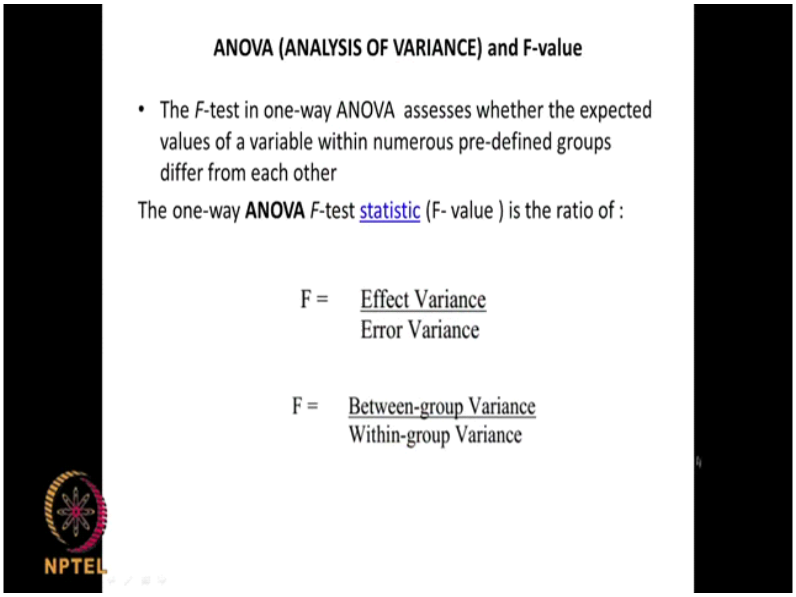
Biostatistics and Design of Experiments
Prof. MukeshDoble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 15
ANOVA

Let us continue on the course on Biostatistics and Design of Experiments. We will continue on ANOVA. ANOVA is nothing but analysis of variance. It is extremely powerful, we are going to spend lot of time on this ANOVA and we will do many problems. Like I said there are something called one-way ANOVA, two-way ANOVA, three-way ANOVA and so on and many of the software packages cannot really do such large different ways of ANOVA but we are going to do it through fundamental approach.

(Refer Slide Time: 00:41)



ANOVA (ANALYSIS OF VARIANCE) and F-value

- The *F*-test in one-way ANOVA assesses whether the expected values of a variable within numerous pre-defined groups differ from each other

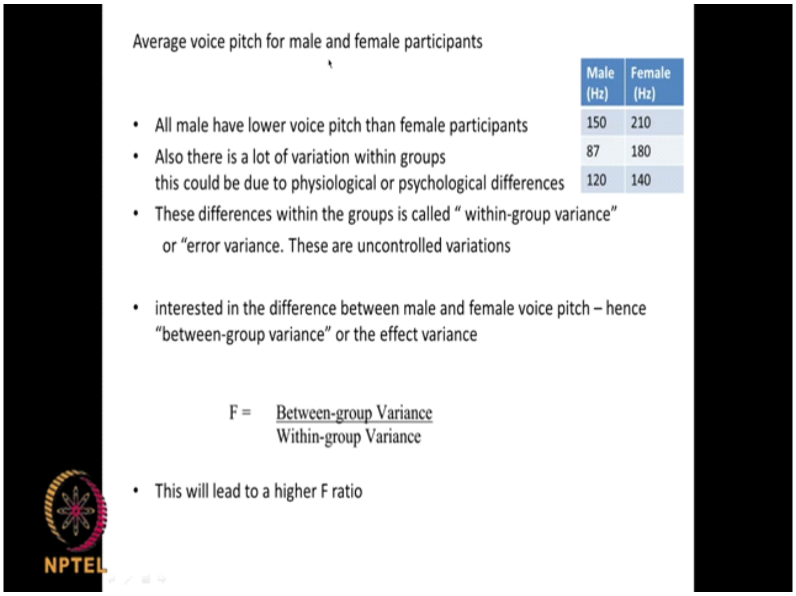
The one-way **ANOVA** *F*-test statistic (F- value) is the ratio of :

