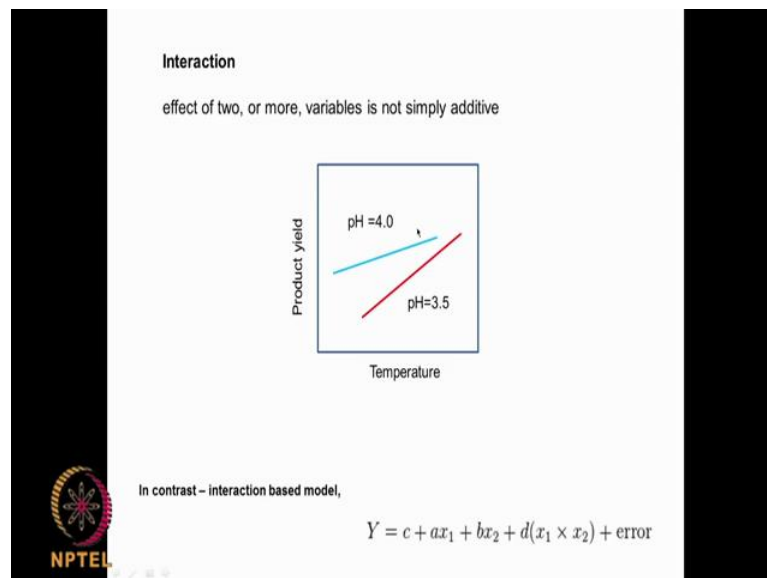


Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 18
ANOVA

Welcome to the course on Biostatistics and Design of Experiments. We are talking about ANOVA, two-way ANOVA, and the most important point which I introduced in the previous situation is interaction between two variables or two main effects or two independent variables.

(Refer Slide Time: 00:27)




So, if it is not an additive, suppose when I am changing temperature, and when I am changing pH, at one particular value of pH it appears to go like this. In another value of pH instead of going almost parallel, it seems to be sort of converging. So obviously, there is an interaction between pH and temperature. Or you may have a situation like this. At one pH, the yield goes up as I change temperature; but at another pH, yield is dramatically raising. Obviously, there is an interaction between pH and temperature. So, interactions are very important I said and we need to study interactions; only when you do the replications, you will be able to study interactions. Without replication of data, you will not be able to study interactions. Let us look at a one more problem, which again looks at two-way ANOVA; it is called two-way ANOVA.

(Refer Slide Time: 01:25)

Production of AHL, a quorum sensing molecule is being optimised using four media and five organisms. The experiments were replicated twice and results of AHL production (in mg/L) is reported.. Perform a Two way ANOVA with interaction

Organism	Media			
	A	B	C	D
1	20,20	19,20	23,22	18,16
2	18,20	18,19	20,21	15,17
3	19,19	20,18	20,22	18,19
4	22,24	20,22	25,26	21,21
5	19,20	20,18	22,20	18,18



So, here, in this particular problem, we are looking at the production of Quorum sensing molecule. Quorum sensing is a very very important concept in bacterial biofilm formation. So, the bacteria goes and attaches to surfaces, and it could be any surface. It could be an inert material, synthetic material, biological material. So, when the bacteria settles, and it produces **EPS** and proteins, carbohydrates, they form layer which is called biofilm. And biofilm is very very important for the growth of microorganisms and so on actually.

And a molecule AHL, Acyl Homoserine Lactone, is involved in identifying the Quorum. So, bacteria produce this AHL, and when there is sufficient amount of AHL that is observed by the bacteria, they know that the amount of bacterial is high. So, they go into a different phynotype and they start forming biofilms. So, that way it is called a signaling molecule, AHL. And AHL has many implications. So, if I can prevent AHL, I may be able to prevent the biofilm formation, the AHL of one bacteria could be toxic to another bacteria. That is more like a defense mechanism and so on actually. There is some interest in looking at bacterial AHL production and use this as an anti bacterial compound and so on actually.

