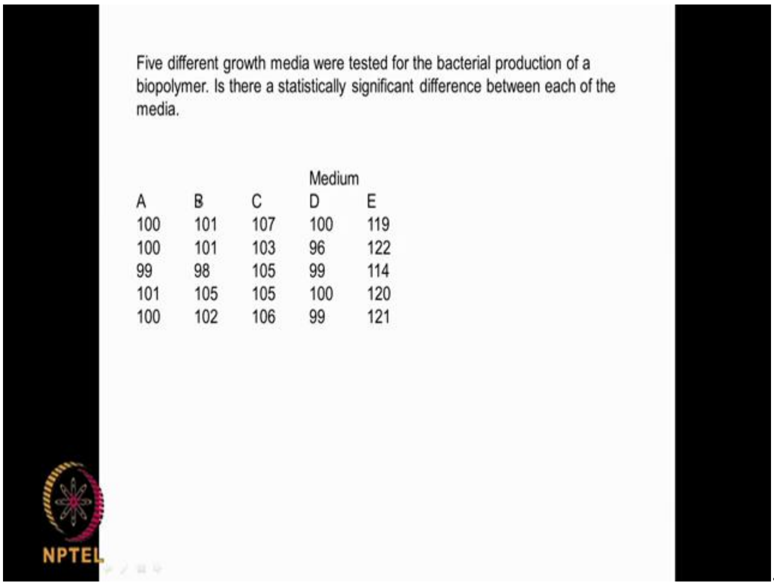
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 16
ANOVA

Welcome to the course on Biostatistics and Design of Experiments. We will continue with problems on ANOVA. I am spending lot of time on ANOVA, with examples, because that is a very, very important topic, and you should be able to do problems related to ANOVA without any difficulty. That is why I am trying to give as many worked out examples as possible with ANOVA. Because we are going to, now look at one-way ANOVA, later on we will look at two-way ANOVA, and then, we will look at interaction between those two different groups or two ways and so on actually. So, let us again continue. So, five different growth media were tested for bacterial production of a biopolymer.

(Refer Slide Time: 00:53)

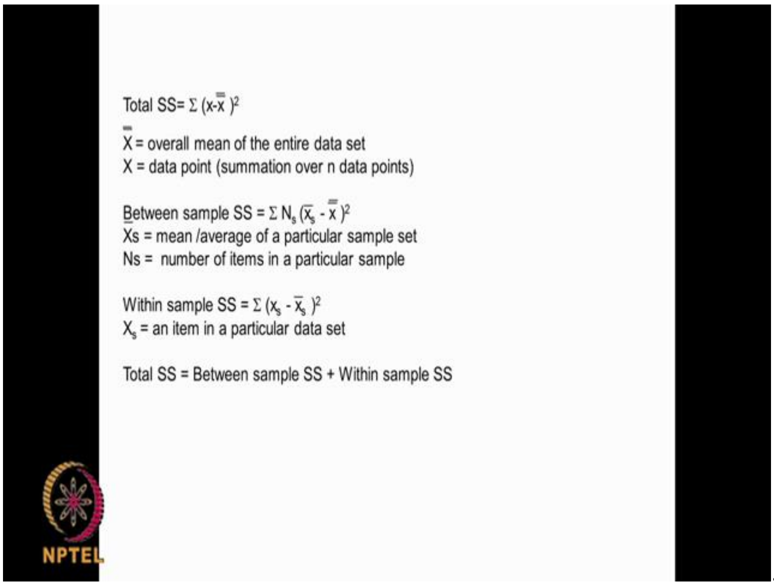


.

Because biopolymers like glucans are produced with bacteria, like you have linear glucan which are extra cellular, you can have cyclic glucan which are intra cellular; so lot of bacteria produces biopolymer. For examples PHB is a biopolymer. Just like the glucans PHB is a biopolymer; there are many biopolymers. So, we tested five different media for the production of these biopolymer and you can see the results. Five different experiments. So, you have five different experiments; five different media. Now we want to know these are the results of the biopolymer production. It could be in terms of say milligram per liter; so that comes to about 0.1 gram, it is quite low; whereas if take you linear glucans you may get in 30, 40, 50 grams per liter.

Some glucans are produced with very low quantities. So, 100 milligrams per liter is what you are seeing, we are having five different media. So, this is now called a one-way ANOVA. Because we are looking at only one way or one parameter or one group, but these are all repeats of this. How do you do it? So, we have talked quite a lot. We should be able to do this problem without any difficulty, but let me also workout once more.

(Refer Slide Time: 02:22)



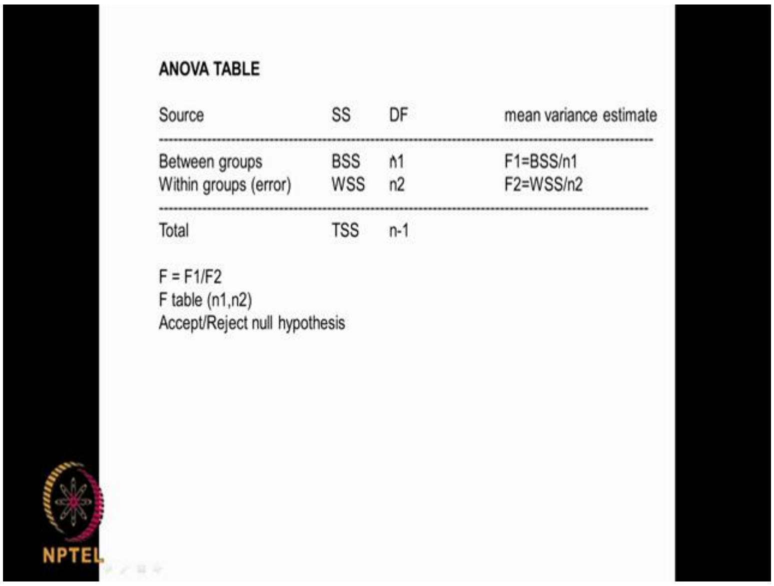
The total sum of squares is X; that means, each of these in this particular example we have 5, and we have 5, so 5 x 5, so each of these data - $\bar{\bar{X}}$; $\bar{\bar{X}}$is the overall average; I will call it the global average; $\bar{\bar{X}}$is a overall mean; that means you add up all the 25 data points, divide it by 25, you will get the overall average. So, this is called the total sum of squares. Now, between sample sum of squares we have five different situations, right. We have five different media. So there is going to be an average for each one of them. We have a global or total average; so subtract from each one of these average, square it up, then multiply by 5, because we have 5 elements in each one of this. Then, when you add up that will give you sample sum of squares. We have five sets of samples or five sets of groups or five sets of independent variables tested; so in this particular case we are going to multiply by 5. So for each one of them, there is going to be an average, and then, you subtract from the total overall average, square it up.