$$F = \frac{\text{Effect Variance}}{\text{Error Variance}}$$

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

So in the ANOVA we are looking at variances and then we use F-test. So what is the F? Between-group Variance divided by Within-group Variance. Within-group Variance we always call it error, so Between-group is the effect of each group or each treatment or each drug or each operator or each instruments and so on. So Within-group is when I do many times I will have certain have some variation I am looking at what is error. When we calculate this F then that is called F-test statistic and we compare it with the table and see whether the F value which we calculate is less than the table value. If it is less then we will not reject the null hypothesis, if it is more then we will reject the null hypothesis. So here generally null hypothesis will be variance from 1 is equal to the variance of 2 equal to variance of 3 that is the null hypothesis and the alternate will be not equal to. Let us carry on so what do we do, we calculate lot of sum of squares, if you remember in the previous case we calculated total sum of squares then we calculated the between groups or between sample sum of squares then we calculated within sum of squares and after that we made ANOVA table and then we calculated the F ratio. So let us look at another problem.

(Refer Slide Time: 02:13)



Average voice pitch for male and female participants

Male (Hz)	Female (Hz)
150	210
87	180
120	140

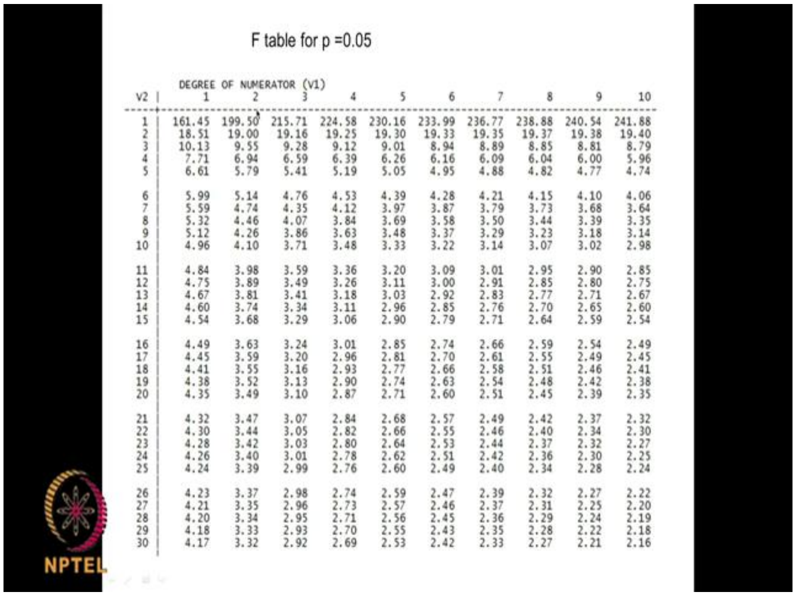
- All male have lower voice pitch than female participants
- Also there is a lot of variation within groups
  this could be due to physiological or psychological differences
- These differences within the groups is called " within-group variance" or "error variance. These are uncontrolled variations
- interested in the difference between male and female voice pitch – hence "between-group variance" or the effect variance

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

- This will lead to a higher F ratio

Now, Let us look at this problem, average voice pitch for male and female participants. So we have 3 data from male and 3 data from female, it is a small data so obviously, the degrees of freedom is less, so the F from the table will be very high. For example, in this particular case if you look at the 2 and 2 sort of thing the F values will be very high, let me show you the table.

(Refer Slide Time: 02:38)



F table for p =0.05

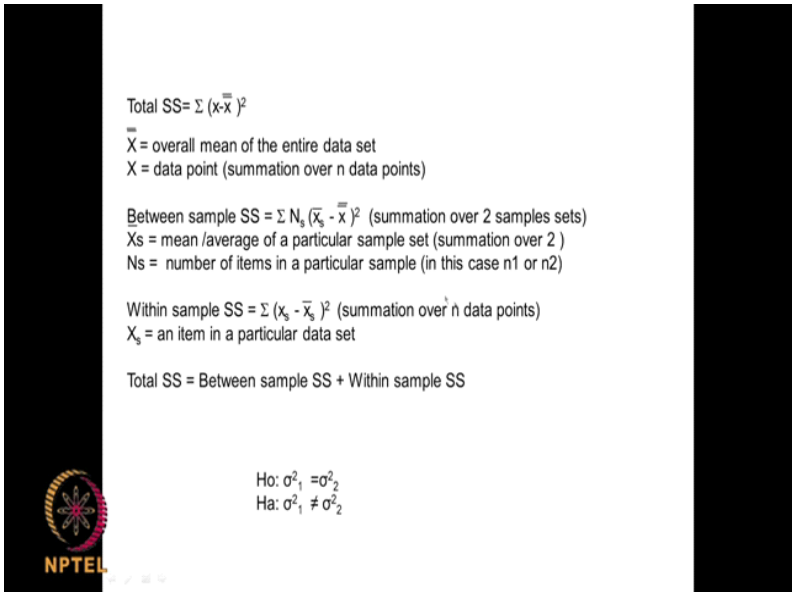
	DEGREE OF NUMERATOR (V1)									
V2	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	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

