

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

**Lecture – 18**  
**Introduction to Analysis of Variance (ANOVA) and Comparing Precisions**

Hello and welcome back to this NPTEL course titled Quantitative Methods in Chemistry. We will be proceeding on week 5 of the course now, and this will be lecture 1. Before we proceed on this week's lecture, let us quickly recap what we learnt in the course in the last week.

**(Refer Slide Time: 00:50)**

**Last week we learnt about:**

- Z-test
- T-test
- Hypothesis testing
- Errors in Hypothesis testing
- Finding out outliers in data using Q-test

**This week we will learn about:**

- Need for conducting Analysis of Variance (ANOVA)
- Introduction to ANOVA
- Protocol for doing "single factor ANOVA"
- Obtaining "Least Significant Difference" (LSD) of means for multiple samples
- Examples of applying ANOVA and LSD


So, last week we got introduced to different statistical test such as Z-test and the T-test. Similarly, we got introduced to the concept of Hypothesis testing, and we also understood what are the potential sources of errors that occur in Hypothesis testing and finally we also learnt about applying Q-test to identify outliers in our data.

So, with this background, we are all set to embark on the next level of applying these understandings to what is known as Analysis of Variance. So, this week we will be essentially focusing on the Analysis of Variance and we will get introduced to the need for conducting Analysis of Variance, which is abbreviated as ANOVA. We will get introduced to the concept of ANOVA and protocol of doing single factor ANOVA.


So, we will understand what this term mean single factor and how do we apply Analysis of Variance to obtain Least Significant Difference when comparing means of multiple samples

and then we will conclude this week's lecture by taking a few examples of applying Analysis of Variance and Least Significant Difference on different test samples.

**(Refer Slide Time: 02:41)**



The need for conducting ANOVA




T-tests compare differences between two samples or populations.  
Question: Can we extend this to comparing n different samples?

Answer: No, because of the Error in Hypothesis testing. Each pair of data set compared brings about Type I or Type II error. We learnt that at 95% Confidence Level, there is 5% chance of committing Type I error i.e. rejecting  $H_0$  when it is true. As the number of data sets increase, the chance of committing this error increases.

If we are performing t-test on 10 samples at 95% CL, the overall probability of committing Type I error will be given by:  
 $1 - (0.95)^{10} = 1 - 0.599 = 0.4$  40% chance - Type I Error

So, there's 40% chance of committing Type I error when applying t-test on 10 samples at 95% CL.

Applying ANOVA reduces this error by condensing the comparison to two numbers.



So, let me begin with the discussion on what is the need for conducting Analysis of Variance. So far, in the last week, we saw that we can apply T-tests to compare differences between two samples or populations. So, the obvious question is, can we extend the T-test to compare n different samples. Suppose, we have 5 or 10 different samples that are being compared, and within these samples, we have individual readings.

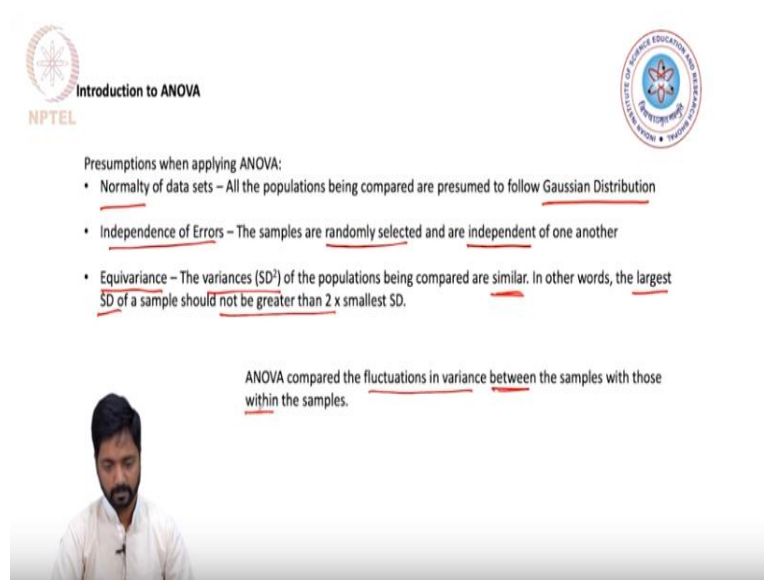
Can we apply T-test to compare the samples and arrive at a particular inference. So, the answer to that question is, no, we cannot apply T-test on very large number of samples because we got introduced to the term of the Error in Hypothesis testing, that is when each pair of data set is being compared, we are committing either Type 1 or Type II error based on the Confidence Level we are at. For example, we learnt in the last week that at 95% Confidence Level, we still have 5% chance of committing Type 1 error and Type 1 error is when we are rejecting  $H_0$  or the Null Hypothesis when it is actually true.