So, production of AHL, this quorum sensing molecule is being optimized using four different media. Media A, B, C, D and five different organisms. Many organisms produce, as I said, lactones are formed as AHL signaling molecule. In some cases, there

are peptides that are formed. And then, even in the lactones, there are different types of lactones that are produced by these organisms. Imagine you are testing five organisms and four different media for AHL production and you **get**, are doing a replicate. Obviously, you have two sets of results, the amount of AHL produced in milligrams per liter. Now you have been asked to perform a two-way ANOVA.

Basically you have 40 data points. So, totally 39 degrees of freedom. Now, let us put it down in the usual form. We have the replicate data also here, replicate data for each one of the organism also here. So we have totally 40 data points.

(Refer Slide Time: 04:19)

	A	B	C	D	Org AVG	
Org1	20	19	23	18		
	20	20	22	16	19.75	
Org2	18	18	20	15		
	20	19	21	17	18.5	
Org3	19	20	20	18		
	19	18	22	19	19.375	
Org4	22	20	25	21		
	24	22	26	21	22.625	
Org5	19	20	22	18		
	20	18	20	18	19.375	
Media Avg	20.1	19.4	22.1	18.1		19.925=Global Avg
Media	0.30625	2.75625	47.30625	33.30625		83.675=Media SS
Org	0.245	16.245	2.42	58.32	2.42	79.65=Org SS
Total	0.005625	0.855625	9.455625	3.705625		
	0.005625	0.005625	4.305625	15.40563		
	3.705625	3.705625	0.005625	24.25563		
	0.005625	0.855625	1.155625	8.555625		
	0.855625	0.005625	0.005625	3.705625		
	0.855625	3.705625	4.305625	0.855625		
	4.305625	0.005625	25.75563	1.155625		
	16.60563	4.305625	36.90563	1.155625		
	0.855625	0.005625	4.305625	3.705625		
	0.005625	3.705625	0.005625	3.705625		196.775=TSS

Now, we can get a media average; just take the average of all these columns - that is called the media average. So, for media A, the average comes out to be 20.1; media B, 19.4; media C, 22.1; media D, 18.1. Similarly, for organism also we can get an average if you go this way. Again 10 data points for organism 1, 10 data points for organism 2. So we get the average for organism. So we can get the media average, we can get the average for organism. We can get the global average; that means, average of all averages. So we get the global average is 19.925.

Now, I need to find out media sum of squares. You all know how to do it, right? We have done it now many times. So, what do I do? So, I have average for each one of the media. So, what do I do? I subtract from this global, square it up, and multiply by 10. Why? Because there are 10 data points here. $10 \times 20.1 - 19.925^2$, $10 \times 19.4 - 19.925^2$, and so

on. So for the media you get 4. You add up all of them; that gives you media sum of squares. Understand? That is called the media sum of squares. Now, let us go for organism sum of squares. So you have organism average here; five different averages. So what you do? You take 19.75 - global, square it up, and multiply by 10. Because you have 10 data points, simple, right? that is what I am doing here $10 \times 19.75 - 19.925$ and so on. So, here $10 \times 18.5 - 19.25^2$, and so on. $10 \times 19.375 - 19.25^2$, 10 into like that you know. $22.62 - 19.9^2$, $10 \times 19.37 - 19.9^2$. So for organism you will get 5, you add up all of that that will give you the organism sum of squares. So we get media sum of squares, we get organism sum of squares.

Now, total sum of squares, you have to take other sum of squares with respect to the \bar{X} . The global average is called the \bar{X} , right? So what do we do? $20 - 19.925^2$. Like that you know $20 - 19.925^2$, $18 - 19.925^2$, $20 - 19$. Like that you get this entire exactly here. And then you add up all of them, that will give you the total sum of square 196.775.

Now how do I get the error sum of squares? Now, you have every time you have replicated; so obviously, this is an indication of the error, right? So, if we take an average, and then the change difference from the average, square it up, that is an indication of the error. So, then, you can add up all of them.