As you can see now, F values are very high especially when you have less number of degrees of freedom and as you go with higher more and more degrees of freedom you will get smaller F. So always remember that it is better to have large a data set but we have problem here, the average voice pitch of 3 male is collected and the average voice pitch of 3 female are collected, there is lot of variation as you can see, you know 87, 150, 120 could be variation because of their health condition or because of the food that they had consumed slightly before and weather condition and so on actually, let us not bother about it. There could be a physiological as well as psychological reason that is why you find such large difference. These are called within-group variance or error variance and when we compare these two we call between-group variance. So obviously, the F value we calculate between-group variance / within-group variance and we have to have a sufficiently large F much larger than the table then only we can say statistically there is a significant difference between the pitch of male vis-a-vis pitch of female.

As you can see, there is the within-group variance or error variance is going to be large because we see 87, 150 whereas here it is not much but here it is huge, here also it is huge 140 and 210. So the error variance or within-group variance is going to be large unless we have sufficiently large difference between the between-groups, our F ratios will not be sufficiently large.

Last time I taught you to how to look at various sum of squares.There are 3 important sum of squares- total sum of squares, between-group sum of squares and within-group sum of squares. Now for total sum of squares the degrees of freedom is, in this particular case 6 - 1.Between-group we have 2 groups, so the degrees of freedom is 1. So the total sum of squares is 6 - 1, 5, between groups is 1. For the errors sum of squares or within the sum of squares it will be 5 - 1, 4 degrees of freedom. We calculate total sum of squares then we calculate between sum of squares and then we calculate within sum of squares and then we make the ANOVA table.

(Refer Slide Time: 05:10)



Total SS= $\Sigma (x-\bar{\bar{x}})^2$

$\bar{\bar{X}}$ = overall mean of the entire data set
X = data point (summation over n data points)

Between sample SS = $\Sigma N_s (\bar{x}_s - \bar{\bar{x}})^2$ (summation over 2 samples sets)
Xs = mean /average of a particular sample set (summation over 2)
Ns = number of items in a particular sample (in this case n1 or n2)

Within sample SS = $\Sigma (x_s - \bar{x}_s)^2$ (summation over n data points)
X_s = an item in a particular data set

Total SS = Between sample SS + Within sample SS

Ho: $\sigma^2_1 = \sigma^2_2$
Ha: $\sigma^2_1 \neq \sigma^2_2$

Let me again show you, there is something called total sum of squares. So total sum of squares is how do we calculate x, x is any value - $x =$ what does $x =$ mean $x =$ is overall mean of the entire data because you will get a mean for the male pitch, you will get a mean for the female pitch that will be called x bar, whereas when you take the mean of those 2 means that will be called $x =$. So the total sum of squares - x - $x =^2$ that means you calculate the overall mean of the entire data set and then you subtract each of these 6 data points, subtract, square it that will give you the total sum of squares and then degrees of freedom is 5 because there are 6 data points. Understand?. Now between samples we have 2 samples,one is the male other is the female. There is an average for male then there is an average for female. Agreed?. So you subtract that with the overall mean, square it and then you multiply by number of data points in that for example, there are 3 male so we multiply by 3. So for male what do we do? We take the male average, subtract from the overall average, square it, multiply by 3 plus you take the female average, subtract it from the overall average, square it and then multiply by 3. So that is how we calculate between sample sum of squares, now within sample this gives you the error within sample is you know the male average, so subtract each one of the male terms, square it, add it up that is 3 times we do that then you take the female average, subtract each one of the female data, square it up and sum it up, we get within samples sum of squares.