Now, there is something called within samples sum of squares. This is an indication of the error. Although you should get this as the value, you are going to get different value, right. For example, we have 100, 100, 99, 101, 100; so there will be an average. So, what is the difference squared from the average or mean? You square it up, for each one of them, you take the subtraction from the mean of that particular group, square it up, then add up all of them. So, when you do that, that is called within samples sum of squares, this is an indication of error.

So, ideally we should have this within sample sum of squares, as small as possible. So, that the between sample sum of squares look large. Because we are going to later on calculate the F value; that means, divide the sum of, the mean sum of squares of between divided by within.

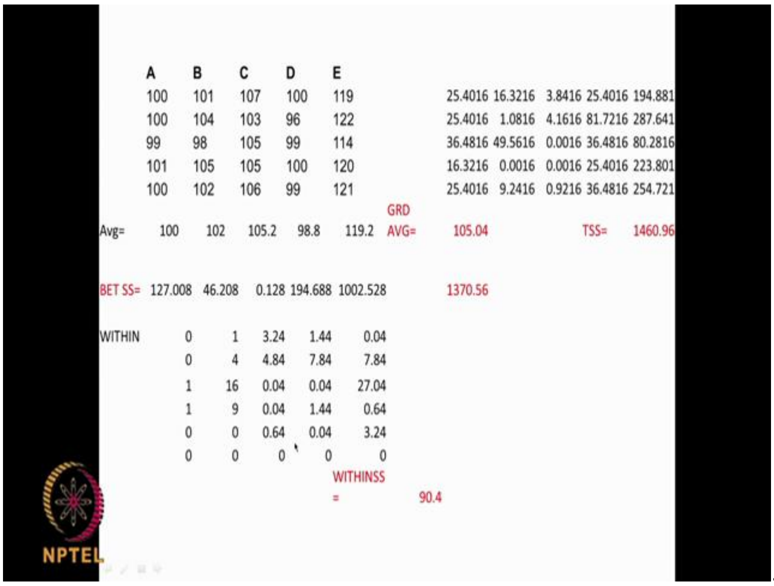
Now, total sum of squares is also equal to this plus this between sample and within sample. So, sometimes we do not have to calculate this, from these two calculations we can do that. Then we create the ANOVA table, before that let us look the hypothesis. The null hypothesis is $\sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2_4 = \sigma^2_5$; that means, all these five sets of samples come from the same population. So, the variances are equal or the alternative could be one of them are not equal. So, that is the alternate hypothesis. We can have the p value and this will be a two-tailed test. Between sum of squares you calculate from here, and then, within sum of squares which is called an error, calculate from here.

(Refer Slide Time: 05:33)



The degrees of freedom is n1 in your problem, we have five sets of groups or five sets of media we are testing, so degrees of freedom will be 4. Now, the total will be 25 is the total number of experiments minus 1 is 1 that is 24, this will be 4; so error will have the degrees of freedom as 20. Now, how do you calculate F1? F1, what you do is, you do… F1 is nothing but mean variance. So mean means BSS / n1, mean means WSS / n2. Now, F is given by F1 / F2 you are dividing by error. So, if the error is very small, F will come out to be very large, and then, we compare with the table, F for n1 and n2 degrees of freedom, and then, you accept or reject the null hypothesis.

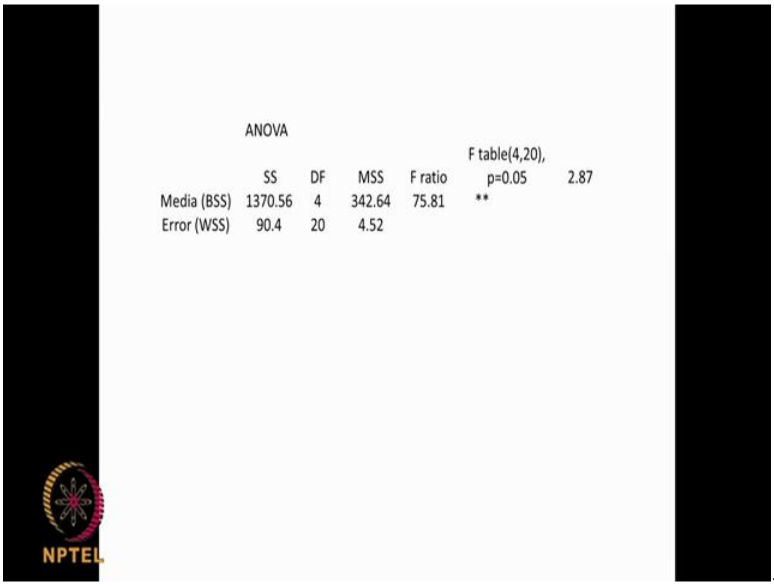
(Refer Slide Time: 06:26)


.

Let us go back to our problem. So, we have the data here given, average of each one of these media will be like this. So, we get this as 100, 102, 105.2, 98.8, 119.2. So, we take the grand average or this is the overall average. So, between sum of squares is very simple, what do you do? You take this minus this, square, multiply by 5. Why 5? There are 5 terms here, right? And then you take this minus this, square, multiply by 5, we get like this right.

This minus this, square, multiply by 5, this minus this, square, multiply by 5, then you add up all of them. So, 5 x 100 - 105.04 ^{2}. 5 you are multiplying because there are 5 times like that. Then you add up all of them, that will come to between sum of squares. Now, how do you calculate within sum of squares? That is simple. I said look at each one of the average, and then see each term how much it is varying. So, 100 - 100 ^{2} 0, 100 - 100 ^{2} 0, 99 - 100 ^{2} is 1, 101 - 100 ^{2} 1, like that. So, this will be 119 - 119.2 ^{2}, like that. Then add up all of them; that will give you within. How do you calculate the total sum of squares? Each term 100 you subtract from the grand average or the global average, square. 100 - 105.04 ^{2}, like that, you know. So, you will get 25 terms, add up all of that, that will give you total sum of squares. You see if I add the between sum of squares and within sum of squares, I will get the total sum of squares, 1370 + 90 is 1460. Then you create your ANOVA table. How do you make the ANOVA table?

(Refer Slide Time: 08:17)


ANOVA

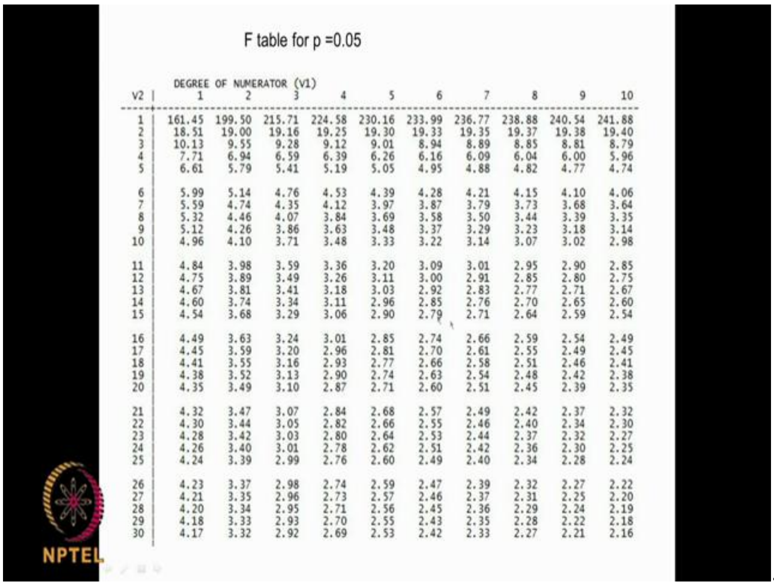
	SS	DF	MSS	F ratio	F table(4,20), p=0.05	2.87
Media (BSS)	1370.56	4	342.64	75.81	**	
Error (WSS)	90.4	20	4.52			

.

