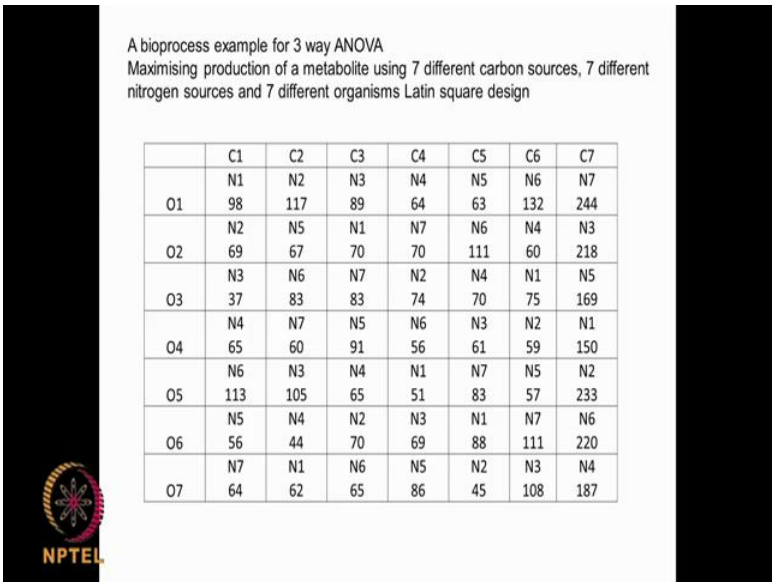
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 19
ANOVA

Welcome to the course on Biostatistics and Design of Experiments. We will continue on the topic of ANOVA. Let us look at one more problem.

(Refer Slide Time: 00:20)



	C1	C2	C3	C4	C5	C6	C7
	N1	N2	N3	N4	N5	N6	N7
O1	98	117	89	64	63	132	244
	N2	N5	N1	N7	N6	N4	N3
O2	69	67	70	70	111	60	218
	N3	N6	N7	N2	N4	N1	N5
O3	37	83	83	74	70	75	169
	N4	N7	N5	N6	N3	N2	N1
O4	65	60	91	56	61	59	150
	N6	N3	N4	N1	N7	N5	N2
O5	113	105	65	51	83	57	233
	N5	N4	N2	N3	N1	N7	N6
O6	56	44	70	69	88	111	220
	N7	N1	N6	N5	N2	N3	N4
O7	64	62	65	86	45	108	187

So, look at this problem, this has got three different parameters or we call it independent variables or we call it groups. Imagine I am doing a bioprocess experiment. This is a three way ANOVA. I want to maximize production of a metabolite. The three main effects are, carbon, nitrogen and organism. So, how do I go about doing it there? There are so, many ways suppose I change one parameter at a time and then I may get about may be 21 experiments 7, 7, 7 like that. But then that is not a very good approach, if you want to look at interactions you may have to change many parameters at a time, correct?.

So, there are many approaches, one is called factorial design, then another one is called fractional factorial design, then there are some second order designs and so on. But one of the designs is called the Latin square design. Look at this design. What is happening is I am putting the 7 carbon sources as heading for the column, and 7 organisms here. And

then, we are changing nitrogen in certain fashion. So, N1, N2, N3, N4, N5, N6, N7 and then you have different nitrogen for these oxygens and carbons. So, but the total number of N1s will come to be 1 in this column and total number of a N2 will come out to be 1, N3 will come out to be 1 and so on actually.