As I said the degrees of freedom for total is 6 data points minus 1 is 5, for between sample we have male and female 2 - 1 is 1, 5 - 1 is 4 that will be the within samples sum of squares.You can also cross check, total sum of squares will be sum of between samples sum of squares + within samples sum of squares. The null hypothesis will be

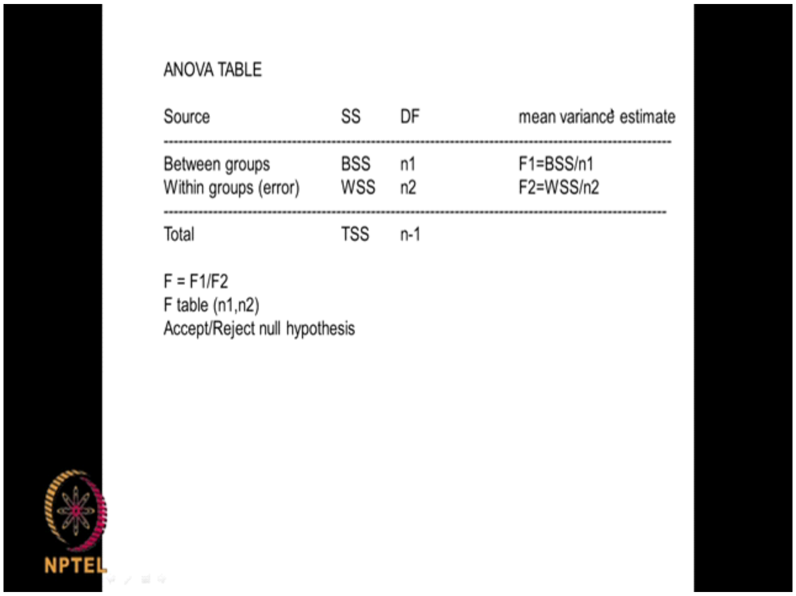
$$\sigma^2{}_1 = \sigma^2{}_2$$

that means variance for male is equal to variance for female and the alternate will be

$$\sigma^2{}_1 \neq \sigma^2{}_2$$

Once you do this calculation we need to prepare this something called ANOVA table.

(Refer Slide Time: 07:39)



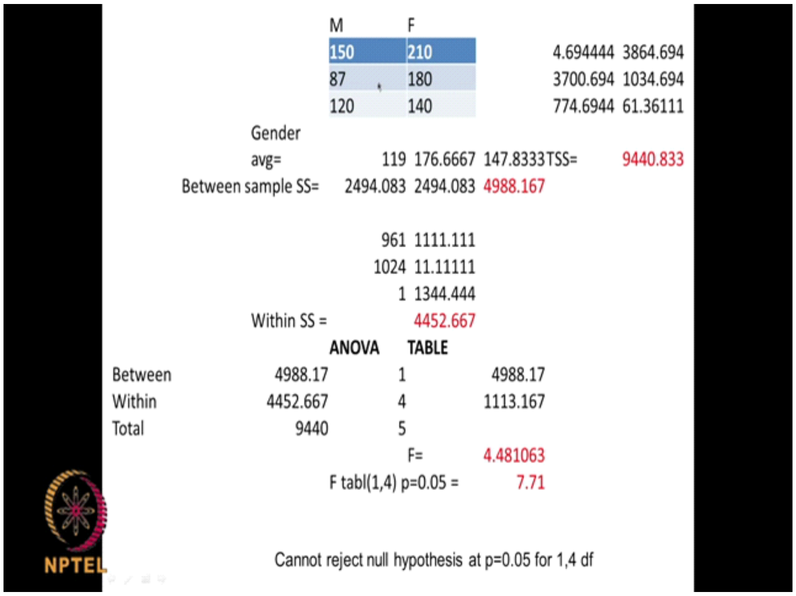
ANOVA TABLE

Source	SS	DF	mean variance estimate
Between groups	BSS	n1	F1=BSS/n1
Within groups (error)	WSS	n2	F2=WSS/n2
Total	TSS	n-1	

F = F1/F2
F table (n1,n2)
Accept/Reject null hypothesis

That is very very important that gives you the sort of summary of all our calculation. So between sum of squares as I said you do it like this and degrees of freedom will be in this case 1, then within sum of squares you do it like this and then total sum of squares you do it like this, in this particular case the degrees of freedom will be 5. So 1, 5 - 1 will give you n2. Now if I want to calculate mean so what do I do, B SS / n1 and W SS / n2 will give you the mean for these 2. Now this is error this is variance between, F1 /F2 will give me the F ratio then from the table, F table for n1 and n2 degrees of freedom I see whether they have calculated is greater than F table or F calculated is less than. So you have calculated less, obviously there is no reason for you to reject the null hypothesis, if the F calculated is greater then there is a reason for you to reject the null hypothesis.

