

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

Lecture – 19
Protocol for Undertaking ANOVA

(Refer Slide Time: 00:28)

Protocol for undertaking ANOVA
NPTEL

Presumptions: Normality, Independence of Errors and Equivariance

If we are analysing I different samples:



Samples	1	2	3	...	I
Data points	N_1	N_2	N_3	...	N_I
Means	\bar{x}_1	\bar{x}_2	\bar{x}_3	...	\bar{x}_I
Std. Dev.	s_1	s_2	s_3	...	s_I
Variance	$(s_1)^2$	$(s_2)^2$	$(s_3)^2$...	$(s_I)^2$

So as we can see that in Single Factor ANOVA, there is actually only one variable that is being varied in our analysis. So with this background, let us start applying our understanding to how ANOVA is to be conducted. So, the protocol of undertaking ANOVA. So, please remember that we still have our additional presumptions that there is normality, there is independence of error, and there is Equivariance.


So, these three presumptions still hold for us and in ANOVA, if we are analyzing I different samples, then we can have I have made it in a tabular form that we have samples as 1, 2, 3 up to I, and these samples individually have data points N_1 , N_2 , N_3 so on and so forth and for the i th data set, it is N_i . So, these individual samples will also have their means, which will be given by the \bar{X}_1 , \bar{X}_2 , \bar{X}_3 so on and so forth.

And since these are samples, they will also have their own standard deviation values, which we denote as S_1 , S_2 , and S_3 up to S_i for the i th sample and when we square the standard deviation, we get the variance of the data sets. So with these points in our hand, how do we conduct an analysis of variance.

(Refer Slide Time: 02:36)



1. Calculate the **Grand Average** of the data set. It is a **weighted average** of the individual sample means.
2. Obtain the square of **deviations between the group means and the grand average**. This is called **Sum of Squares due to Factor (SSF)**. Degrees of Freedom = **$I - 1$**
3. Obtain the square of **deviations within the groups**. This is called the **Sum of Squares due to errors (SSE)**. Degrees of Freedom = **$N - I$**
4. Calculate **Sum of Squares Total (SST)** = $SSF + SSE$. Degrees of Freedom = **$N - 1$**
5. Calculate the **Mean Square Factor (MSF)** = $SSF / (I - 1)$. MSF is an estimate of variance due to error plus variance between the groups i.e. $\sigma_e^2 + \sigma_f^2$.
6. Calculate the **Mean Square Error (MSE)** = $SSE / (N - I)$. MSE estimates only the variance due to error, i.e. σ_e^2 .
7. Calculate $F_{\text{test}} = MSF / MSE$ and compare with F_{crit} .



So, that is shown in the next slide, which you can see now and what we need to do is, we need to calculate what is known as the Grand Average, and I will come to it to see how it needs to be calculated, but for you it is important to understand that the Grand Average is actually a weighted average of individual sample means. So this is essentially an average of the whole data set and in the second step, what we do is, we obtain the square of deviations between the group means and the grand average.

So this will be called the Sum of Squares due to Factors. Now, let us go to the board and understand how these two things can be calculated. That is how Grand Average and the Sum of Squares due to Factor can be measured or calculated. So, going to the board, first let us see how Grand Average is to be calculated.

(Refer Slide Time: 03:45)

Grand Average $\bar{\bar{x}}$

$$\bar{\bar{x}} = \left(\frac{N_1}{N}\right)\bar{x}_1 + \left(\frac{N_2}{N}\right)\bar{x}_2 + \left(\frac{N_3}{N}\right)\bar{x}_3 + \dots + \left(\frac{N_I}{N}\right)\bar{x}_I$$

Total no. of data points that we are dealing with

2. Sum of Squares due to factors
SSF.

So, we denote the Grand Average by the symbol $\bar{\bar{x}}$ to indicate that it is the Grand Average and as we mention that $\bar{\bar{x}}$ will be a weighted average of the individual needs. So, it would be N_1/N where N is the total number of data points that we are dealing with. So, we keep doing this so as you can see that each of these are weighted average, weighted in terms of the number of data points within individual samples.

So, we calculate the Grand Average, towards the end this would be $N_i \text{ sample} / N \times \bar{x}_i$. So, this way we calculate our Grand Average, and the second part was the Sum of Squares due to factors, which we abbreviate as SSF. So, let us go to the next page and see how the Sum of Squares due to Factors needs to be calculated.