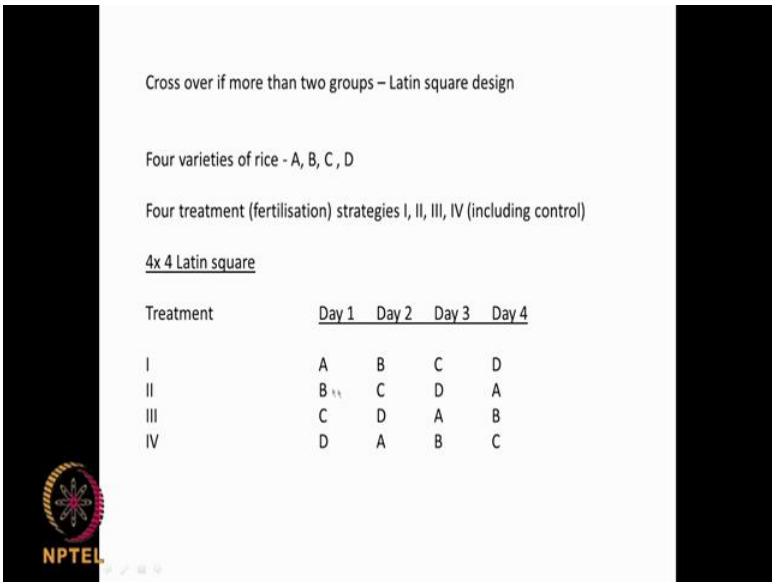
So, that way there is some sort of asymmetric. So, if you take this particular experiment, I will be doing with the carbon 1, it is called level one or it could be a carbon with that particular concentration, then nitrogen at a particular concentration and if the organism 1. Now suppose I take this particular experiment, I am doing a carbon at a particular level of 7, nitrogen at a particular level 4 with organism seven. So, these are the total number of experiments. So, what I am doing is, I am doing totally 49 experiments right? I am doing totally 49 experiments, and these are the results given here.

We are looking at a metabolite which we want to maximize, it could be in quantity of milligrams per liter here. So, we have it here 98. So when I use a certain amount of carbon level 1, nitrogen level 1 with organism one, I get 98 milligrams per liter of my product. So suppose if I take this, I am taking carbon at level 7, nitrogen at level 4 with organism 7, I get 187 milligrams per liter. So now I want to look at the main effects. What is the effect of carbon? What is the effect of nitrogen? What is the effect of organisms? And so on. Here, so we are doing 49 experiments.

If you remember long time back, I introduced concept of Latin square. We were doing Latin square design it is called 4 by 4 Latin square. We were looking at 4 varieties of rice A, B, C, D and 4 different fertilization strategies.

(Refer Slide Time: 03:44)

.



So, what do I do is, I take variety A, B, C, D I give a treatment 1, day one, day two, day three, day four consecutively. So, the treatment one is given to rice variety A on day one. On day two, treatment 1 is given to rice variety B. And on day three you give the treatment 1 to rice variety C. And on day four you give the treatment one to right variety D. So, on day one, A will be getting treatment 1, B will be getting treatment 2, C will be getting treatment 3 and D will be getting treatment 4. So this is 4 by 4 and because we have 4 different varieties of rice and 4 different treatment strategies, we have totally 16 blocks here. We can see this A, B, C, D and beauty is, it is very symmetrical. So if you add if you look at it there A will appear only once, B will appear only once. Similarly if you go this way A will appear only once, B will appear only once, C will appear only once. So, here you are doing 16 experiments.