(Refer Slide Time: 09:02)



		M	F			
		150	210		4.694444	3864.694
		87	180		3700.694	1034.694
		120	140		774.6944	61.36111
	Gender					
	avg=	119	176.6667	147.8333	TSS=	9440.833
	Between sample SS=	2494.083	2494.083	4988.167		
		961	1111.111			
		1024	11.11111			
		1	1344.444			
	Within SS =		4452.667			
		ANOVA	TABLE			
Between	4988.17	1		4988.17		
Within	4452.667	4		1113.167		
Total	9440	5				
			F=	4.481063		
		F tabl(1,4) p=0.05 =		7.71		

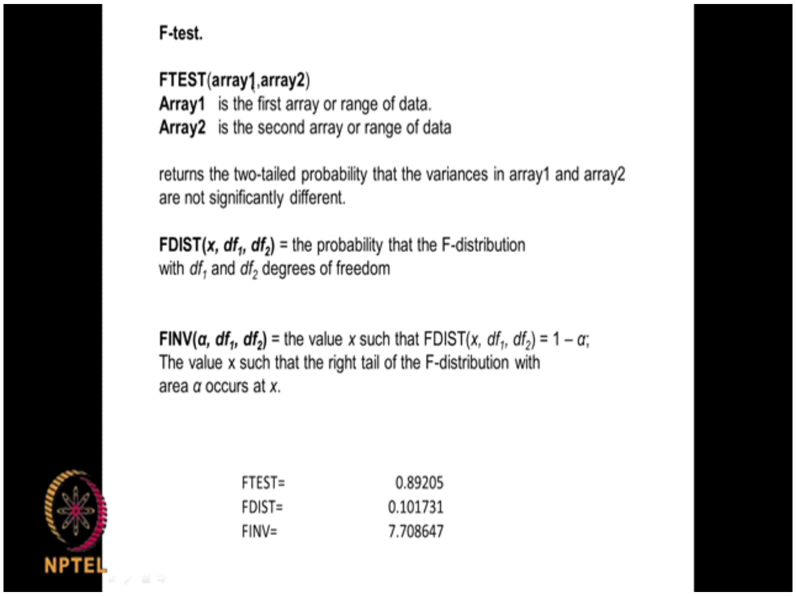
Cannot reject null hypothesis at p=0.05 for 1,4 df

Let us look at this problem 150, 87, 120 that is the male pitch and 210, 180, 140 is the female pitch. I add all these 3, divided by 3, I will get the average this is the male average, similarly add all these 3, divided by 3, this is female average. If we take an average of 119 and 176 I will get the global average that is the overall average.Now how do you calculate total sum of squares? Remember this equation, so you take each one of the data points subtract from the global average, square it up and keep on adding. So I take 150 - 147.8, subtract, square, 87 - 147.8^{2}, 120 - 147.8 ^{2}, 210 - 147.8 ^{2}, 180 - 147.8 ^{2}, 140 - 147.8 ^{2}, add up all of that will give me 9440. Understand?.This gives you the total sum of squares. Now I want to calculate between sum of squares, how do you calculate between sum of squares? if you remember you take each of the average, subtract it from the global average, square it and then multiply by the number of terms there. In this particular case, what do I do 119 - 147.8, square it, multiply by 3 because there are 3 data points, 176.6 -147.8, square it, multiply by 3, you get this.Now you add both these through that will come to 4988. So this gives you the between sample sum of squares, degrees of freedom will be 1, 40 SS degrees of freedom will be 5. How do you do within sum of squares?You know the average, you know each value so 150 - 119 ^{2}, 87 - 119 ^{2}, 120 - 119 ^{2,} 210 - 176.6 because this is for the female, 210 - 176.6 ^{2}, 180 - 176.6 ^{2}, 140 - 176.6 ^{2}, if you add up all these it comes to 4452 point. This is called within, this is an indication of the error,if you add between and within you should get the total. Now can you see that? So we have 8 +2 10, 5 + 3, 13 + 1 14, 4 + 9, 13 +