So, the sum of square is 1370 from here between that is media and then sorry, within is

90.4. So 90.4 and then total sum of square is 1460. Now media - there are 5 media, so we have 4 degrees of freedom, and then, total you have 25 data points. So, 24; 24 - 4 is 20. So, error has 20. So, how do you calculate mean? This divided by this, this divided by this. Then, how do we calculate the F ratio? We take the mean of this divided by the mean of this. So, we get F ratio. Do you understand? So that way F value is much larger than the F value we calculated here, 2.87.

(Refer Slide Time: 09:10)

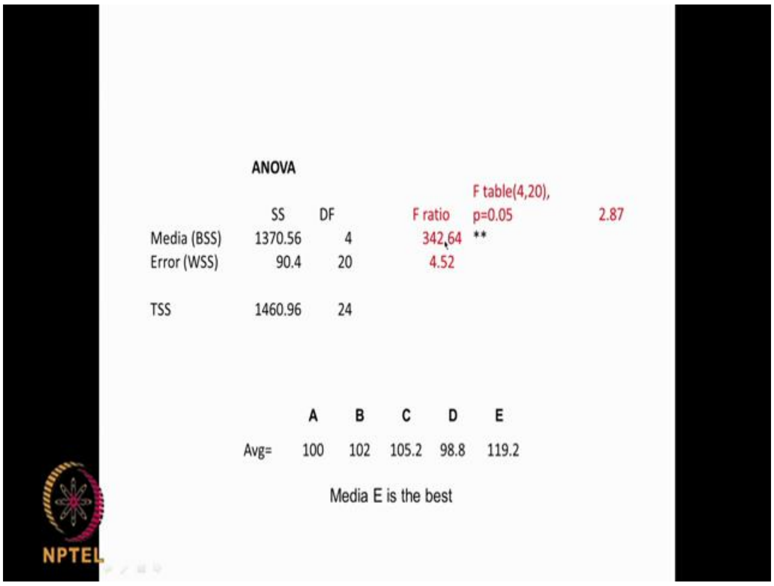


F table for p =0.05

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

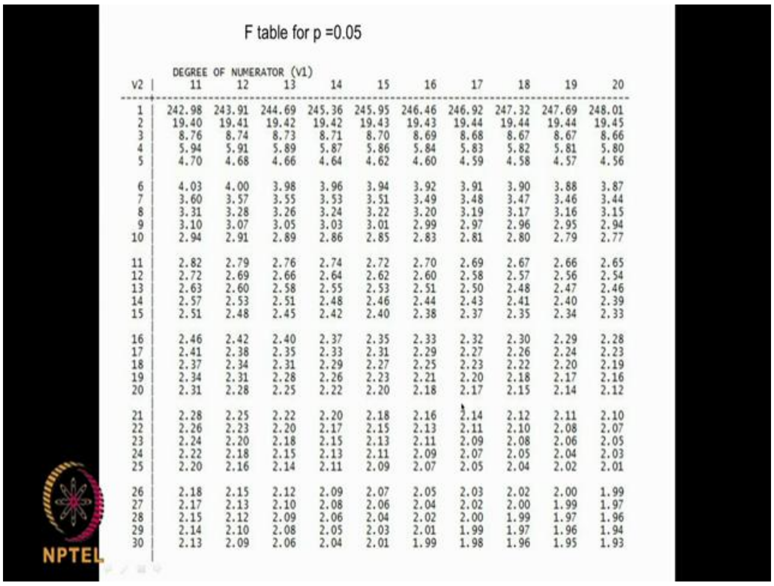
It is the table value for 4 and 20 degrees of freedom; for 4 and 20 degrees of freedom 2.87. Whereas we calculate 342.6. So, we can reject the null hypothesis.

(Refer Slide Time: 09:17)



If you look at the media average the E gives you the highest. So, obviously, I would say media is the best. Because E gives you the highest average value; you understand? E gives you the highest average. That is how you do this problem. So, we have done many times the one-way ANOVA as it is called. Because here we are considering only one media, that is why it is called the one-way ANOVA. So, we make use of these tables.

(Refer Slide Time: 09:52)



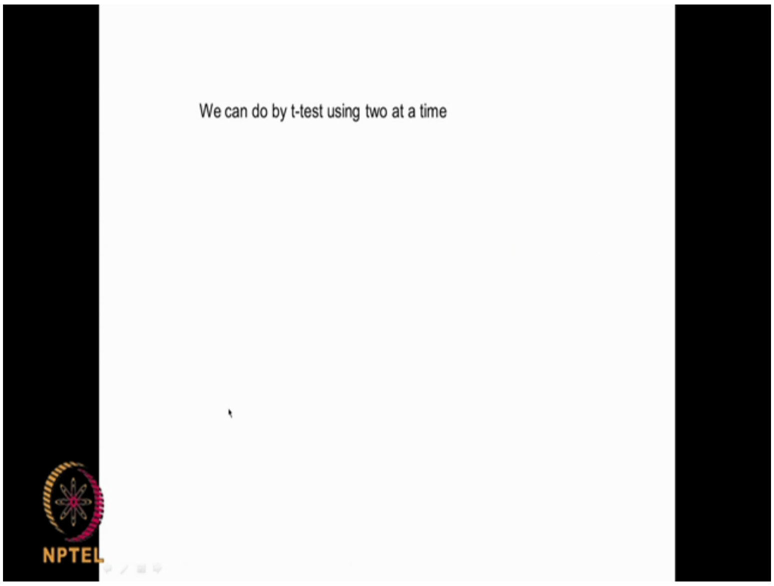
F table for p =0.05

DEGREE OF NUMERATOR (V1)

V2	11	12	13	14	15	16	17	18	19	20
1	242.98	243.91	244.69	245.36	245.95	246.46	246.92	247.32	247.69	248.01
2	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66
4	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80
5	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56
6	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87
7	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44
8	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15
9	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94
10	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77
11	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65
12	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54
13	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46
14	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39
15	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33
16	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28
17	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23
18	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19
19	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16
20	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12
21	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.11	2.10
22	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10	2.08	2.07
23	2.24	2.20	2.18	2.15	2.13	2.11	2.09	2.08	2.06	2.05
24	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03
25	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01
26	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99
27	2.17	2.13	2.10	2.08	2.06	2.04	2.02	2.00	1.99	1.97
28	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96
29	2.14	2.10	2.08	2.05	2.03	2.01	1.99	1.97	1.96	1.94
30	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93