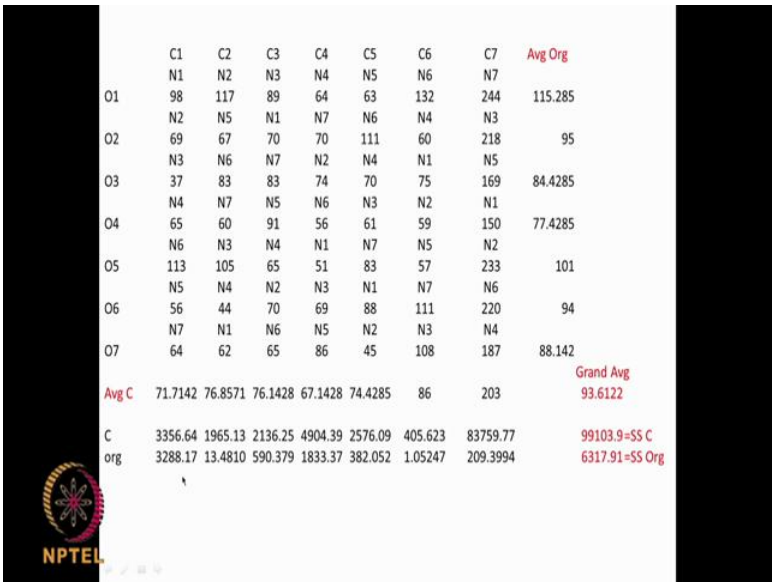
Now in our problem it is not really completely symmetric with respective A, B, C because the oxygen, sorry organism you are trying it out. All the organisms, carbons are being tried out. So the organism verses carbon is being tried out in different combinations. For example, organism 2 is being tried out on with carbon amount 1, carbon amount 2, carbon amount 3 and so on, but if you look at the nitrogen amount, it is not exactly as much. So, ideally if you want to do complete factorial, it will be like a cubical 7 X 7 X 7 type of that is a huge number. 7 X 7 is 49 and then when you multiply

again by 7 that is going to be a big number, right? So, that sort of approach we are not going to do whereas we are doing much less number of experiments. That means we are doing here only 49 experiments here in this particular case. And although the organism and the carbon amounts are being done in a full factorial that means, all the values, nitrogen is not completely being looked at.

For example, if you take nitrogen 1, you look at nitrogen 1 is being tried out only with the organism 1 at a carbon level 1. But it is not tried out at all carbon levels. But still it is very very useful because, you are going to do very less number of experiments, but you will get lot of live information on the main effects. Let us go back to the problem and this is called a Latin square design, we are looking at a carbon at 7 different levels or 7 different sources of carbon, nitrogen may be at 7 different sources of nitrogen like ammonium based, inorganic or organic, different types of nitrates, sulphates, phosphates and so on.

Carbon could be east media or LB media and so on organisms, different types organisms here. So, if we take 49 experiments, let us go back to ANOVA and see how to work at it. So, as you know you know how to do the sum of squares with the respect to the carbon, sum of squares with respect to the nitrogen, sum of squares which is respect to the organism, total sum of squares, you subtract each one of the individual, group sum of squares from total sum of squares, that will give you an error sum of squares. So, that is the general approach by which one goes about handling this problem. Let us look at the problem.

(Refer Slide Time: 07:35)



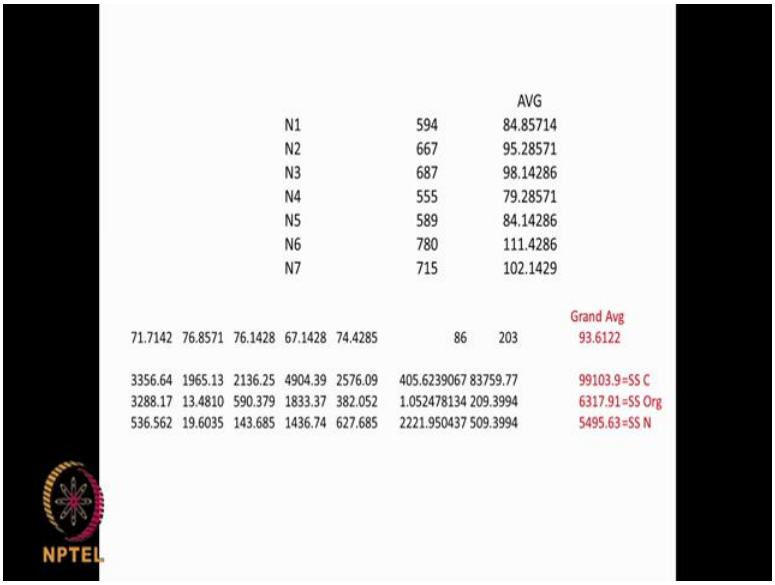
	C1	C2	C3	C4	C5	C6	C7	Avg Org	
	N1	N2	N3	N4	N5	N6	N7		
O1	98	117	89	64	63	132	244	115.285	
	N2	N5	N1	N7	N6	N4	N3		
O2	69	67	70	70	111	60	218	95	
	N3	N6	N7	N2	N4	N1	N5		
O3	37	83	83	74	70	75	169	84.4285	
	N4	N7	N5	N6	N3	N2	N1		
O4	65	60	91	56	61	59	150	77.4285	
	N6	N3	N4	N1	N7	N5	N2		
O5	113	105	65	51	83	57	233	101	
	N5	N4	N2	N3	N1	N7	N6		
O6	56	44	70	69	88	111	220	94	
	N7	N1	N6	N5	N2	N3	N4		
O7	64	62	65	86	45	108	187	88.142	
									Grand Avg
Avg C	71.7142	76.8571	76.1428	67.1428	74.4285	86	203		93.6122
C	3356.64	1965.13	2136.25	4904.39	2576.09	405.623	83759.77		99103.9=SS C
org	3288.17	13.4810	590.379	1833.37	382.052	1.05247	209.3994		6317.91=SS Org