(Refer Slide Time: 06:22)

$$SSF = N_1(\bar{x}_1 - \bar{\bar{x}})^2 + N_2(\bar{x}_2 - \bar{\bar{x}})^2 + N_3(\bar{x}_3 - \bar{\bar{x}})^2 + \dots + N_I(\bar{x}_I - \bar{\bar{x}})^2$$

Square of deviations between the groups.

Degrees of freedom = $I - 1$

So what we have here is that we say that the Sum of Squares due to Factors will be the number of readings in the first sample. So, please note that it is a square factor, so each of the terms will be squared here and this is essentially the deviation from the Grand Average, which we did not take \bar{X} double bar from the individual sample means. So, we continue this and we get for the i th sample, this will be $N_i \times \bar{X}_i - \bar{X}$, which is the Grand Average and this is how we will be calculating the Sum of Squares due to Factors.

So, the Sum of Squares due to Factors considers the square of deviations between the groups. So, please keep this in mind and for Sum of Squares due to Factors, we will have the Degrees of Freedom as I , which is the number of samples which we are dealing minus 1. So, for example if we have 5 different samples, the Degrees of Freedom, would be $5 - 1$ is equal to 4. So, by doing this exercise, we will be able to calculate the Sum of Squares due to Factors. Now, let us go back to the desktop and see what we need to do next.

Next, we need to obtain what is known as the Sum of Squares due to variations within the group. This is nothing but the errors that we have been discussing in this course, or fluctuations within a particular sample that we are dealing with. So, this is called as Sum of Squares due to Errors, abbreviated as SSE, and here the Degrees of Freedom is N , which is the total number of readings minus I , which is the number of samples that we are dealing with. So, let us look at how the Sum of Squares due to Errors is calculated.

(Refer Slide Time: 09:48)

The image shows a blackboard with handwritten mathematical formulas. At the top left is the NPTEL logo. The main formula is:

$$SSE = \sum_{i=1}^{N_1} (x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{N_2} (x_{2j} - \bar{x}_2)^2 + \sum_{k=1}^{N_3} (x_{3k} - \bar{x}_3)^2 + \dots + \sum_{g=1}^{N_I} (x_{Ig} - \bar{x}_I)^2$$

Below this, the formula for Degrees of Freedom is written as:

$$\text{Degrees of Freedom} = N - I$$

Arrows point from the text below to the variables in the formula: "Total no. of readings" points to N , and "No. of data sets or samples present" points to I .

So, Sum of Squares due to errors will be calculated as the summation of deviations in the first group from the mean value of sample 1, squared, because it is a squared term here and this

continues, so, i then j , is equal to N^2 because this is the second sample, and for the third sample, similarly we will have $X_{3k} - \bar{X}_3$ squared, so on and so forth and finally, we will have it for our I th sample, let us call it $y=1$ to N_i x summated as $X_{iy} - \bar{X}_i$.

So, we calculate the individual deviations or errors in our individual samples and summate the square of it. So, from that we will get the Sum of Squares due to errors and as I already mentioned, the Degrees of Freedom here for Sum of Squares due to errors is nothing but the total number of readings $N - I$, which is the number of samples that we are dealing with. So, let us rewrite this, N is the total number of readings, and I is the number of data sets or samples present or being compared.

So, we calculate the Sum of Squares due to errors, then in the next step, we calculate the mean square values, so let us again go back to the desktop, and what we see here is that after we calculate the Sum of Squares due to Factors and Sum of Squares due to errors, we can calculate the Mean Square Factor which is nothing but the Sum of Square due to Factors, divided by degrees of Freedom SSF.

And in between, we have also seen that there is a term known as the Sum of Squares due to Total, which is essentially the total sum of squares and nothing but the Sum of Squares due to Factors plus Sum of Squares due to errors, and since it is a total of all the data sets and data points that we are dealing with. For this, the degrees of freedom is nothing but $N-1$, where again N is the total number of data points that we are dealing with.

So, it is important to note here that the Mean Square Factor, which we denote or abbreviate as MSF is an estimate of variance that comes due to the errors within the samples plus the variance between the groups. So, as I have been emphasizing that in Analysis Of Variance, we will be comparing these two things that is the errors within the sample versus the errors between the samples

So, MSF or the Mean Square Factor takes into account, both the variances coming from the error plus the variance between the groups, which we have denoted as σ_E squared, which is the variance due to errors + σ_F squared, which is the variance due to Factor under consideration. Now, let us see how do we calculate the Mean Square Error. It is again

very straightforward, that you have the SSE, which is the Sum of Squares due to errors divided by the degrees of freedom for the Sum of Squares due to errors, which is N-1.