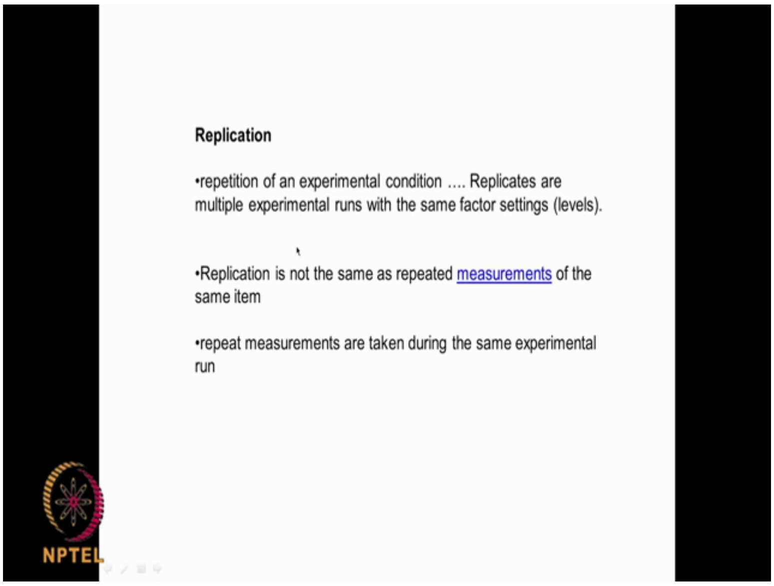
We have looked at these tables many times. So, you should be comfortable for 95 degrees of freedom, as well as, sorry for 95 % confidence, as well as 99 % confidence, and then, we have the numerator degrees of freedom, denominator degrees of freedom and so on.

(Refer Slide Time: 10:07)



We can also use t-test for this, but then how do you do the two sample t-test? We have five media, right. So, we have to do two-two at a time, so it is time consuming. So, as you can see one-way ANOVA is more powerful than the two sample t-test. Because it can look at all the many samples in one go. That is the advantage of this actually.

(Refer Slide Time: 10:32)

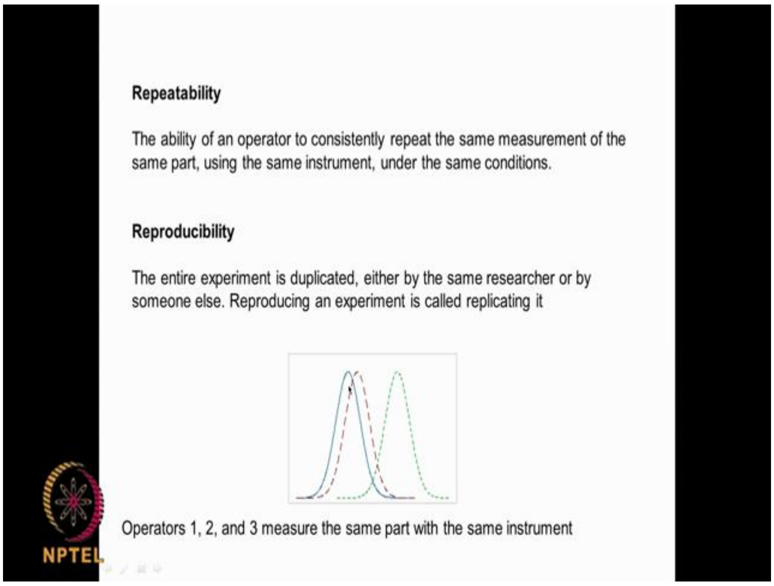


Now, let me introduce something called Replication. Replication is a repetition of an experimental condition. Like in when I am doing a bio process experiment, what do we do? I do the reaction with the media one and finally find out the biopolymer quantity. Again, I do the reaction, from the beginning. So, that is repetition of an experimental condition. So, multiple experiment runs with the same factors. So, if I am using some pH, am using some temperature, I am just changing media so A. Then I will do the whole experiment, extract the polymer, weigh the polymer, and come up with 100 milligrams per liter. Again, I do the same experiment in the fermenter, and again I extract the polymer, weigh it, and again I get another 100, like that, it is called Repetition, sorry Replication.

But that is not same as repeated measurements. What do you do in repeated measurements? Suppose I am measuring the amount of biopolymer in a high pressure liquid chromatography, I take the sample, measure it, get the amount of biopolymer. I again take a sample measure it. So, that is not same as replication; we have to remember that. So, whereas in replication you do from the beginning, keep all the parameters, and then do the whole experiment, and finally, get the amount. Whereas in repeated measurement you take one sample, then again another, take a second sample, third sample, and looking at the amount of say biopolymer, but you are not repeating the experiment from the beginning. Replication as Repeated measurement of the same item is called a repeat, whereas Replication is multiple experimental runs.

(Refer Slide Time: 12:25)

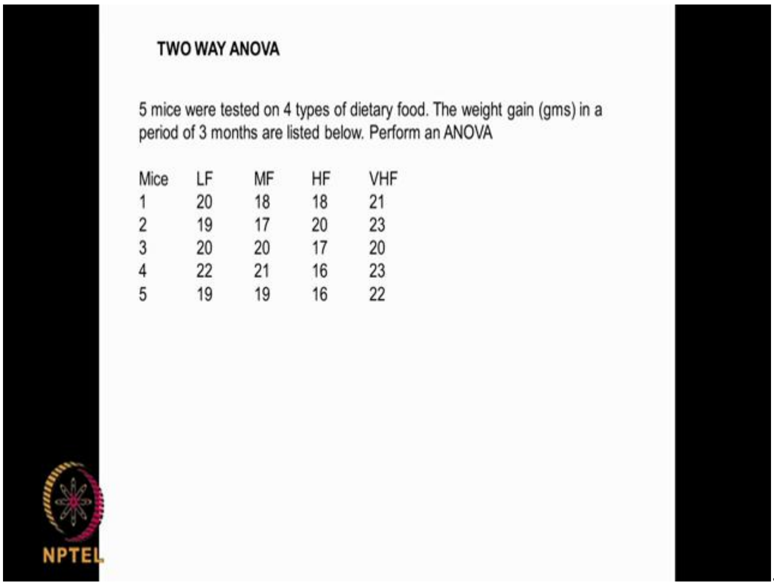
.