So again go back to the problem here. So, if we look at the row, you will get an average this is the average for organism 1. This is the average for organism 2. That means, I am adding up all these terms and dividing by 7 and this is for the organism 3, I am adding up all these, dividing by 7, adding up all these dividing by 7, you got the hang of it. Now for carbon I would have just add up the columns and divide by 7. So, this is the average for carbon 1, this is the average for carbon 2, this is the average for carbon 3, 4, 5 and 6 and 7. Now understand, average for carbon, average for nitrogen. Now we can get a grand average, what is grand average? Average of all these averages. So, it could from the organism if I add up all and then divide by 7, I should get or from carbon,if I add up all these and divide by 7, I should be able to get that. So, straight forward this is the grand average. So, we have 7 values for average organism along the rows, 7 for carbon along the columns.

Now we need to get sum of squares for, we will, we will look at nitrogen in the next slide, but we need to get the sum of squares for carbon, sum of squares for the organisms. So, it is very straight forward. So, what do we do we multiply by 7 here because there are 7 items here, we take the grand average is $93.61 - 71.71^2$, 7 X because there are 7 items, $93.61 - 71.7^2$. So, next item $7 X 76.85 - 93.6^2$ this item $7 X 76.1 - 93.6^2$ and so on. This item $7 X 86 - 93.6^2$ and this item $93 - 203^2 X 7$. So, we add up all these things that will give you the sum of squares for carbon. Do you understand?

We have the averages for carbon, we subtract each one of the item, with the grand average, square it up every time, we had a multiply by 7 because there are 7 terms here. Now for the organisms what do we do? We again multiply by 7 because, there are 7 items here. 115.28 -grand average 93.61 ^{2}, 7 X 95 - 93.2 ^{2}, 7 X 84.4 - 93.6 ^{2},7 X 77.4 - 93.6 ^{2}, 7 X 101 - 93.6 ^{2} finally, 7 X 88.1 - 93.6 ^{2}. So, if you add up all these you will get the organism sum of squares. So, you know how to do the carbon sum of squares and now the organism sum of squares. Now let us look at the nitrogen sum of squares. So, it is very straight forward. So, we need to first get the 7 different averages for nitrogen.

(Refer Slide Time: 10:58)



So, nitrogen 1, how do you do that? Nitrogen 1 is here, nitrogen 1 is here, nitrogen 1 is here, nitrogen 1 is here, nitrogen 1 is here, nitrogen 1 is here, nitrogen 1 is here. So, you add up all these items, so it is not in a column or in a row, it is in different places. It is in a different places, understood?. So, we add up all these, divide by 7 that will give you nitrogen 1. Then nitrogen 2 here, here, here, here, here, here, here ,add up all these divide by 7. Nitrogen 3 so we have here, we have here, we have here, we have here, we have here, we have here and we have here, divide by 7 nitrogen 3. So, like that we can do for all the nitrogen. So, we will get 7 different nitrogen averages.