So, as discussed previously, the mean square error, which we abbreviate as MSE, estimates only the variance due to the errors present within a particular sample that is nothing but the sigma E squared value. Now, when we have the MSF and the MSE calculated, we will then use these terms to calculate what is our F-test value, which is nothing but a ratio of MSF and MSE, and then as previously, we will be comparing it with a particular critical value of F, which will tell us whether the null hypothesis holds here or not.

(Refer Slide Time: 16:19)

The slide displays an ANOVA table with the following structure:

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

Handwritten notes on the slide include: "F_{test} vs. F_{crit} (γ₁, γ₂, 95% CL)" with an arrow pointing to the F column. The presenter's video feed is visible in the bottom right corner.

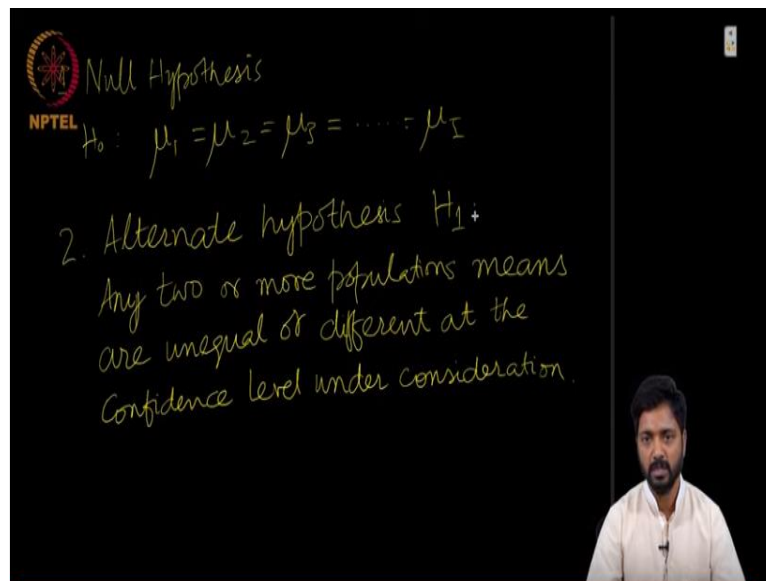
So, what we have to do after we have performed our ANOVA analysis, is to prepare what is known as ANOVA table where we summarize our calculations based on the Source of Variation, which is in the case of Factor Effect between the groups and also the variation that is present within a group, which is called the Error Effect. Then, in this column, we tabulate the Sum of Square.

So, for the Factor Effect, this would be Sum of Squares due to Factors, and for the Error Effect, this would be Sum of Squares due to Error, and obviously when we total both of these, when we add both of these, we get Sum of Squares due to Total and as we discussed in the previous slide, that for the Sum of Squares due to Factor, the Degrees of Freedom is nothing but the number of samples – 1.

Similarly, for Sum of Squares due to Errors, it is the total number of measurements minus the total number of samples that we are dealing with, that is $N-I$ and for the Sum of Squares due to Total, this is simply $N-1$, where N is the Total number of data points, that we are dealing with. So, based on these we can calculate the MSF value, the Mean Square Factor value as well as the Mean Square Error value, and we calculate the F-test value as a ratio of the MSF/MSE.

So, once we plug in these values, our ANOVA table is ready and finally, we will compare the F-test versus the F-critical at 2-degrees of freedom, which will be the $I-1$ and $N-I$, at 95% and 99% Confidence Level. So, based on that we will be able to estimate whether our Null Hypothesis holds true or not. So, at this point, I would also want to explicitly state what our Null Hypothesis and Alternate Hypothesis would be for Analysis of Variance. So, going back to the board.

(Refer Slide Time: 19:26)



Tell what is the Null Hypothesis, which is nothing but the population means of the samples being compared are estimated to be or presumed to be similar. So, if you are dealing with I samples, all of these means are estimated to be similar. Now, it is important to note what the Null Hypothesis is contrasted to the Alternate Hypothesis. So the Alternate Hypothesis here is that any two or more populations means are unequal or different at the Confidence Level under consideration.

So, please note that in Analysis of Variance, we cannot tell precisely, which of the two or three or N number of samples will have the population means different. The Alternate

Hypothesis, which we denote as H_1 , is nothing but that any two or more population means are unequal. What we will do later is to use what is known as the least significant difference to estimate, which exact samples are actually different from each other. That would be part of the exercise for the next lecture, but let us continue with applying ANOVA on the real data set.