1 14, 4 + 4 +1, 9 4 4 0, that is the cross check, so between we take it as 4 9 8 8, within we take it as 4 4 5 2 and then total is 9 4 4 0, then degrees of freedom like I said there are male and female. So 1 degree of freedom for total 6 data points, 5 degrees of freedom so within will come to 5 - 1 is 4.Now mean sum of squares 4 9 8 8 / 1, 4 4 5 2 / 4 that gives you this now the F ratio is between divided by within, within is a indication of error, so it comes out to be 4.481.Now if you go to table you go to p = 0.05 what do you take? You take 1 , 4 degrees of freedom so 1 ,4 comes to 7.71,so 7.71 is your table value, calculated is 4.48 at 95 % confidence, there is no reason for you to reject the null hypothesis, cannot reject the null hypothesis this is at 1 , 4 degrees of freedom at p = 0.05. Understand?. How to do this? It is not very difficult it looks tricky, but it is not difficult. So we get the total sum of squares, we get the between sum of squares and then we get the within sum of squares. You can always subtract total minus between to get this also just like a degrees of freedom you subtract total minus between to get this so no problem. So you can avoid these set of calculation but I am teaching you how to do that that is very important. I am teaching you how to do that and so we get between and then we get total and then we get mean sum of squares by dividing by degrees of freedom and then F ratio is given by error will come in the denominator and between the groups will come or between the groups or between the gender male and female will come on the top. We get F ratio of 4.48 table gives you at 1 and 1 , 4 degrees of freedom 7.71. So there is no reason for you to reject the null hypothesis. Understood?. It is not a difficult problem to do.

Excel also can do the same thing there are different ways by which we can do it through Excel.

(Refer Slide Time: 14:19)



F-test.

FTEST(**array1**,**array2**)
Array1 is the first array or range of data.
Array2 is the second array or range of data

returns the two-tailed probability that the variances in array1 and array2 are not significantly different.

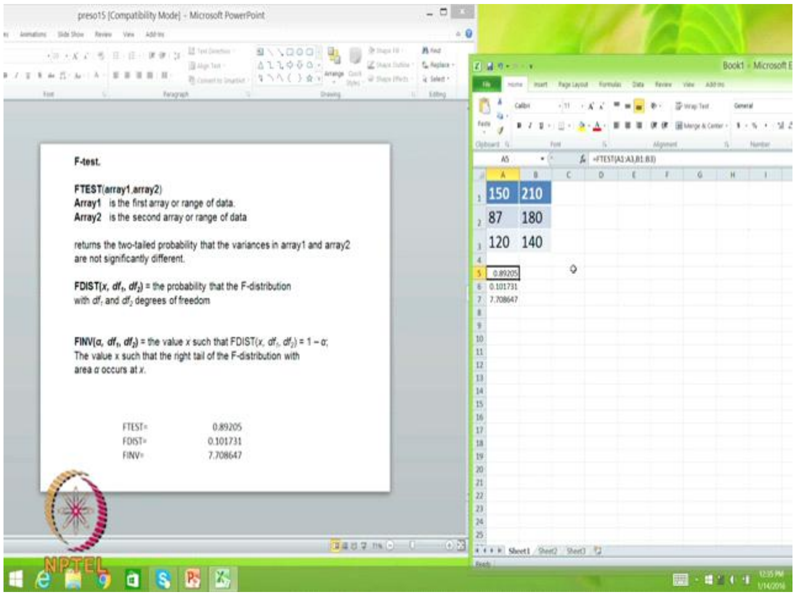
FDIST(*x*, *df*$_1$, *df*$_2$) = the probability that the F-distribution with *df*$_1$ and *df*$_2$ degrees of freedom

FINV(*α*, *df*$_1$, *df*$_2$) = the value *x* such that FDIST(*x*, *df*$_1$, *df*$_2$) = 1 – *α*;
The value x such that the right tail of the F-distribution with area *α* occurs at *x*.

FTEST=	0.89205
FDIST=	0.101731
FINV=	7.708647

There is a command called FTEST where you give array 1, array 2 and it calculates the F ratio and then it tells you, actually it does not do it through ANOVA but it does it through F test. F test is nothing but ratios of the 2 variances and then of course, you remember this FDIST command it gives you the probability given the ratio for 2 degrees of freedom and FINV command when you give the probability it will calculate the F, F value for the 2 degrees of the freedom this is exactly like table. Let us look at it also, right, how to go about doing this problem using Excel.

(Refer Slide Time: 15:03)



F-test.

FTEST(array1,array2)
Array1 is the first array or range of data.
Array2 is the second array or range of data

returns the two-tailed probability that the variances in array1 and array2 are not significantly different.