So, nitrogen averages are little bit tricky because the organism average is along the row, carbon average is along the column, but nitrogens are spread all over the place you see N1, N1,N1, N1, N1, N1, N1. So you have to pick up all these randomly put, it is not very

random because it is beautifully arranged. So, that as we can see in each row you will have only 1 nitrogen. So, it will not be two nitrogens, only one will come here. So, it is very beautifully done that is why it is called Latin square design.

We look at these type of designs later on when we look at design of experiments. So, we pickup like that and then add it up, sum it up and then divide by 7 that is how you do that for each of these nitrogen. So, nitrogen is little bit tricky.

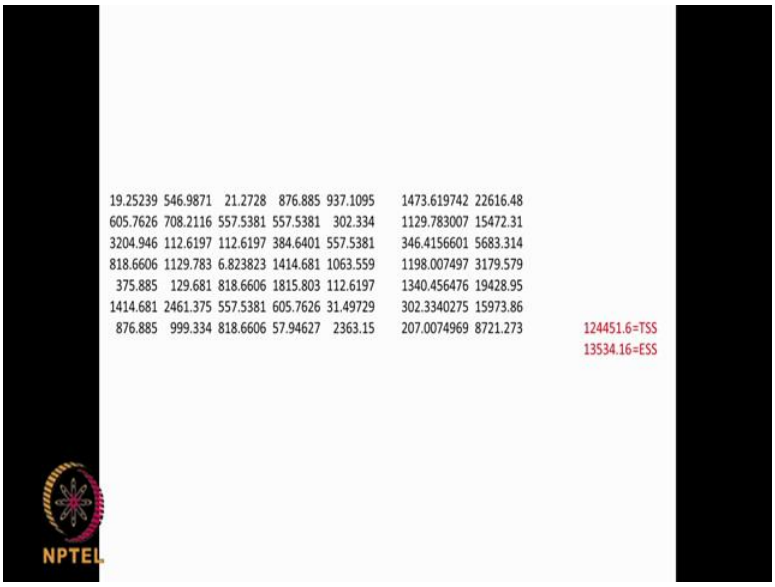
Now how do you do take the average, sum of squares for nitrogen? So we know the average, this is the average 84.8 and then you know the grand average, you know the grand average, so what do you do? So, you know the average, you know the grand average, so each one of them you subtract, square it up, multiply by 7, so you should get this particular row from nitrogen. So, each of the averages are like this 84.8 this column gives you the sum of nitrogens, then divide by 7 you get this. So, 84. So, how do you do? 84.85 or 86 - 93.6 multiplied by 7, 95.2 - 93.6^{2} multiply by 7, 98.1 - 93.6^{2} multiplied by 7. Like that you do and then you add up will get the sum of squares for nitrogen.

So, these are the main parameters or main groups or main independent variables different names I am giving. So, you did experiment by varying a carbon source or carbon concentration you did experiments by varying nitrogen source or nitrogen concentration, you did experiments by varying organism, 7 type of organism, 7 levels of carbons or 7 different carbon sources, 7 levels of nitrogen or 7 different types of nitrogen source. So, we got sum of squares for carbon, sum of squares for organism, sum of squares for nitrogen because these are the three main effects. So, look at the different terminologies by which I am addressing this. We can call it main effects, we can call it groups, we can call it independent variables. So, we have three independent variables- carbon nitrogen and organism or main effects are carbon nitrogen and organism principle effects carbon nitrogen and organism. These are the three independent variables in our problem. So, different names are there.