(Refer Slide Time: 07:54)

	A	B	C	D	Org AVG					
Org1	20	19	23	18	18					
	20	20	22	16	19.75					
Org2	18	18	20	15						
	20	19	21	17	18.5					
Org3	19	20	20	18						
	19	18	22	19	19.375					
Org4	22	20	25	21						
	24	22	26	21	22.625					
Org5	19	20	22	18						
	20	18	20	18	19.375					
Repeat avg	20	19.5	22.5	17		Rept SS	0	0.25	0.25	1
	19	18.5	20.5	16			0	0.25	0.25	1
	19	19	21	18.5			1	0.25	0.25	1
	23	21	25.5	21			1	0.25	0.25	1
	19.5	19	21	18			0	1	1	0.25
							0	1	1	0.25
							1	1	0.25	0
							1	1	0.25	0
							0.25	1	1	Error SS
							0.25	1	1	0
										21.5

So, again I am showing the same data here. If I take the average, average of 20 X 20 is 20; 18 and 20 is 19. And then 19 and 19 is 19; 22 and 24 is 23; 19 and 20 is 19.5.

Similarly, here 19 and 20 is 19.5. So you will get five data set. These are average of each one of these groups. So, how do I take the sum of squares? So $20 - 20^2$, $+ 20 - 20^2$, $18 - 19^2$, $20 - 19^2$, $19 - 19^2$, $19 - 19^2$, $22 - 23^2$, $24 - 23^2$ like that you know. Because these are indications of the error, because these are deviations from the mean. So you have this box where you have repeated, repeated, repeated, and the deviation from the mean of this each group is the repeat sum of squares or error sum of squares. And if we add up all these, you will get the error sum of squares here, you understand?

So we get the total sum of squares. You know how to do the total sum of squares. You are subtracting from the global average; you know how to do the media sum of squares because we have four media, each media will have its own mean; you subtract from the global; square it up, multiply by 10, because 10 data point. Now media sum of squares, organism sum of squares. For each of the organism you get an average. In this case there are five organisms. So, each of these you subtract from the global, square it up, multiply by 10. Why 10? Again there are 10 data points in this particular situation. So, you add up all of them; that will give you the organism sum of squares. Total sum of squares, you subtract each one of the term with the global average, square it up, then add whole lot.


Now, how do you get the repeat? You have done it twice; each experiment is done twice. Obviously, there is an average, and the difference between the average and individual value square is a measure of variance of that particular box. Understand? I will call this a box; each one is a box; each one is a box here. Now, you add up all of them; that will give you the error sum of squares. So we have total sum of squares; we have the media sum of squares; we have the organism sum of squares; we have the error or repeat sum of squares. If you add all these four and subtract from total sum of squares that should give us interaction sum of squares. So, we have four, we have four media, so we will get 3 degrees of freedom.

(Refer Slide Time: 10:44)

ANOVA						
	SS	DF	MSS	Fixed Effect model	Random / Mixed effect model	F table (p=0.05)
Media	83.675	3	27.89	25.945	28.008	3.1
Organism	79.65	4	19.912	18.52	19.995	2.87
Organims*Media	11.95	12	0.995	0.9263	0.9263	2.28
Error	21.5	20	1.075			
Total	196.775	39				

In Fixed effect all MSS are divided by Error MSS

In Random/Mixed model Interaction SS is divided by ErrorSS but Main effects are divided by InteractionSS



And media sum of squares as you see 83.675. So we put that as 83.675. There are five organisms, we have 4 degrees of freedom. Organism sum of squares is 79.65, 79.65; total is 196.775, 196.775. Degrees of freedom is 39 because we have five organisms and four media. So 5×4 it is 20 repeated each one twice. So, 20×2 is 40. $40 - 1$ is 39. So, organism into media you get it as 3×4 , 12. Now, error sum of squares we got it as 21.5, 21.5. We know total 39, media we know 3, organism is 4, organism into media is 3×4 , 12. So $12 + 4$ is 16, $16 + 3$ is 19, $39 - 19$ is 20. This is how we calculate the degrees of freedom.