So the important point or take home message is that as the number of data sets increase, the chances of committing Type 1 error also increases. So, we will not be able to apply T-test on a large number of samples, and for example, I have calculated what is this error if we are dealing with 10 samples at 95% Confidence Level, and we are applying T-test and comparing different pairs of samples.

So, the overall probability in this case of committing Type 1 error, will be  $1-0.95$  is our Confidence Level, to the power of 10, which is the number of samples that are being compared. So, when we do these calculations, we figure out that there is about 40% chance of committing Type 1 error. Now, this is quite significant and that means that any prediction that we make or any inference that we have based on this statistical analysis will be erroneous by about 40%.

This cannot be ignored and that is the reason why we need to apply ANOVA in such cases, because applying ANOVA reduces this error by condensing the comparison to two numbers and we will see as we proceed through the lecture how this is done and what results do we obtain through Analysis of Variance.

**(Refer Slide Time: 06:43)**



The slide is titled "Introduction to ANOVA" and features the NPTEL logo on the left and the University of Pune logo on the right. It lists three presumptions for applying ANOVA:

- Normality of data sets – All the populations being compared are presumed to follow Gaussian Distribution
- Independence of Errors – The samples are randomly selected and are independent of one another
- Equivalence – The variances ( $SD^2$ ) of the populations being compared are similar. In other words, the largest SD of a sample should not be greater than 2 x smallest SD.

Below the list, a text box states: "ANOVA compared the fluctuations in variance between the samples with those within the samples."

In the bottom left corner, there is a small video inset showing a man with a beard and dark hair, wearing a white shirt, speaking.

So, let us get introduced to Analysis of Variance a little bit. So, there are certain presumptions, actually 3 main presumptions that are being made when we apply Analysis of Variance on a particular data set.

First of this presumption is the concept of Normality of data sets. So, what we imply here is that all the populations that are being compared are presumed to follow Gaussian Distribution.

So, we got introduced to the concept of Gaussian distribution of large data sets early in this course. Another point here is that of independence of errors. Now, what does this imply. This imply that the samples that we are comparing are randomly selected and are indeed

independent of one other. So, the errors that propagate in sample 1 will have no bearing on the errors that are present in sample 2 or sample 3 and so on.

Another very important concept, which is presumed here is that of Equivariance. So, what is presumed before we perform Analysis of Variance is that the variances, which is the standard deviation squared, please note the squared. So, the variances of the populations being compared are presumed to be similar. In other words, the largest standard deviation of a particular sample should not be greater than 2 times the smallest standard deviation of a different sample.

So, this is one important exercise that we need to do before we apply Analysis of Variance on our samples. So, in nutshell in Analysis of Variance what we attempt to do is to compare the fluctuations in variance between the samples with those within the samples. So, comparison between with within. So, we will see how this is done and what does this yield us to.

**(Refer Slide Time: 09:04)**

**Comparison of Variances using F-test**

NPTEL F-statistics is named after Ronald A. Fisher who made seminal contributions to the Statistics.