So, what is the dependent variable? your yield. The yield of your metabolite. So, we calculated the sum of squares for carbon, sum of squares for organisms, sum of squares for nitrogen. Now the question is how do you calculate total sum of squares? So, it is very simple total sum of squares we all know how to do that, right? We have the grand average, so, each item is subtract, square it up, you will get 49 different terms I add up all

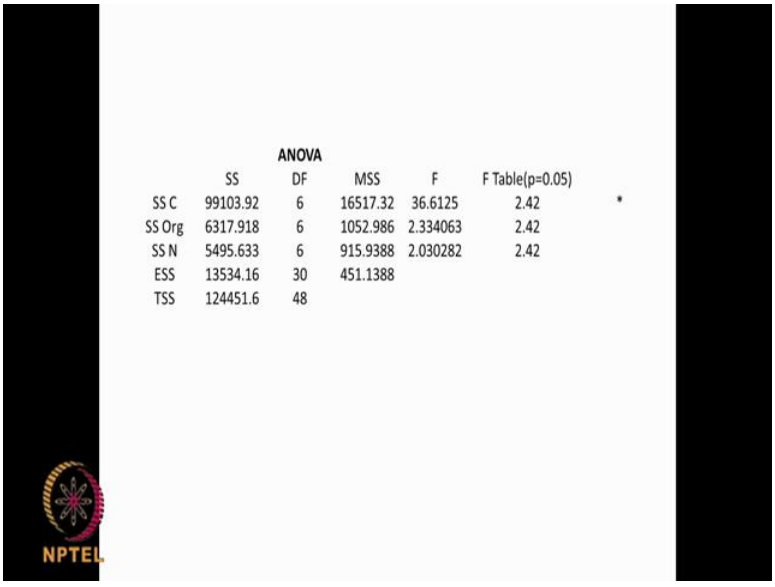
of them that will give you total sum of squares, do you understand?

(Refer Slide Time: 15:32)



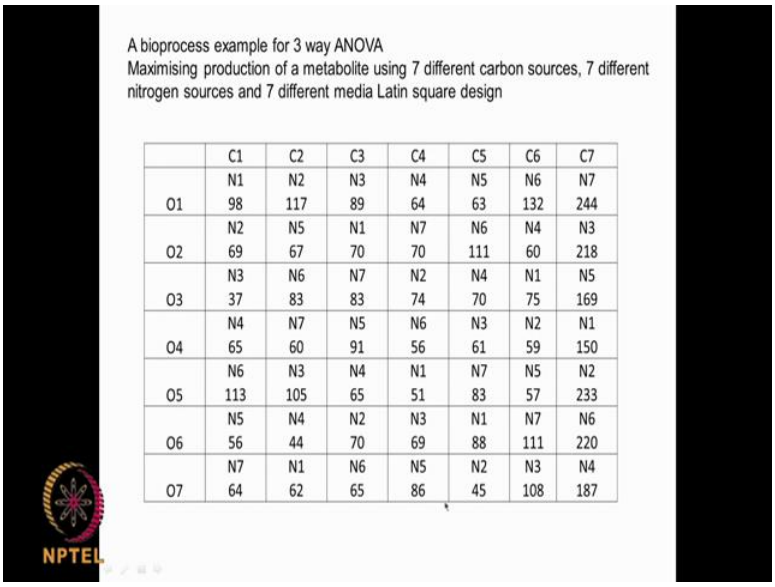
So, each one of these items we subtract, square it up, so that is the variation with respect to the, the grand average. Now how do you calculate the error some of squares? So, we have the principal effects or the main effects, three main effects carbon, organism, nitrogen. We have the total sum of squares, we subtract each one of them and we get something called error sum of squares. That is called the error sum of squares. Now will go the ANOVA table.

(Refer Slide Time: 16:05)



So we go to the ANOVA table, prepare your ANOVA table. Sum of squares for carbon 99103, 99103, sum of squares for organisms 6317.9, sum of squares for nitrogen 5495.6, total sum of squares is 124451. So, how do you get the error some of squares? We subtract each one of this from this, this is an indication of the error. Now there are totally 7 X 7, 49 data points. As you can see you member this right our table. 7 X 7, so it is a 49 data points.