FDIST(x, df_1, df_2) = the probability that the F-distribution with df_1 and df_2 degrees of freedom

FINV(α, df_1, df_2) = the value x such that FDIST(x, df_1, df_2) = $1 - \alpha$;
The value x such that the right tail of the F-distribution with area α occurs at x.

FTEST=	0.89205
FDIST=	0.101731
FINV=	7.708647

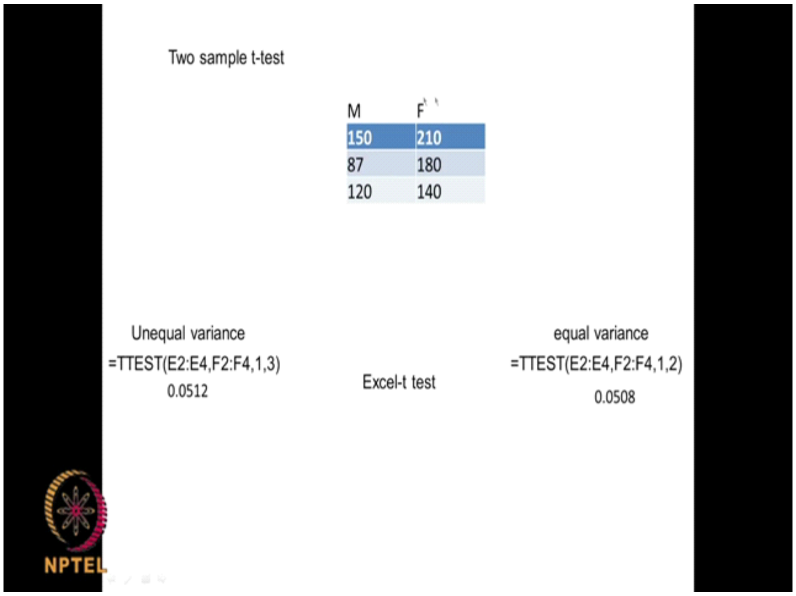
A5 =FTEST(A1:A3,B1:B3)