UNIVERSITY OF PUNE

Protocol:

1. Establish a Null Hypothesis – The population variances being compared are equal. In other words,  $s_1^2 = s_2^2$ .
2. Apply test statistics :  $F_{\text{test}} = s_1^2 / s_2^2$  (Note the square of std. dev.). The larger SD is kept on numerator.
3. Compare the  $F_{\text{test}}$  with  $F_{\text{critical}}$  at a particular Confidence Level. There will be two degrees of freedom involved, for sample 1 and sample 2, respectively.

$\nu_1$   $\nu_2$

At this point, it is important that we look into how variances are compared in statistics and for that, we employ a test, which is called the F-test or the F statistics employed when we are comparing the variances or in other words, the precision of two measurements.

So, just a brief sight, the F-statistics is named after Ronald A. Fisher who made very significant and seminal contributions in the field of statistics. So to honor him, the F-test is named after him. So, here the protocol that we apply for undertaking the F-test is that we establish a Null Hypothesis as always, so the Null Hypothesis presumes that the population

variances being compared are equal, so in other words, if we are comparing two samples, S1 and S2, and they have standard deviation, s1 and s2, then the square of the standard deviation values are similar.

Next, we apply the test statistics, we calculate the F-test value, which is nothing but the ratio of variance 1/variance 2. Again, now please be aware and please note that we are squaring the standard deviation to compare the variances, which is what variance is actually about. So, also note that when you are calculating the F-test value, you keep the larger standard deviation value on the numerator.

So, this is a protocol that needs to be followed to obtain a F-test value and this F-test value will be compared again with a statistical table of F and we compare it with a critical value of F, which is now found at a particular Confidence Level and now there would be two degrees of freedom associated here. One with regards to the sample 1, which will be at the numerator, which is often denoted by Nu1, and similarly there will be a degree of freedom that is associated with sample 2, which is denoted as Nu2.


(Refer Slide Time: 12:05)

The image shows a statistical table for the F-distribution at a 5% significance level. The table is organized with degrees of freedom (Nu1) in the columns (1 to 20) and degrees of freedom (Nu2) in the rows (1 to 100). A red arrow points to the 10th column, and a handwritten note '99% CL' is present. The table is part of a presentation slide, with logos for NPTEL and the University of Engineering & Technology visible. A video inset shows a presenter in the bottom right corner.


So, just to give a quick view of what these statistical tables look like. This is one F-distribution table, which is shown here and the critical value of F are tabulated at 5% significance level or in other words, 95% Confidence Level. Now, this is the Confidence Level, which we have been working quite often in this course, so I also want you to note that you have a Nu1 value, which is in this case tabulated up to 20 Degrees Of Freedom and similarly you have Nu2 value, which is tabulated up to 1000 Degrees Of Freedom.

So, this is the table that we have at 95% Confidence Level. Now, going to 99% Confidence Level or 1% Significance Level, we have another table, which is again similarly tabulated with the 2 Degrees Of Freedom being mentioned in the table. So, we can utilize the F-table and conduct the F-test to compare the variances or the precisions between two samples that are being compared.

**(Refer Slide Time: 13:48)**




Example of applying F-test to compare the precision of two samples



Question: A standard method based on 1001 data points for estimation of Na ions has a SD value of 0.18 ppm. A new method based on 21 data points had SD value of 0.21 ppm. Can we say at 95% confidence level that the new method is more imprecise than the standard method?

Answer:

1. Apply null hypothesis :  $H_0 = SD1 \sim SD2$ .
2. Calculate  $F_{test} = (0.21)^2 / (0.18)^2 = 0.0441 / 0.0324 = 1.36$ .
3. Compare with  $F_{critical} (95\%, 20, 1000) = 1.58$ .
4. Since  $F_{test} < F_{critical}$ , null hypothesis cannot be rejected. That means, at 95% CL, the two methods being compared have similar precision.