(Refer Slide Time: 16:45)



So, we do a degrees of freedom for total sum of squares is 48. We have 7 different carbon sources or a carbon concentrations, so, degrees of freedom is 6. We have 7 different organisms, so degrees of freedom is 6. So 7 different nitrogens so we have degrees of a freedom is 6. So, error sum of squares 18 6 + 6 plus 6, 48 - 18 gives you 30. That is the error sum of squares. So, how do you get the mean sum of squares? You divide this by, ÷ 6 , ÷ 6, this ÷ 6, this ÷ 30. So, mean sum of squares how do you do the F table, sorry, F ratio? This divided by this gives you 36, this divided by this gives you 2.3, this divided by this gives you so these are the F values.

Now we need to compare with the F table for 6 into 30 degrees of freedom for a probability of 0.05. For a probability of 0.05. So, let us go to the table 6 and 30 degrees of freedom. So, 6 we have we have 30 we go down down down 2.42. So, 2.42 is your F table value, your F calculated value is 36.6, 2.33, 2.03. So, this is larger than the table value, so we can reject the null hypothesis, so we can say that the carbon have a

significant effect on the yield. We reject the null hypothesis, so we say carbon has a significant effect on the yield. So, we look at the average of these carbons. Look at there is know, 71, 76, 76, 67, 74. So, this carbon seven are, this big looks very big. So, this particular carbon gives you the maximum yield of your metabolite.

Now let us at the other terms, sorry organism. 2.33, 2.42. So, although we will say that it does not have an effect that means, organism there is no reason for you to reject the null hypothesis, but still it is pretty close you know, you understand? 2.33 and 2.44, 2 is very far. So, I may look at it more in detail, may be carry out more experiments and so on. So, although first hand we can say organism does not have an effect because there is no reason for you to reject the null hypothesis, but still it is not very far. So, you better watch out, may be do more experiments to get more data.

So, let us go back to the organisms. So, this gives you 115, 95, 84, 77, 101, 94. So, all those statistic, statistically the table value is higher than the calculated value, so you do not reject the null hypothesis, this organism gives you best so far in this list. So, we may do more experiments to be sure that organisms whether they play or do not play an effect.

Now let us go to the nitrogen. Nitrogen is 2.03, 2.42. So, we can say nitrogen does not play a role, but again they are not very far away. So, we may go and look into the nitrogen, these are the various nitrogen. Look at these averages this nitrogen 6 seem to be giving a larger number when compared to these numbers. So, one may try to do more experiments with nitrogen six to be sure, whether to accept the null hypothesis or reject the null hypothesis. Do you understand?

So, on first hand we can say only the carbon has an effect because the table value is less than the calculated value. So, only carbon has a significant role; that means, we reject the null hypothesis and we accept the alternate hypothesis. Organism does not seem to play a role, because F calculated is less than the table value so, but then as I am saying then differences are not very large. So, one may try to do more experiments to be sure about it and on just by a casual look, this organism is giving the highest number of the metabolite. So, one may do more experiments to be sure about it and same thing with nitrogen, either you can say nitrogen does not play a role or 2.03 and 2.42 are not very far. So, do I again look at the nitrogen data and say this looks like giving maximum, so

maybe I do more experiments to be sure whether categorically I will accept the null hypothesis or reject the null hypothesis. But if you go very strictly by the statistical rules we reject the null hypothesis only for the carbon where as we accept the null hypothesis for the organism and nitrogen.

Now another interesting point we need to consider is, can I look at interactions?. The problem is we will not be able to look at interactions. Why? I mentioned it long time back. So, if I want to look at say for example, carbon-oxygen-organism interaction and I have 6 degrees 6 degrees. So, carbon-organism will become 36 degrees of freedom, but I have only 30 for error, so I will not be able to look at the interactions. That means my number of experiments are not sufficient. Suppose if I had repeated the experiments, if I had replicated the experiments, not repeated, please note you should not use a word repeat although, we in English, we use causally repeat, we replicate the experiments twice; that means, each of the experiments we, from the beginning we completely do it with these combinations two times, this combination is two times. So, you will end up having 49 X 2 experiments. So, your degrees of freedom could in will be 98 then for interaction of carbon-organism 36. So, even if you subtract you will get some error degree of freedom, but as of now because we have not replicated the experiments. If I cannot look at say carbon-organism interactions because, that will require thirty six degrees of freedom, I have only 30 degrees of freedom remember that.