	A	B
1	150	210
2	87	180
3	120	140
4		
5	0.89205	
6	0.101731	
7	7.708647	

So Excel, you go to Excel it will save this, control, we say FTEST it gives you the probability, the two tailed probability to variance of array 1 and are not significantly different, it gives you the probability and as you can see the probability is 0.89. So there is no reason for you to reject the null hypothesis. Now we can do it by FDIST also, but in the FDIST you are giving the ratio.In this particular problem the ratio is 4.48, 4.48 is what you have to give and the degrees of freedom for each 1 of them, so FDIST 4.48 1 comma 4,oh sorry FDIST 4.48 1 comma 4,0.10. So for the ratio it gives you some probability 0.1,so we cannot, given the probability.It can also calculate exactly like that table using the FINV command, so 0.05 1 comma 4, it is 7.7. So you have to reach ratio of 7.7 whereas the ratio is calculated is 4.48 as I shown here, there is no reason for you to reject the null hypothesis. We can do it using Excel also but Excel does not do it through ANOVA,Excel does it just by the FTEST, FTEST is nothing but it calculates the two tailed probability for array 1 and array 2. It does not look at the errors sum of squares, it does not look at between sum of squares but it just comparing 2 variances of 2 data sets and doing the calculation. If I have 3 data sets I cannot use the FTEST command that is available in Excel,so I have to do it through ANOVA I am going to show you some problems there only if, if you have 2 data sets we can do ratio 1 / by the ratio of 2 but unfortunately if there are more than that then it is not possible, now we can also do it by two sample t test. Right?. We have 1 data set, another data set,so can we do it by two sample t test. You all know t t e s t. So data set 1, data set 2, it is a one tail or two tail, so no difference, difference and all that. So we can say one tailed distribution, unequal variance, we can say 3, 3, it is greater than 0.05, so there is no reason to reject the null hypothesis. Let us do, I did unequal variance let me try equal variance also, one sample t test equal variance also let me see x, it is still slightly greater than, so you do t test you may start wondering may be it is very close border because 0.05 for a one sample t test but the F test shows a very big difference right 4 point F test shows a large difference calculated F is 4.48 whereas we expect it to be larger than 7.70 for a statistically significant difference. So your variance of the within is very very large so obviously because of that we are not able to differentiate between the 2 data sets, that is obvious, the variances are very large, you are not able to if you improve on the variances for this is, this is coming out for a two tailed test actually, whereas the previous it is for one tailed test because here as you can see I put 1 here where as here I put 2. So variances are large that is why we are not able to differentiate between the 2 data sets that is one thing because the error variance comes in the denominator and so the F value comes out to be small that is why it is not significantly larger than the table F actually. So you need to remember this, your variances are extremely large.Look at this 87, 150 similarly 140, 210, so almost 70 difference in the pitch that means about 30 % difference. When you collect data sets, remember, you need to collect large number of data sets that is one thing and try to have more consistent data otherwise if you are going to have large variations you will not able to see between group variations at all, you know these error variance may take over and you could be in real problem of not able to see differences in the between group. Let us recall, in this particular problem we found out 3 types of sum of squares, one is called the total sum of squares and that is obtained by subtracting each one of the data point with the global average. How do you get the global average?Average for male, average for female and then take the average of average that is global in a way that is total sum of squares, straight forward.Now there is something called between samples sum of squares, for each samples you have an average, you see the difference of each one of these. So you have taken an average, there is global average. This 119 is different from 147,so you subtract this, square it up, this has happened because of 3 data point, multiplied by 3, now you subtract the female average in 147, square it up, multiply by 3, you add all these that will give you between samples.If you want within samples 119 - 150 ^{2}, 87 - 119 ^{2}, 120 - 119 ^{2}, 210 - 176 ^{2}, 180 - 176 ^{2}, 140 - 176 ^{2}, you add up that will give you within.Within is an indication of the error that comes out here, between is the effect of the group or treatment or drugs or operators or patients whatever it is, then the total one important point you need to remember is, if you add up these two, you should get the total and the overall degrees of freedom 6 data point minus 1, 5, between has a degrees of freedom of 1 because we are talking about 2 genders not male-female. So 1 within will come to be 5 - 1 is 4, one important point you need to keep in mind is this error or within should have sufficiently large degrees of freedom that is very very important. So do not have a very small number here then you are not able to really the quantify extent of error.Once we get the mean sum of squares, divide the between by error to get the F value than you compare it with the table F value, that is how you do this problem. It is not very difficult. So main take away from this is because we have in a small data set we are not able to really decipher, although you see the gender average wise the female has a higher pitch than the male we are not able to say there is a statistically significant difference because of the variance here. So variant I talked in the first or second class that variance is more important than mean which is coming out very clearly here.We see there is a very large difference actually 119 and 176 but statistically we are saying there is no statistically significant difference. Why? Because of these large variance and we need to keep that point in mind that variance is very very important that will play more important in statistically, identifying significant differences or not finding significant differences rather than the mean and ANOVA is completely based on the variances and F test is completely based on variances. So we need to keep that particular point in mind.

Let us continue more on these ANOVA problems in the next class also because I would like you to have a good confidence. Now FTEST which is available in Excel can just do a F test alone, that means it will take a ratio of 2 data sets.Suppose I have 3 data sets and FTEST, this function cannot do anything. We need to go through ANOVA and there are commercial softwares which can do that and then FINV I talked about, is exactly like that F table given the degrees of freedom, given the p that is 0.05, it will tell you what will be the F value. FDIST tells you the other way and gives you the probability, given the F ratio. The excel functions cannot really do too much of ANOVA at all in this.Now of course there are different other ways of doing that if in order to get ANOVA table we can use a regression relationship to get the ANOVA.We will look at it much later when we go around that is a problem with F test.

(Refer Slide Time: 25:45)



Two sample t-test

M	F
150	210
87	180
120	140

Unequal variance
=TTEST(E2:E4,F2:F4,1,3)
0.0512

Excel-t test

equal variance
=TTEST(E2:E4,F2:F4,1,2)
0.0508

Whereas, if you now of course you can use this a two sample t test also to do it, but then the two sample t test has limitation suppose if I had between groups male, female, infant then two sample t test means I have to take 2 such of samples one at a time and perform this type of t test which is not very comfortable and also t test looks at means, whereas F test looks at variance also.

We will continue with more problems as we go along as I said in the course of, in the next class as well.

Thank you very much for your time.

Key words: Variance, Null hypothesis, between group variance, within group variance, total sample sum of squares, between sample sum of squares, within sample sum of squares, global mean, alternate hypothesis, degrees of freedom, error, one way ANOVA, ANOVA table.