Let us actually apply this F-test to compare the precision of two samples. So, suppose we are dealing with a standard method, which was based on say 1001 data points for estimation of Na ions in water and this standard method has a standard deviation value of 0.18 ppm. Now, we come up with a new method, which is based on only 21 data points and has a standard deviation value of 0.21 ppm. So, we see that for the reference of standard method, the standard deviation value of 0.18 seems a little lower than the standard deviation value of the new method, which is 0.21 ppm.

So, the question that we need to answer is, can we say at 95% Confidence Level, that the new method is more precise than the standard method. So, how do we approach this problem is that we first apply the Null Hypothesis and the Null Hypothesis presumes that the standard deviation for the reference and test methods are similar, that means that the 2 methods are indeed similarly precise in giving their results with regards to the sodium ions present in the samples being analyzed.

Now, we calculate the F-test value here. So, again I want you to note that we have put the larger standard deviation value of 0.21 on the numerator and that is squared. Similarly, the smaller standard deviation value of the reference method of 0.18 is put in the denominator and again it is squared. When we undertake this calculation, what we get is a F-test value of 1.36.

Now, this F-test value needs to be compared with F-critical value of 95% Confidence Level and 20 Degrees Of Freedom for our new method, which is giving a higher standard deviation value and 1000 Degrees Of Freedom for the old or the reference method, which was giving a lower standard deviation value of 0.18 ppm.

Now, let us quickly go back to the F-table, at 95% Confidence Level and see what the values would be. So, at 95% Confidence Level, Nu1 of 20 and Nu2 of 1000 gives us a F-critical value of 1.58. When, we insert this value here, and compare the F-critical value with the F-test value what we obtain is a test statistics is less than the critical statistics. So, what we can say is that under these conditions, the Null Hypothesis cannot be rejected. That means at 95% Confidence Level, the 2 methods being compared have similar precision. So, we see how the F-test is being applied in comparing the 2 different methods or samples.

**(Refer Slide Time: 18:18)**

The slide is titled "Single Factor ANOVA" and features the NPTEL logo in the top left and the Indian Institute of Technology Bombay logo in the top right. Below the title, it lists "Questions that we want to address:" followed by two questions: "1. Does the yield of a reaction change with temperature?" and "2. Does the spectral profile of the molecule change with pH?". The words "yield" and "spectral profile" are circled in red, and "temperature" and "pH" are underlined in red. Red arrows point from the underlined words to the circled words. Below the questions, definitions are provided: "Factor – the variable that is being changed.", "Level – Individual measurements within each factor.", and "Response – The change in the value of measurement being performed." At the bottom, a text box states "Single factor ANOVA implies that there is only one factor that is being varied." A video feed of a speaker is visible in the bottom left corner.

Now, let me take you to the concept of Single Factor ANOVA. So, we will get introduced to the Single Factor ANOVA in this lecture and in the next lecture, we will be applying the Single Factor ANOVA on a real data set. So, the questions that we can address with Single Factor ANOVA can be that whether the yield of a reaction changes with temperature or with

a spectral profile of a molecule changes with pH. So, the terms highlighted in red here are our variable and this is also known as the factor in Analysis of Variance.

So, the factor is being varied in our analysis and what we do is, we have the spectral profile or the yield of the reaction as the different levels. That means, that at a particular temperature, we calculate the yield at multiple times, similarly we change the temperature by a certain unit and again calculate the yields multiple times. So, the yields of the reaction in this case is the levels that is being changed. Now, the response is nothing but the change in the value of measurement being performed.

For example, when we are comparing the spectral profile of a molecule with a change of pH, it is the wavelength which will be the response and the spectral profiles generated at different pH will be at different levels, and pH is a variable here, so it will be the factor. So, a Single Factor ANOVA, in fact, implies that there is only one factor that is being varied in these analysis. So, what we will do is, we will undertake a detailed example of how Analysis of Variance is to be conducted for a Single Factor in the upcoming lecture.