In previous one class also I did talk about if you do not replicate you will not look at interactions. In same thing here also if you are not replicating, you will not be able to really see if you want to see say interaction of one these variables, because each of the interactions will require 36 degrees of freedom. So, we can just combine together and we will call it only the error, we will call it the error sum of squares in this particular case and we can get a good F value for the carbon, for the organism, for the nitrogen.

But if you want to really look at interactions then obviously, we may have to do replications and so on actually. So, you understand this problem its very interesting problem it is a 3 / 3 ANOVA problem and it is called a Latin square design. We are varying three variables- one is the carbon source, another is the nitrogen source, another is the organism. So, although there are 7, 7, 7 we are not doing 7 X 7 X 7 experiments. We are only doing 49 experiments. So, organisms and carbon completely is done in detail. Whereas the nitrogen is done only once, for each row we have only one nitrogen.

Otherwise if you want to do complete factorial design it will be like 7 X 7 X 7, like a cubic, right?. Whereas here it is like a square and that is why it is called Latin square design and nitrogen for example, appears only once in each of these row that way it is very symmetric. So, I do the experiments these are the results, this the carbon nitrogen and organism are called the principle effect or main effect.

We then looked at how to calculate the average for carbon, which is easy along the column, average for organism along the row. Once we have the global average, we subtract each one of the average, square it up, multiply by 7 because there are seven items, we get the sum of squares for carbon and organism, carbon and organism. But for nitrogen it is little bit tricky because nitrogen one is coming here, here, here, here, here, here and here. So, we need to take an average of these seven items which are spread in different place. For nitrogen 2 also we do the same thing we have to the average of this, this, this, this, this, this and this now like that. Then again we can take the sum of squares for nitrogen, the same way for the other two from by subtracting it from the global average, square it up and multiply by 7. Then total sum of squares is from each of the grand average or global average you subtract each of these item and square it up.

So if you go to our ANOVA table, you have the sum of squares for carbon, sum of squares for the organisms, sum of squares for the nitrogen and then you have the total sum of squares then subtract each one of these main effects from the total, you get the error sum of squares, degrees of freedom 49 experiments, so you have 48 degrees of freedom you subtract each one of these 6 - 6 - 6 you get 30, why 6?, we have 7 carbon sources, 7 nitrogen sources and seven nitrogen organisms. Then mean you know how to do this divided by this, this divided by this, this divided by this this, divided by 30. Now the F value, you every each of these divided by error sum of squares. So, F table for 6 and 30 degrees of freedom comes out to be 2.42. So, it is very clearly showing that, carbon we can reject the null hypothesis accept the alternate, so there is a variation in the carbon sources. But we cannot reject the organism and nitrogen. But then on second thought if you look at these numbers 2.33 and 2.42, we can say they are very close so obviously, I need to do more experiments to be 100 % sure, whether I completely accept the null hypothesis or not and same thing here also 2.03 and 2.42 is not very far away. So, do I do more experiments on the nitrogen to be very sure to accept or reject the null hypothesis.

So, we looked at very interesting 7 by 7 Latin square design it is a three way ANOVA. But we are not able to see interactions because we did not do any replicates. So, the number of degrees of freedom is also very less if you want to bring in the interaction effects also. We will continue with the ANOVA in the next class also.

Key words: Factorial design, fractional factorial design, independent variables, ANOVA table, Latin square design, three-way ANOVA, F value, Principal effect, F table, Interaction effect, error, sum of squares, replication, error sum of squares.