There are two more terms; one is called Repeatability and the other is Reproducibility. So, Repeatability is more like testing the instrument. I take a sample, check it out on the instrument; then again I measure, take the sample, check it out on the instrument. So, that is more like checking the repeatability of the instrument at the same condition. Whereas Replication is more like Reproducibility; the entire experiment is repeated; entire experiments; from the beginning you keep all the variables, set all the variables, and do the experiment, do the extraction, measure the biopolymer using a chromatography and so on. So, the entire experiment is duplicated by the same researcher or someone else. So, that is called Reproducibility. So, Reproducibility is same as Replication, but Repeatability is repeating the analysis twice or thrice and so on actually. So, Reproducibility you are trying to see say seeing whether operator 1, 2, 3 are able to reproduce a same experiment. I give the same experiment to three people and see whether they are able to reproduce it.

So, I may have a situation where one operator, second operator almost able to reproduce where as third operator is not able to reproduce. So, the Repeatability is different from Reproducibility. So, in Reproducibility you are starting from the beginning, and doing the whole experiment, and measuring your dependent variable. That is also called Replication, where as in Repeatability we may take only one sample and check the measurement. It is more like you may just do the assay twice, thrice, from different samples or you may make instrument measurements. So, it checks your error or consistency. So, the Repeatability tells you the consistent of the repeated measurements of an instrument or a part at same condition, whereas Reproducibility is the entire experiment is duplicated. Are you able to reproduce? Is somebody else able to reproduce your experiment? Somebody else in a different location is able to reproduce? So, both are very important; Repeatability and Reproducibility both are very very important actually. So, in the example which we looked at the biopolymer production in the previous ANOVA examples. So, it is a Reproducibility, it is Reproducibility or it is also called the replicated experiment.

(Refer Slide Time: 15:01)

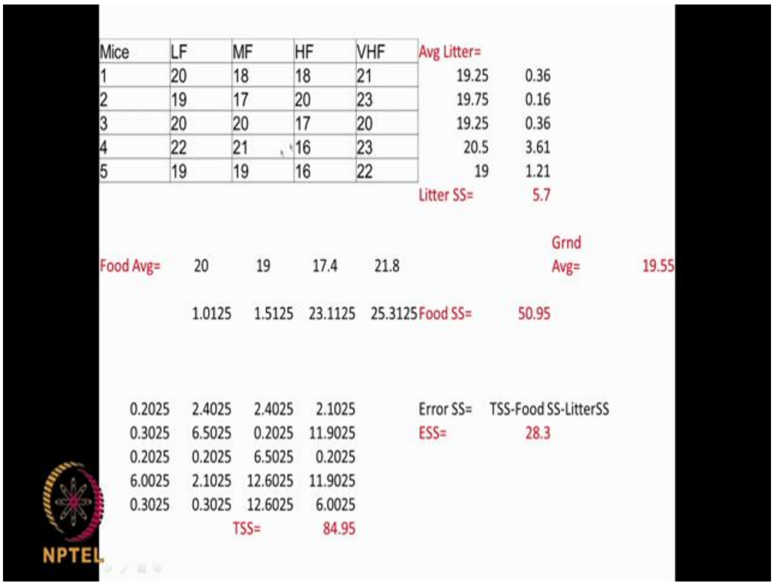


Now, let us look at this particular problem we can look at it as a replication. So, we have five mice, they were tested on four different types of dietary food, and then after three months their weight gained was measured. So, mouse one was given LF gained the weight of 20, when it was given is low fiber, medium fiber, high fiber, very high fiber. So, each mouse from a litter was tested, and then, the weight gained was measured actually. So, these are the experimental results we have.

So, we can do the same ANOVA. So, here we can do two types of ANOVA. One is based on the type of food, another is based on the mouse. Because each mouse could have come from a different litter. So, we could have one way based on the type of food -

whether it is a low fiber, high fiber, medium fiber, very high fiber. The other one could be based on the litter from which the mouse have come. So, is there going to be difference in type of food? Is there going to be a difference in the litter from which the mouse has come?

(Refer Slide Time: 16:28)



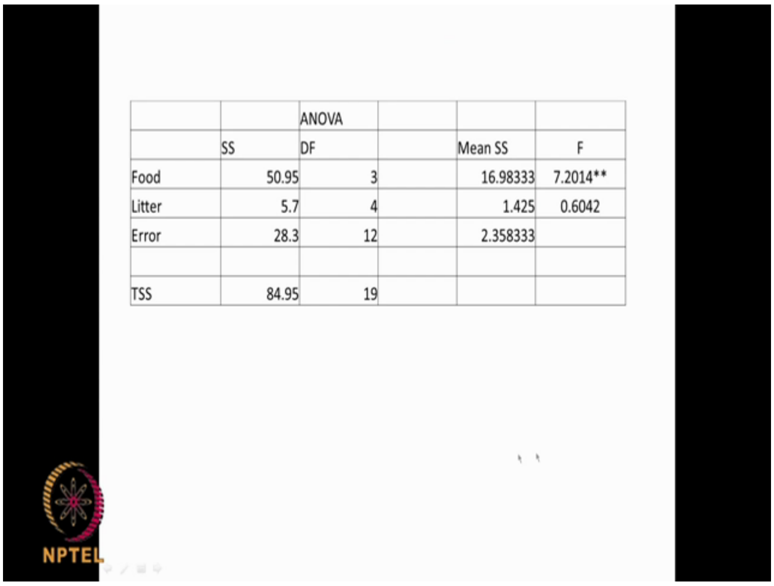
So, we can have two different situations. I will call it two-way ANOVA. So, we have the mouse, a litter, these are the food. So, I can have average for food; I can have an average low fiber food, medium fiber food, high fiber food, very high fiber food. The grand average or overall average again I take an average of all these. Now, we can also have average along the rows. So, this is an indication of the litter of the mouse. So, 19.25 is the average of these four items, 19 this is 19.75 of this average of this row and so on. And when we take an average of all these I should be getting the grand average.

Now we have a food average, we have the litter or mouse based average. Now, how do we calculate sum of squares for food? 20 - 19.5, 20 - 19.55, square it up, multiply by 5. Why do I if multiply by 5? Because there are 5 terms. So, $19 - 19.55^{2}$, multiply by 5, $17.4 - 19.55^{2}$ multiply by 5, $21.8 - 19.55^{2}$. Then add all these I will get food sum of squares. Now for litter also I can do. What do I do? How do I get this term? 19.25 is the average of this row right, multiplied ,sorry minus 19.55 square it up, multiply by 4. Why do I need to 4? Because I have four terms here, right. So, $19.75 - 19.55^{2}$ multiply by 4, $19.25 - 19.55^{2}$ multiply by 4, $20.5 - 19.55^{2}$ multiply by 4, $19 - 19.55^{2}$ multiply by 4. So, if I add up all these, I will get litter sum of squares.

How do I calculate total sum of squares? So, from the grand average I subtract each one of this terms, square it up. 20, 20 -19.55 ^{2}, 19 -19.55 ^{2}, 20 - 19.55 ^{2}, like that you know. Then add up all the terms, I will get the total sum of squares. So, we have food sum of squares, one based on litter, one based on total sum of squares. So, if I subtract food sum of squares, litter sum of squares from total I should get the error sum of squares, that is coming to 28.3, you understand. So I have got one more variable situation here. Not only the food, I also want to see whether the mouse responds differently.

