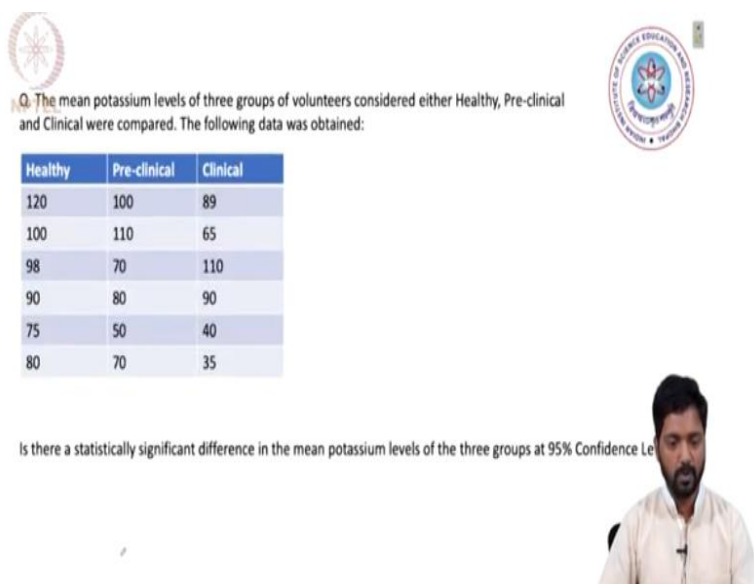


Quantitative Methods in Chemistry
Dr. Aasheesh Srivastava
Dr. Bharathwaj Sathyamoorthy
Department of Chemistry
Indian Institute of Science Education and Research - Bhopal

Lecture 20
Protocol for undertaking ANOVA - Part 02

(Refer Slide Time: 00:32)



Q. The mean potassium levels of three groups of volunteers considered either Healthy, Pre-clinical and Clinical were compared. The following data was obtained:

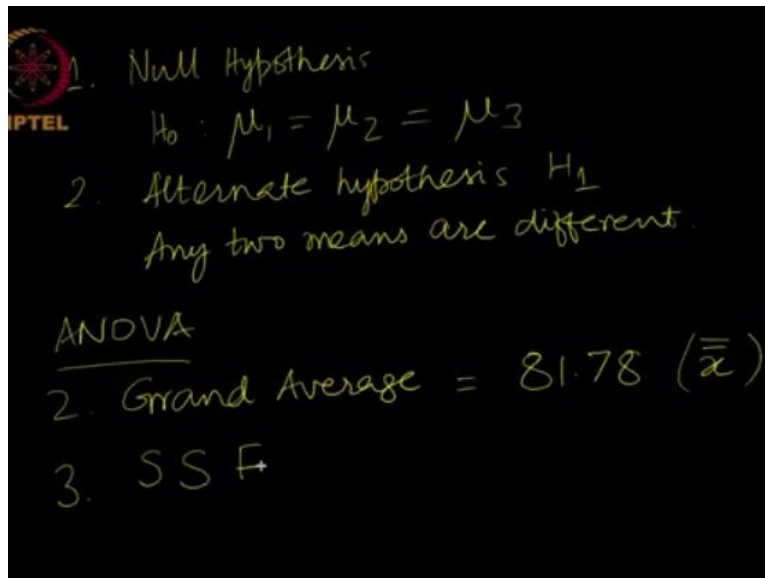
Healthy	Pre-clinical	Clinical
120	100	89
100	110	65
98	70	110
90	80	90
75	50	40
80	70	35

Is there a statistically significant difference in the mean potassium levels of the three groups at 95% Confidence Level?

Let us go back to the presentation and what I have here for you, is sort of a sample question where we are comparing the mean potassium levels of three groups of volunteers, which are considered either to be healthy, preclinical or clinical and their potassium levels are compared and the question that we would want to address is given below and that is, is there a statistically significant difference in the mean potassium levels of the three groups at 95% confidence level?

So we are having six readings in each group now. So in the healthy group, we have six readings, similarly in the preclinical and clinical and we want to address if there is a statistically significant difference in the mean potassium levels of these three groups. So we will perform the analysis of variance now on this data set to understand, how this calculation is to be executed.