Now, each of the sum of squares we divide by the degrees of freedom to get mean sum of squares. $83.6 / 3$, $79.6 / 4$, $11.9 / 12$, $21.5 / 20$. You understand? Now, how do I calculate F value? Here I am introducing something different. I am talking about fixed effect model, and random or mixed effect model. So, we have been all the time dividing by error, right? That is called a fixed effect model. So $27.89 / 1.075$, $19.9 / 1.075$, $0.995 / 1.075$. So, these are the F ratios.

Now, in the random or mixed effect model, what do you do is, the interaction is divided by error, but main effects are divided by interaction. So, in the random or mixed effect $27.89 / 0.99$, $19.91 / 0.99$ and $0.99 / 1.075$. So there is a difference between fixed effect model and the random effect model. In the fixed effect model, you divide every MSS with the error. Whereas in the random effect model only the organism media, that is

interaction effect, you divide by error, where as the main effects are divided by the interaction effect. Now, we go to the F table, we have 3 and 20 degrees of freedom or 4 and 20 degrees of freedom for 0.095 3 and 20 degrees of freedom 3.1, 3.1. 4 and 20 degrees of freedom, 4 and 20 degrees of freedom is 2.87, 2.87 or even 12 X 20 degrees of freedom.

(Refer Slide Time: 13:39)

F table for p=0.05

v2	DEGREE OF NUMERATOR (v1)									
	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

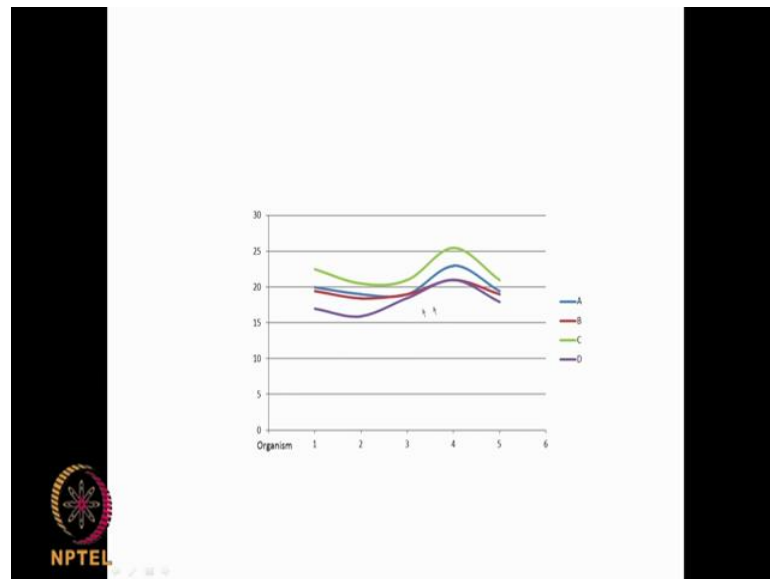
Next page 12 X 20 degrees of freedom is 2.828. These are actually valid for fixed effect model. For random effect model, i will be... it is not 3 X 20 degrees of freedom, it will be 3 X 12 degrees of freedom, because in random effect model you are dividing by the interaction; so it will be 3 X 12 degrees of freedom. So, in 3 X 12 degrees of freedom, if you are looking at, you will use 3.49. So, this F table for random effect model will be 3.49. And 4 X 12 will be 3.26. 3.49 and 3.26. Do you understand? Because your degrees of freedom have gone down. So 3.26 and 4.2.

So, if you look at this say fixed effect model, you see that the F value which you calculate from ANOVA as against a table value. So media has a very strong effect. So obviously, there we take the alternate hypothesis media as a strong effect and then organism also has a strong effect here. Whereas, interaction does not have any effect; whether it is fixed effect or the random effect model. So, out of these if you look at all these data, we have these four. We have these four different media, we have these five different organisms. So, this particular media C seems to be producing largest amount of

this particular molecule. And similarly, organism four seems to be producing largest amount of this molecule.