So, just like I take average for the food, I take the average for the mouse, each one of the mice, because each one of the mice could be coming from a different population of the species. Then, I subtract, square, multiply by 4, because there are four terms, then add up all of them I will get litter sum of squares. Just like food sum of squares I get litter sum of squares. You know how to do the total sum of squares. So, error sum of squares comes out to be TSS minus food sum of squares minus litter sum of squares. Now I make my ANOVA table, look at this ANOVA table.

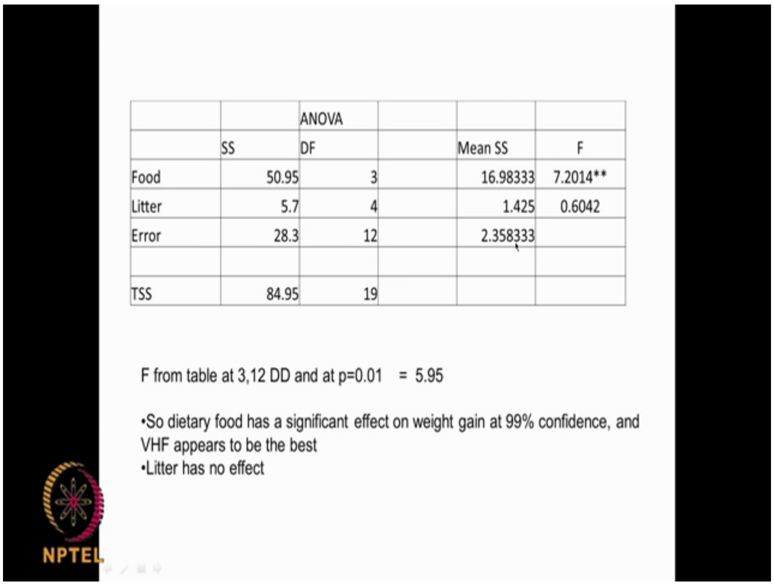
(Refer Slide Time: 19:50)



		ANOVA			
	SS	DF		Mean SS	F
Food	50.95	3		16.98333	7.2014**
Litter	5.7	4		1.425	0.6042
Error	28.3	12		2.358333	
TSS	84.95	19			

It has one extra term, food sum of squares is 50.94, litter sum of squares is 5.7, error sum of squares comes out to be 28.3, total is 84.95. Now there are four different foods. So, DF is 3; there are 5 different litter or mouse. So, 4, there are 20 data sets. So, TSS is 20 minus 1. So, error is 19 - 4 - 3, 12. So, mean 50.95 / 3, 5.7 / 4, 28.3 / 12. So, if I go on to get F 16.9 / 2.35.

(Refer Slide Time: 20:38)



		ANOVA			
	SS	DF		Mean SS	F
Food	50.95	3		16.98333	7.2014**
Litter	5.7	4		1.425	0.6042
Error	28.3	12		2.358333	
TSS	84.95	19			

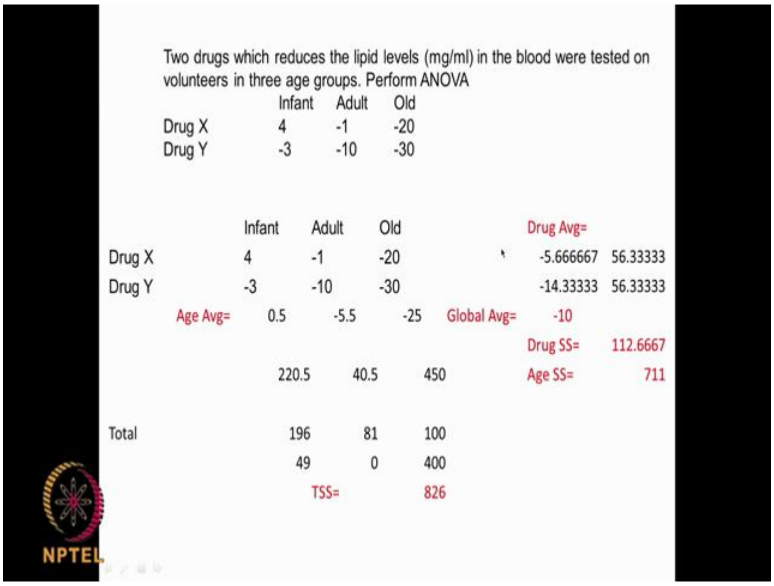
That comes to 7.2. Then for litter 1.4 / 2.35, 0.6. So, for 3 and 12 degrees of freedom, 3 and 12 degrees of freedom, at 99 for example, 3, let us go to 3 and 12 degrees of freedom for 99 is 5.95. 5.95, you get an answer of 7.2. Your calculated value is much larger, so reject the null hypothesis. So, dietary food has a significant effect gained at 99 % confidence interval.

Now let us look at this one. How did you get this? Litter divided by this 0.6. So, this has a 4 and 12 degrees of freedom; let us go back to our F table at 99 %, 4 and 12 degrees of freedom, 5.412. So, we need to get 5.412; you got only 0.6; obviously, litter has no effect, only food has an effect on the or dietary fiber has an effect on the weight gain, but the litter from which each mouse has came has no effect.

So, you are very happy about it, because you do not want that effect coming into the picture. And out of these food, which one is the best? Obviously this, VHF is the best because the average is 21.8 here, right. So, VHF is the best food for weight gain for the mouse, and food has an effect here, where as litter does not have any effect. We will call this as two-way ANOVA. Because we have two different situations - one is food, four columns we have low fiber, medium fiber, high fiber, very high fiber; and then five different litters of mouse. So, we checked whether the food has an effect, calculating F value; litter has an effect, calculating the F values. So, every time we divided by the error. So, we find the food has much has an effect, because of the F value we calculate is higher than the table value, whereas litter has no effect because the F value which you calculated is much less than the table value.

So, one important point you need to know is your errors sum of squares is small, so when you divide we get large F. In order to get good error sum of squares, I mean small error sum of squares, your DF for error also should be large. Then only we get a very small mean sum of squares. Do you understand? Do you understand the story behind this?

(Refer Slide Time: 23:30)



Let us look at another problem, which deals with a two-way situation. So, we have two drugs; it is being tested on three different age groups - we call it the old age group, we call the adult age group, we call the infant age group. So, we have the age group coming three different levels of age group; two different drugs - drug X and Y. They are suppose to reduce the lipid levels mg by ml in the blood. So, we tested on six volunteers. And these are the results. It reduces in some cases; it increases slightly in this case. Now, you can look at one is related to age, another one is related to drug; one is related to the age, another one is related to the drug. Now, if I look age average. So, just column wise I can put, right, 4 ± 3 / 2, - 1 ±10 / 2. So, these are the age average.