(Refer Slide Time: 01:55)



So to begin with first of all state your null hypothesis. It is very important that in the beginning of this exercise or in the t-test the null hypothesis is explicitly stated and for us here a null hypothesis denoted by H_0 will be that the population means of three samples being compared are similar. The alternate hypothesis for us will be that any two means are different. It can be even any three of the means are different, but at least any two of the means are thought to be different in the alternate hypothesis H_1 .

Now let us execute the analysis of variance for this data set. First of all in the step 1, we have already told what the null and alternate hypothesis is, in the step 2 we calculate the grand average for this data set, which will be nothing but the sum of all the readings divided by the total number of readings and the grand average here turns out to be 81.78 and we denote it as $\bar{\bar{x}}$. Now in the third step, we will be calculating the sum of squares due to factors and for that let us go to the next page.

(Refer Slide Time: 04:15)


$$\begin{aligned}
 SSF &= N_1 (\bar{x}_1 - \bar{x})^2 + N_2 (\bar{x}_2 - \bar{x})^2 + N_3 (\bar{x}_3 - \bar{x})^2 \\
 &= 6(93.83 - 81.78)^2 + 6(80.00 - 81.78)^2 \\
 &\quad + 6(71.50 - 81.78)^2 \\
 &= 6 \times (12.05)^2 + 6(-1.78)^2 + 6(-10.28)^2 \\
 &= 6 \times 145.20 + 6 \times 3.168 + 6 \times 105.678 \\
 SSF &= 1524.3
 \end{aligned}$$

So sum of squares due to factors will be the mean values of the data sets, individual data sets. Let me put this in the previous page. So for our first data set \bar{x}_1 value turns out to be 93.83. Similarly for the second data set, the average value turns out to be 80.00 and finally for the third data set, which is the clinical sample, we have the average value of 71.50. So with these numbers in our hand, we calculate the sum of squares due to factors, which is nothing but $n_1 \bar{x}_1 - \bar{x}$ squared plus $n_2 \bar{x}_2 - \bar{x}$ squared plus $n_3 \bar{x}_3 - \bar{x}$ squared.

So when we plug in the numbers, we get 6 into 93.83 minus 81.78 squared plus similarly the six readings in the second data set and the mean in the second data set was 80.00, from which the grand average is subtracted, which is 81.78 squared plus again 6 into for the third data set the average value is 71.50 and from this the grand average value of 81.78 is to be subtracted and squared. So when we do this we get 6 into 12.05 squared plus 6 into minus 1.78 squared plus 6 into minus 10.28 squared and on solving this we get this finally gives us.

I will leave the rest of the calculation to you. This finally gives us the sum of squares due to factors as 1524.3 rounded off and next, we will calculate the sum of squares due to errors.

(Refer Slide Time: 07:34)



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


$$SSE = 8199$$

$$SSE = (N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2 + (N_3 - 1)SD_3^2$$

Which was nothing, but the summation of the deviations present within each sample. So for sample 1, i is equal to 1 to N1, for sample two j is equal to 1 to N2 and similarly for sample three k is equal to 1 to N3 and when we do this exercise, what we get is the value of sum of squares due to errors coming out as 8199. So it is important at this point to mention that the sum of squares due to errors, which is given as the deviations squared can also be written in a different manner.

And that is the degrees of freedom of individual samples multiplied by the standard deviation squared. How does this come about?

(Refer Slide Time: 09:55)



$$SD = \sqrt{\frac{\text{Dev.}^2}{N-1}}$$

$$\text{Dev.}^2 = (N-1) \times \underline{SD^2}$$

$$5. MSF = \frac{SSF}{I-1} = \frac{1524.3}{2} = 762.1$$

$$6. MSE = \frac{SSE}{N-I} = \frac{8199}{18-3} = 546.6$$

Now we know that the standard deviation is nothing but square root of deviation squared by the degrees of freedom and when we open it up, we get the deviation squared will be $N - 1$ into the standard deviation squared. So from this, we will be able to calculate the sum of squares due to errors by either utilizing the standard deviation values of individual data sets or manually calculating the individual deviation values.

So both the protocols will give us the same value of the sum of squares due to errors. Now in the next step, we need to calculate what are the, so let us see what that number is. So in the fourth step, we had already calculated the sum of squares due to errors and in the fifth step, we calculate the mean square factor, which is nothing but sum of squares due to factors divided by $i - 1$.