So, as you can see, it is very difficult to use two sample T test. So we have to solve it by ANOVA and ANOVA also gives this two-way ANOVA. We are looking at the main, principal effects on the organism and media and interaction which we saw, that is the interaction between the organism and media. In this particular case, we do not see any interaction because these calculated value is much smaller than the table value. Understand? So, if you plot all these data, you see that different media, different organism 1, 2, 3, 4, 5.

(Refer Slide Time: 16:29)



So, they are almost like parallel, you know. So, they are not crisscrossing or going up and down. So all the organisms approximately perform in the additive way towards these media.

If you compare this particular graph as against your previous problem if you remember. They were going almost like this, you know, two drugs, and male and female. THZ is performing like this; whereas, your Metformin was performing like this. So Metformin was performing much higher effect on female when compared to on the male. So both the drugs are almost similar, little small differences on male, whereas Metformin was performing very very differently and compared to THZ on female. So there was an interaction. The lines were not parallel.

Whereas in this particular case, you see that the lines are **parallel**, almost parallel. This is the organism 1, 2, 3, 4, 5, and these are the different media, as you can see here, different media. So, the effects are additive between the media and the organism, and it is not a non-linear type of dependence or an interaction type of dependence.

Now, the important point here is we had enough data sets, so we are able to understand interaction also. Suppose, I did not do a replication. I did only one experiment in each case. What happens? So I have done only one experiment in each case. I did not do a replication. So, total number of data points are 20. So my degrees of freedom for total is 19.

(Refer Slide Time: 18:20)

Suppose we did not repeat the experiments				
	A	B	C	D
Org1	20	19	23	18
Org2	18	18	20	15
Org3	19	20	20	18
Org4	22	20	25	21
Org5	20	18	20	18

	DF
Media	3
Organism	4
Organisms*Media	12
Error	0
Total	19

Media there are 4 media, no change. So, degrees of freedom is 3, organism 5, organism degrees of freedom is 4. So, if you want to study **organism into media 4 X 3, 12. So, 12 + 4 is 16**, 3 error becomes degrees of freedom is 0. That means, I cannot understand error. So, the interaction should be called as error. Then it becomes easy for me to calculate F value. So, do you understand? So, interaction gets mixed up with the error or interaction gets confounded with the error. I am introducing a new term called confounding. So, interaction becomes confounded with the error. So we cannot separate out error and interaction, that is media and organism. Because we do not have... if I take the interaction as 12 degrees of freedom that is **3 X 4**, error becomes 0. Because you did not have enough degrees of freedom. Or you call these error, then you cannot study the

interaction. Say either you call this error, then you cannot study interaction because you have only 19 degrees of freedom.

So, if you repeat the whole lot, when like in the previous case I showed you, if you repeat the whole, what do we have? We have 40 data points. So obviously, your total is 39, media remains same 3, organism remains same 4, **organism into media** interaction becomes 12. So obviously, error **12 + 4** is 16, **+ 3** 19, **39 - 19** is 20. You understand? So, this problem also shows you the importance of replicating the experiment. If we do not replicate, it is not possible for you to look at interaction effects. So, you need to always replicate your data, so that you can study the interaction effect also. Otherwise, interaction effect gets confounded into error. You will not have sufficient degrees of freedom to perform these type of calculations and that becomes a big problem actually. So, we cannot talk about interactions at all.

So, we looked at different types ANOVA so far. We looked at one-way ANOVA, then we looked at two-way ANOVA, two-way ANOVA with replication, without replication, then we looked at two-way ANOVA with interaction. That means, the principal effects are getting non-linearly associated with each other. It is not a simple additive type of relationship, but it is a non-linear relationship, that is called an interaction effect. In some cases, the interaction could be positive; that means, $x_1 x_2$ could be positive; in some cases, $x_1 x_2$ could be negative also; both the possibilities are there, and that is where the interaction plays a very important role.