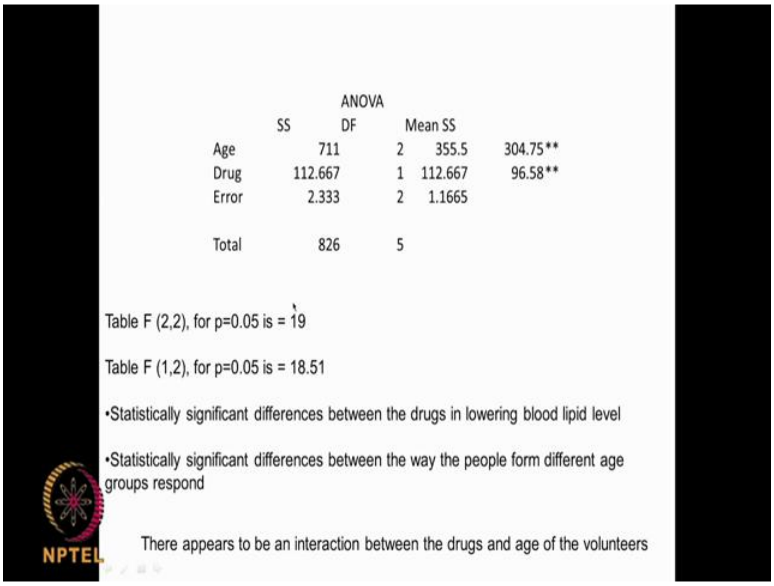
Then if we took the global, this is the global average, total average, this plus this plus this divided by 3, do you understand? Now drug average. Drug average means you add up all these divided by 3 because here we have three situations. So, it is - 20 - 21 / 4, 17, 17 / 3 will be - 5.66. For this - 30, -10, - 3, that comes to -s 43 / 3, - 14. So, this is the drug average for each one of the drug. Now, how do you cal calculate the drug sum of squares? You take this age sum of squares, you all know. You have the age average - 10^2, multiply by 2 because there are only 2 for this, then add it up. So, you understand? 0.5 this, - 10, - 5 that is 10.5^2, multiply by 2. So, these are all this sum of squares we get, we add up everything, this will give you the age sum of squares.

Let me again show it to you drug average, how do you get it? How do we get the drug average? We have this plus this plus this divided by 3. How do you get drug sum of squares? We have a 5.66 here, global is - 10. So, - 10, - 5^2. You are multiplying by 3 because there are three terms here. So, what do you do? Like that you get. Then for this what do you do - 10, - 14^2 X 3, you get this term. So, that is how you get. How do you get the age average? 2, because there are two terms, 10 minus, that is - 10 - 5^2, so you get this term. Then in this particular case - 5.5 - 10^2 multiply by 2. In this particular case - 25, - 10^2 multiply by 2 add up all of them. We will get age sum of squares, then similarly you got drug sum of squares. Now, total you all know how to do that, right? Very simple.

So, - 10 is your global average - 4^2, -10, - 1^2, - of - 1 is + 1 so 9 9 81. Then add up all these; that will give you total sum of squares. So, you got total sum of squares, you got drug sum of squares, you got age sum of squares, if you subtract both from this you will get error sum of squares. You understand? So, let us go to our ANOVA table; total sum of squares, age sum of squares, drug sum of squares, error is got from subtracting these two. Now, total is six data points. So, degrees of freedom is 5, there are 3 ages so degrees of freedom 3 - 1, there are 2 drugs so you get 1. So, error is 5, - 2, - 1 that comes to 2.

So, mean sum of squares 711 / 2. Drug is 112 / 1, error is 2.33 / 2, 1.16. Now for age F 355 / 1.1. For drug 112 / 196. Now, for this we get F 2 , 2 at 95 %. Let us look at it, 2 , 2 at 95 %, 2 , 2 at 95 % you see, 19 degrees of F value comes out to be 19, right? 19. So, this is large.

(Refer Slide Time: 28:39)

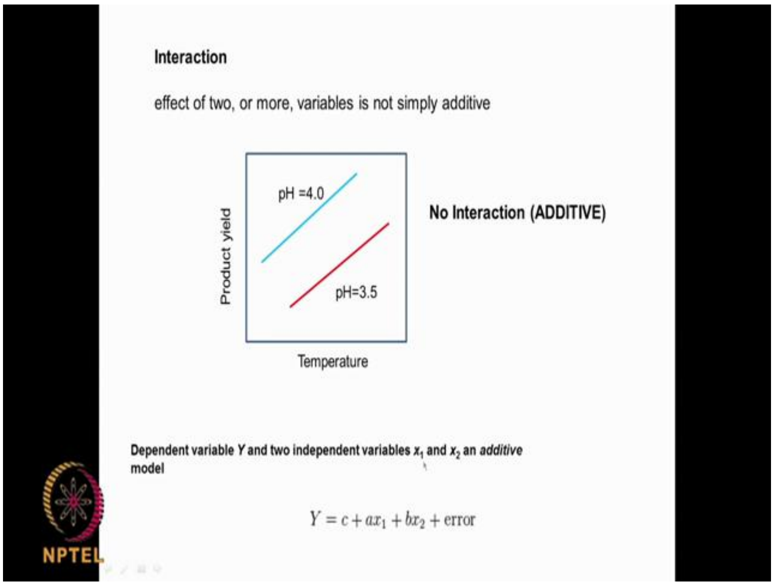


So, obviously, age has a significant effect because we have to reject the null hypothesis at 95 %. Similarly, for 1 , 2, 1 , 2 F, I got 18.5 from the table and here we have 96.58; so obviously, we reject the null hypothesis. So, statistical significant differences between the drugs in lowering the blood lipid level; statistically significant differences between the way the people from different age groups respond. So, obviously, the people respond differently for the drug. There is a statistical significant. There is a statistically significant difference between the drugs because both are significant. So, here this is called a two-way ANOVA. One way is the drug or one level is drug; other way or other level is the age.

We have three age levels infant, adult and old; and we have two drugs one is X and Y - both are significant. Because when I divide by the error both come out with the F value comes out to be very large, when compared to the table value.

Now, I want you to think another important thing. There appears to be an interaction between the drug and the age of the volunteers. Why do I see this? Have a look at this. Now, in infant drug X is increasing, whereas drug Y is decreasing in infant. If you go to adult, drug Y seems to have bigger effect than drug X. And if you go the old age, drug X and Y are almost same 20 and 30. So obviously, it is not only drug has an effect on… there is a variation in the age group, there is a variation in the performances drug, but there seems to be some sort of an interaction the way drug acts on different groups. The way drug X acts on different groups, the way drug Y acts on different groups. So, let us look at this; look at this drug X average its - 5.66, drug Y average - 14.3, you see. So, it seems to be acting very differently on different populations. So, that is called Interaction.

