. So we insert healthy, preclinical and clinical, and like in the last class, we will enter the individual numbers for the groups.

Similarly, for the preclinical sample, these numbers were 100, 110, 70, 80, 50, and again 70. For the clinical samples, the so-called clinical samples, these values were given as 89, 65, 110, 90, 40, and 35. So we have understood our data, which needs to be analysed through ANOVA. Now let us perform the analysis of variance using the data analysis tool here. So once you click on the data analysis tool, what opens is a new window, where analysis tools of a wide variety are listed together.

So we go on the top and choose single factor ANOVA, which is what we have been performing throughout in this week. So once we choose single factor ANOVA and click ok, a new window pops up and it requires us to give the input range. Now let us click on that icon and choose the relevant cells and the cells are now chosen appropriately, we go back and what I want you to notice is that here.

We have ticked the icon, which says that there are labels in the first row, because in deed we have the labels like healthy, preclinical, and clinical present in the first row. So also note that there is an alpha value, which is by default put as 0.05. This means that we are dealing with a significant level of 5% or a confidence level of 95%. This is the default value, which is already present in the software. Now along with that, we need to give an output range here.

So let us click at the output range and choose a range where the output of the results are to be summarized. We click in the sheet and then simply click ok. So as you can see now that the Excel has already calculated for us all the parameters that are required for us for ANOVA. You can see that it has already labeled the groups as the healthy, preclinical or clinical and correctly the count in each of these data sets is of 6. So 6 measurement values in each data sets.

We also see that the sum within each data set is provided. For example, if you summate all these values in the first column, you should get a value of 563, similarly 480 for the preclinical sample, which is now highlighted here and for the clinical sample, this number is the summation of the numbers given in the third column.

So if we look at the actual numbers or the average values here for these data sets, it seems that indeed the average value of the potassium levels in the healthy samples are significantly higher than that in the clinical samples, while the preclinical samples are somewhere in between. However, you would note that the variance here, which is again automatically calculated is quite large in these samples and as you may recollect, variance is nothing but the standard deviation squared value.

So if we do the square root of these numbers, we will get the standard deviation values for these data sets. So the point to be noted here is that for all these data sets, we have large variances. For example, for the first one it is about 260, for the last one which is the clinical samples, this number is indeed pretty high at 900. That means the standard deviation value will be close to 30. So what we are dealing with a sample in clinical sample, we have an average value of 71.5.

But the standard deviation value would be about 30. Now, if you look at the ANOVA table, which is shown below, we have listed the sources of variation, which will be between the groups or within the groups and on the next column, we have the sum of square values enlisted and so we have in the first row in the sum of square column, it denotes the sum of squares due to factors, that is due to variations between the groups and the second row indicates the sum of squares errors, which is occurring due to the differences within the group.

Now, we have also enlisted the degrees of freedom here. So the degrees of freedom is 2, because we are dealing with three different samples and this degrees of freedom is for the variations between the groups. So there are three groups of samples that we are analyzing. So the degrees of freedom here is 2. Similarly, there is the degrees of freedom within the group, which is nothing but N minus i and N being 18 here, i being 3.

So the degrees of freedom in total come as 15 for within the group value. Also listed in this table is the mean square values, which is nothing but the sum of square values divided by the degrees of freedom. So essentially 762.388 is the value that comes when we divide 1524 by 2 and similarly dividing 8198.33 by 3 gives us the mean square error values of 546.55. when we undertake the calculation of F value using these numbers.

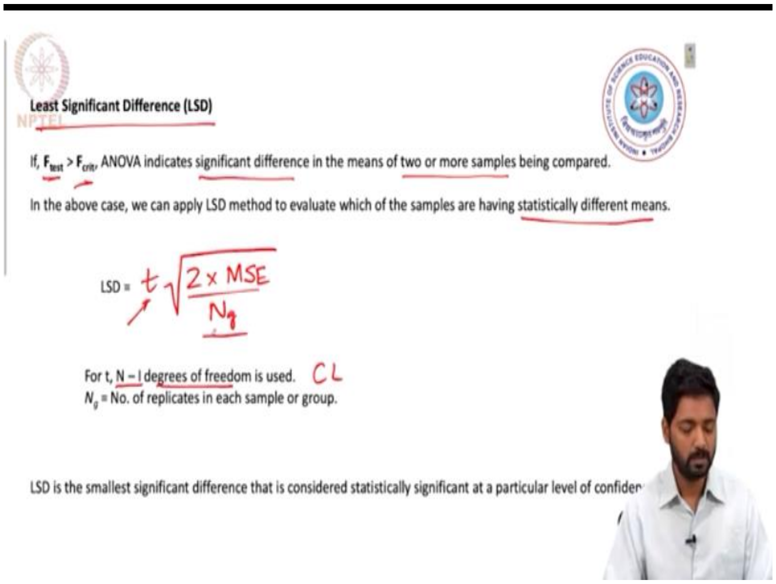
The calculated F value comes as 1.39 and the ANOVA table also enlists the F critical value for this data set, which will be enlisted at 95% confidence level or 5% significance level, which is what we choose during our calculations and based on 2 and 15 degrees of freedom, this F critical value turns out to be 3.68. Please also note that we have this P value listed here as well. P value is how or when this F calculated value.

How often will this come due to the random fluctuations in the data and we see that the P value is significantly higher at 0.278. That means there is 27.8% chance that this F calculated value would have come due to random fluctuations. Now note that we wanted to see whether these numbers are different at 5% significance level. So obviously, we see here that the P value of 0.278 indicates that the differences between the healthy, preclinical and clinical samples are not significant.

And that is also clearly shown through the F critical value being larger than the F calculated value. So let us go back to the presentation. So just to summarize, after clicking the Excel add-ins, the analysis tool pack was chosen and in the data tab, we will get the data analysis icon activated on choosing the analysis tool pack. We choose the ANOVA single factor for performing ANOVA and finally provided the input and output options and the ANOVA table was displayed.

So I hope this will make things much easier for you in using ANOVA in your day to day samples. However, for the conduct of this course and in the tests that will be conducted, you will be required to perform ANOVA through your hands.

(Refer Slide Time: 14:54)



Now let us go to the concept of least significant difference, which is enlisted here in the next slide and here what we understand is that if there is F test value, which is coming greater than the critical F value, then our analysis of variance indicates that there is a significant difference in the means of two or more samples that are being compared. However, at this moment, we are not sure which of these samples actually have different means.

For that, LSD method will be applied to evaluate which samples that are being compared have statistically significant different of the means. So LSD can be calculated as. So we use the T table here to find out what the T value is and for that, we will utilize the confidence level, that is usually 95%. It can be 99% also or 99.9% based on with what confidence we want to report our inference.

So the degrees of freedom here for the T value to be inserted for calculating LSD is N minus i and I would also want you to note that Ng is the number of replicates in each group. So the example that we recently covered using Excel had 6 replicates in each group. So for that, the Ng value was 6. So to summarize, LSD will be the smallest significant difference that is considered statistically significant at a particular level of confidence.