If you want to really find out effects for interaction generally what do we do? If we have the effect, principal effect **1** with certain degrees of freedom DF 1, principal effect 2 certain degrees of freedom DF 2, the interaction effect will be **DF 1 X DF 2**. So, when I am calculating the error degrees of freedom by subtracting all these degrees of freedom from the total, I should have at least 1 or 2 degrees of...1 degrees of freedom for error. Then only I do F calculation. Because the sum of squares of error means sum of squares of error comes in the denominator. So, I should have some number and if I want to calculate mean sum of squares for error, obviously I should have some degrees of freedom, at least one degree of freedom; otherwise it is not possible to calculate mean sum of squares for error.

So, Replication is a must. I should have enough data points. So that I get degrees of

freedom for each one of these terms. And that is why I showed if I am doing it twice. If I am doing it twice, each of the experiment I get sufficient degrees of freedom for error. Whereas if I am doing only once, then obviously, I am in big trouble, I am not able to get any degrees of freedom for error as I am showing here actually; as shown in this particular example.

(Refer Slide Time: 22:39) **this slide is not related to the present video, in this place**
(Refer Slide time : 18:20) has to be placed

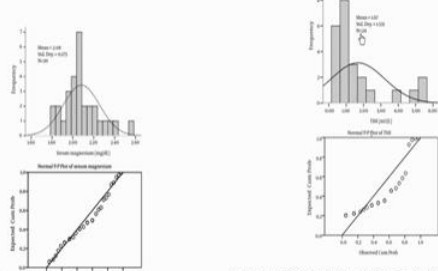
Test for Normality


The null hypothesis that the data come from a normally distributed population

The test results indicate whether one should reject or fail to reject the null hypothesis

Types of normality tests

Anderson-Darling test compares the empirical cumulative distribution function of the sample data with the distribution expected if the data were normal. If the observed difference is adequately large, reject the null hypothesis



 NPTEL

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/>

New slide

Suppose we did not repeat the experiments

	A	B	C	D
Org1	20	19	23	18
Org2	18	18	20	15
Org3	19	20	20	18
Org4	22	20	25	21
Org5	20	18	20	18

	DF
Media	3
Organism	4
Organims*Media	12
Error	0
Total	19

So obviously, Replication plays a very important role. And another thing is, higher the number of degrees of freedom for error because when you calculate the mean sum of squares of error, you are dividing by the degrees of freedom; so the means sum of squares of error becomes very very small. So, when that number becomes very very small that goes into the denominator when you are calculating F ratio. So obviously, the F values become very large. If the F values become very large with respect to the table value, then we can say that there is a statistically significant difference; that means, we will be able to reject null hypothesis.

If your degrees of freedom for error is very small, because it comes in the denominator for mean sum of squares for error, that term becomes large, and again, that comes in the denominator when you calculate F; so your F values become small; so obviously, those F values will be smaller than the table F value; that means, you will not have any reason for rejecting the null hypothesis. You will not be able to see effects of various factors whether its drugs or whether it is fermentation pH or fermentation temperature or carbon amount and so on, if your degrees of freedom for error is not sufficiently large or if you have too much of variations when you do a replicate experiment. So, in these two situations, you will not be able to see effect of various factors or effect of various independent variables. So, you need to keep these points in mind.

So, when I am doing a replication, I should be very sure that the variances or errors involved are small. And when as many replications are always good, because my degrees of freedom increases; that means, my degrees of freedom for error increases, that means mean sum of squares for error decreases; that means, my F values will increase and definitely it will be much higher than my table value.

Thank you very much. So, we will continue in the next class further concepts about the biostatistics.

Thank you for your time. Thank you.

Key words: independent variables, dependent variables, interaction effect, replication, sum of squares, confounding, error, non-linear relationship, replicate, F value, degree of freedom, mean sum of squares for error, global average, fixed effect model, random or mixed effect model