Now the sum of squares due to factors value for this particular data set was found to be 1524.3 divided by 2 here, because we are dealing with only three data, three samples and this comes out to be 762.1. We similarly calculate the mean square error and the mean square error is the sum of squares due to errors divided by the degrees of freedom, which is nothing but $N - I$. So plugging in the values of the sum of squares due to errors, we get 8199 divided by 18.

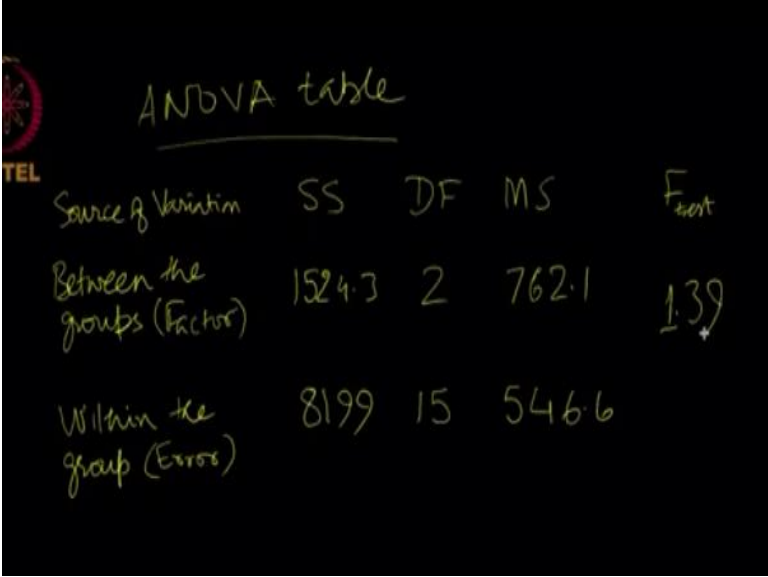
Which are the total number of points in the data set that we are dealing with minus 3 which is the number of samples that we are dealing with in this problem. So this finally solves to 546.6. Now since we have already calculated the mean square factor and the mean square error, it would be straight forward to calculate F value here using these two values.

(Refer Slide Time: 13:17)

And we have the table at 5% confidence level. So the table at 5% confidence level is now on the screens and we are talking about nu 1 value of two and nu 2 two value of 15 that means our F critical value at 5% significance level will be 3.68, which is encircled now. So let us go back to the board and put in the F critical value, which is 3.68. Now we can make an assessment of this data set.

We find that F test value is in fact less than the F critical value and that means that the null hypothesis cannot be rejected, that means we cannot say with confidence that there is any significant difference in the potassium levels of these three samples. So in other words, the healthy, preclinical and clinical samples have similar population mean values at 95% confidence level. Now let us quickly create the ANOVA table for this data set based on the calculations that we have done.

(Refer Slide Time: 16:55)



A handwritten ANOVA table on a blackboard. The title 'ANOVA table' is written in yellow. The table has five columns: 'Source of Variation', 'SS', 'DF', 'MS', and 'F_{test}'. The first row is 'Between the groups (Factor)' with values 1524.3, 2, 762.1, and 1.39. The second row is 'Within the group (Error)' with values 8199, 15, and 546.6.

Source of Variation	SS	DF	MS	F _{test}
Between the groups (Factor)	1524.3	2	762.1	1.39
Within the group (Error)	8199	15	546.6	

So the ANOVA table would look something like this. We will have the source of variation, then we will have the sum of squares values, then we will have the degrees of freedom, the mean square values and finally the F test or F calculated value. So the source of variation when it is between the groups, we are talking about the sum of squares due to factors and the sum of squares due to factors for this data set was 1524.3.

And the degrees of freedom associated with the sum of squares due to factors was 2, a mean square factor as a result 762.1. Similarly the source of variation when it is within the group, what we denote as the errors in the individual samples, the sum of squares due to errors was calculated to be 8199, the degrees of freedom associated with the sum of squares due to errors was 15, consequently the mean square value of error was found to be 546.6.

And finally the F test value turned out in our case to be 1.39. So this is how, you will be able to create ANOVA table for your data set. In the next class, we will be looking at how if we have F test value greater than F critical that is the population means are not similar between the samples being compared, how do we estimate which of the individual samples are actually different. For that we will be using the least significant difference concept and we will apply that to pin out or pinpoint which data sets or which samples are not similar in their population means. Thank you.