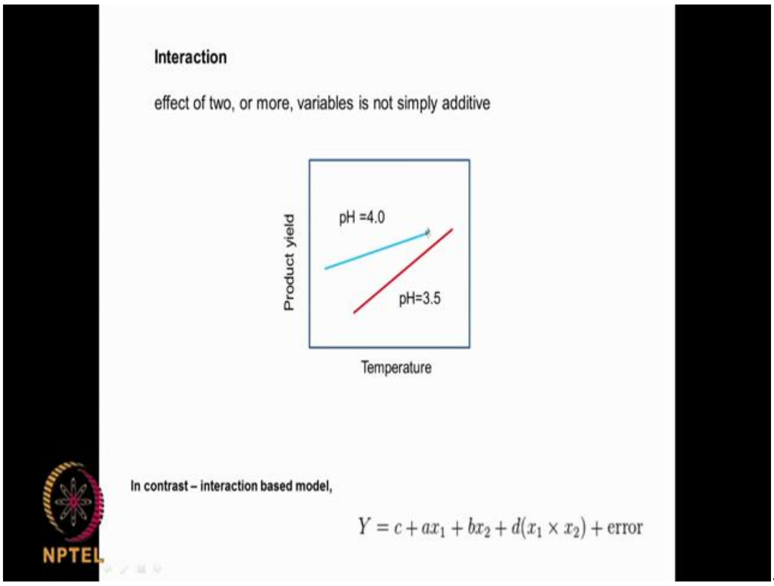
(Refer Slide Time: 31:02)



What is interaction? We will spend some time on what is interaction? Suppose, I am looking at the product yield in a fermentation. I am changing temperature and pH. So, at one pH 3.5, I am changing temperature I get a product yield like this. I am doing at pH is equal to 4, I get the product yield like this. So they look almost parallel. So, if I want to develop a model, x1 is variable 1 that is temperature; x2 is variable 2 pH. I can develop a model regression model

$$Y = c + ax_1 + bx_2 + \text{error}$$

. So, there is no interaction between temperature and pH, and this is called an additive. Because you have ax1 + bx2. When I change temperature at one pH the yield goes up like this. When I change temperature at another pH yield goes up like this, they look almost parallel. (Refer Slide Time: 32:00)

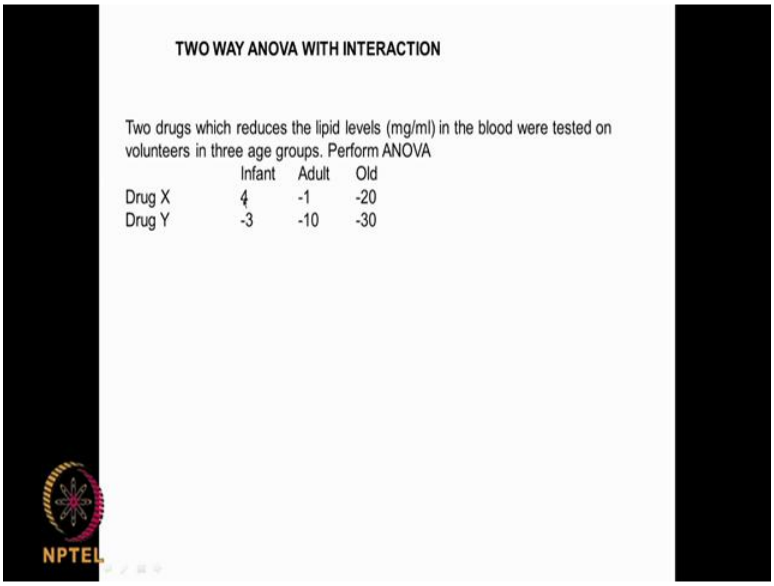


Now you can have situations like this. At pH 3.5 yield is going up, at pH 4 it is not going up as much, but it is sort of flattening. So obviously, at pH 4 depends differently for temperature when compared to pH is equal to 3.5. And this type of situation we cannot write equation like this for Y yield, we may have to write

$$Y = c + ax_1 + bx_2 + d(x_1 \times x_2) + \text{error}$$

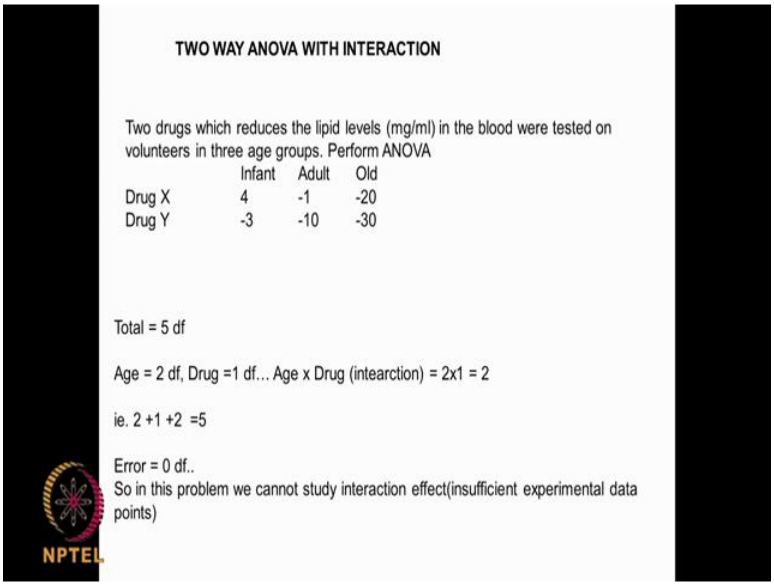
; this is called interaction, (x1 X x2). As you can see here at pH 3.5 consistently going up when you change temperature; whereas, at pH 4 it is going up and then sort of flattening out, right? So, it is not like this situation. So obviously, there is an interaction between pH and temperature or you may have a situation like this pH 3.5 it is going up like this. At pH 4 when you change temperature it is going up very high. Obviously, there is an interaction and you may have to develop model like this for interacting situations like for this or for this. Whereas if you look at this, it can be an additive model; there will not be terms like x1, x2. And that is what is happening in your drug, whereas people also. If you plot same thing, you may get same problem, right?

(Refer Slide Time: 33:24)



Drug X is like this on infant, adult, old; drug Y is like on infant, adult, old.

(Refer Slide Time: 33:32)



And obviously, there is a difference, but we cannot look at really interaction, but there is definitely there is an interaction. I want you to see this. The drug Y behaves very differently between adults, whereas it works very differently with infants, whereas it works very differently on old when compared to drug X. So obviously, there is an interaction between drug X and Y.

We will talk more about interaction in the next class as well. Because that is very very important. Interactions can happen quite a lot between drugs and type of race. Some drugs work differently on Caucasian versus African verses Asian verses European. Some media works differently with different types of bacteria, pH temperature where interactions are possible and so on actually. We will talk about interactions and how to look at interactions using analysis of variance techniques in the forthcoming classes.

Thank you very much.

Key words: Total sum of squares, between sample sum of squares, within sample sum of squares, one way ANOVA, two way ANOVA, replication, reproducibility, repeatability, interaction, additive.